УТОЧНЕНИЕ И ОБОГАЩЕНИЕ ИНДИКАТОРНЫХ СЛОВАРЕЙ ДЛЯ АВТОМАТИЧЕСКОГО ИЗВЛЕЧЕНИЯ ИНФОРМАЦИИ ИЗ НАУЧНЫХ ТЕКСТОВ[1]

EDITING AND ENRICHMENT OF CUE DICTIONARIES FOR AUTOMATIC INFORMATION EXTRACTION FROM SCIENTIFIC TEXTS

Саломатина Н.В. *(nataly@math.nsc.ru),* ***Гусев В.Д.*** *(gusev@math.nsc.ru), Институт математики СО РАН*

Данная работа продолжает исследования авторов в области формирования и использования индикаторных словарей для отражения различных аспектов содержания научных текстов. Рассматриваются вопросы уточнения, обогащения словаря без увеличения объема обучающей подборки и построения квазиреферата текста.

Введение

Одним из методов извлечения информации из текста является использование индикаторных словарей, содержащих подсказки (cue words) об интересующих нас аспектах в виде отдельных слов, словосочетаний или более общих маркеров типа шаблонов (образцов с переменными). Применительно к научным текстам интерес представляют такие аспекты как цель работы, новизна, предлагаемый метод решения, полученные результаты и др. Основы подхода были заложены примерно в 70-е годы прошлого столетия (см. обзор [1] и цикл избранных работ [2], отражающих историю вопроса). Достаточно детальное описание подхода применительно к русскоязычным научным текстам представлено в [3], [4].

Экспериментальные исследования показали (см. [2]), что индикаторный метод дает лучшие результаты, чем частотный, позиционный и другие методы автоматического реферирования (смыслового сжатия) научных текстов, хотя ни один из них по отдельности не гарантирует решения задачи «в полном объеме». В частности, недостатком индикаторного подхода является его зависимость от жанра (для отражения содержания научного текста нужны другие индикаторы, чем для художественного или политического). К тому же формирование индикаторных словарей – довольно трудоемкая процедура, требующая привлечения экспертов и значительных затрат ручного труда. В [5] мы рассмотрели возможность частичной автоматизации этого процесса. Она основана на *вычислении L-граммных характеристик* невысокого порядка ($L = 1 \div 4$) для достаточно представительной подборки предварительно нормализованных научных текстов, их фильтрации и ручном просмотре существенно ограниченного (по сравнению с исходной подборкой) объема данных с целью отбора потенциально возможных маркеров. В качестве таковых в первом приближении выступают цепочки из L подряд следующих слов (L-граммы) с невысокой внутритекстовой частотой встречаемости, позволяющие с той или иной степенью достоверности локализовать в тексте информацию о конкретном аспекте содержания. Например, цепочки «в статье рассматриваются», «в настоящей работе», «ставится задача» сигнализируют, скорее всего, о *цели работы*, тогда как цепочки «впервые», «уникальный», «отличительная особенность», «новый подход к» обычно характеризуют *элемент новизны*.

Целью данной работы является *уточнение индикаторных словарей*, сформированных на начальном этапе (см. [5]), исследование возможностей их *обогащения*, а также *реализация процедуры построения развернутого квазиреферата текста* с использованием словарей данного типа и субъективная *оценка качества* результатов. Существенным продвижением по сравнению с [5] является переход от L-грамм к маркерам более общего вида, допускающим ограниченные замены и вставки, а также расширение числа учитываемых аспектов, что позволяет получать приемлемые по качеству квазирефераты текста.

1. Краткое резюме по работе [5]

В качестве исходного материала в [5] была использована подборка трудов конференции «Компьютерная лингвистика и интеллектуальные технологии» (Диалог'2002), содержащая 146 докладов с суммарным объемом

[1] Работа выполнена при финансовой поддержке РФФИ (проект № 06-06-80467)

порядка 442 тыс. словоупотреблений. Стандартная (но не всегда выдерживавшаяся) структура статьи включала в себя заголовок, ключевые слова, аннотацию на русском и английском языке, текст сообщения и список литературы. В *L*-граммных характеристиках отражалась вся эта информация, в виде совокупности всевозможных *L*-словных цепочек, представленных в подборке, с указанием частот их встречаемости в каждом тексте. Вместо полного просмотра всех текстов с целью выявления релевантных каждому аспекту маркеров эксперту предлагался значительно меньший по объему фрагмент *L*-граммного спектра, с большой вероятностью содержащий в себе искомые индикаторные цепочки, снабженные к тому же полезной частотной и позиционной информацией. Сокращению объема просматриваемого материала способствовало также то, что *L*-граммы были представлены в *нормализованной форме*.

Всего было выделено 12 аспектов содержания[2]: A1 – рассматриваемая проблема (задача); A2 – введение в проблему (история вопроса); A3 – цель исследования; A4 – актуальность; A5 – новизна; A6 – идея решения (обоснование подхода); A7 – предлагаемый метод решения; A8 – особенности решения; A9 – характеризация и оценка полученного результата; A10 – возможности использования; A11 – возможности дальнейшего развития; A12 – итоги, выводы. Очевидно, что многие аспекты коррелированы (например, A1, A2, A3 или A6, A7, A8), поэтому возможно сокращение числа аспектов путем их агрегирования. Это уменьшает степень дублирования одних и тех же маркеров в разных аспектах и устраняет возможность искусственного завышения веса предложения. В то же время достаточно детальная дифференциация аспектов может быть полезна, если кого-то интересует лишь один аспект, например, A6, либо когда из итогового квазиреферата желательно исключить какой-либо аспект, скажем A2, из-за ограничений по объему.

Путем просмотра *L*-граммных характеристик было выделено порядка 700 возможных индикаторов по всем 12 аспектам. Большая часть их – это цепочки длины 2 и 3. Цепочек длины 4 мало. Цепочки длины 1 (отдельные слова), в основном, представлены элементами двух- и трехсловных сочетаний, играющими определяющую роль в аспектном маркере (см. выделенные слова в приводимых далее примерах: представлять *интерес*, *главный результат*, в *ближайшее* время, авторами *предлагается* и т.п.).

2. Уточнение исходного словаря

Как уже отмечалось выше, отбор потенциально возможных маркеров осуществлялся по *L*-граммам, представленным в нормализованной форме с целью сокращения объема просматриваемого материала. Однако при поиске маркеров в исходном ненормализованном тексте такая форма представления может оказаться слишком общей и привести к появлению ложных маркеров или неправильной идентификации аспекта. Например, биграммы «авторами предлагается» и «авторам предлагается» в нормализованной форме будут выглядеть одинаково: автор\предлагаться. Однако первая из них является маркером (аспект A3), а вторая – нет. Поэтому в индикаторном словаре маркер должен быть представлен в ненормализованной или «частично нормализованной» форме, подразумевающей в данном случае возможность употребления существительного в любом числе, но обязательно в творительном падеже.

Для аспектов A9, A12 характерно использование глаголов или кратких форм страдательных причастий в прошедшем времени (решен(а,о,ы), решалась, доказана(а,о,ы), исследован(а,о,ы), исследовался и т.п.). Для аспектов A1, A3 те же маркеры будут представлены в настоящем или будущем времени (решаем(ю), решается, исследуем(ю), исследуется, будет исследоваться и т.п.).

Уточнение многословного маркера фактически эквивалентно *учету синтаксической связи* между составляющими его словами. Наличие знака пунктуации, например запятой, внутри потенциально возможного маркера чаще всего свидетельствует об отсутствии синтаксической связи между его элементами. Примером может служить конструкция типа: «… является сложной *задачей, автоматизация* которой представляется…». Здесь выделенная биграмма, записанная в нормализованном виде (задача\автоматизация), могла бы быть воспринята как ложный маркер. Введение ограничений на связь между первым и вторым словом маркера устраняет такого рода ошибки. В данном маркере слово «задача» может использоваться во всех своих формах, а «автоматизация» – лишь в родительном падеже. Выявление синтаксической связи между элементами маркера играет важную роль и в тех случаях, когда мы допускаем модификацию маркера путем замен и вставок (см. п. 3). Сохранение связи между элементами маркера в этих случаях обычно свидетельствует о том, что маркер сохраняет свою функцию, т.е. является индикатором того или иного аспекта содержания.

3. Обогащение словаря

Если выше речь шла о повышении точности поиска с использованием индикаторных словарей, то здесь мы обсудим возможности повышения другого показателя эффективности информационного поиска – полноты.

[2] В значительной мере они отражают вопросы, которые должен осветить в своей заявке любой претендент на получение гранта РФФИ

Она достигается либо увеличением объема исходной выборки и повторением процедуры обработки (см. [5]), что достаточно трудоемко, либо выявлением характера изменчивости уже отобранных маркеров и целенаправленным их варьированием с использованием «допустимых» редакционных операций. Очевидно, например, что ввиду близости понятий «задача» и «проблема» и при наличии в словаре маркера «важнейшая проблема» мы можем пополнить словарь и маркером «важнейшая задача», если таковой там отсутствует. Аналогично, если почти все глаголы, вошедшие в подсловари для аспектов А1, А2, А3, представлены в прямой и возвратной форме (решать – решаться, предлагать – предлагаться и т.п.), то при появлении нового глагольного маркера (например, анонсировать) естественно добавить в словарь и другую форму (анонсироваться).

Анализ цепочек, отобранных экспертом путем просмотра *L*-граммных характеристик, позволяет выявить следующие приемы варьирования маркеров, используемые при отражении различных аспектов содержания.

1). *Условно синонимичные подстановки* сводятся к замене отдельных элементов маркерной цепочки близкими по смыслу или функциональной нагрузке словами. Так, для оценки *значимости* проблемы (или задачи) используются подстановки следующего типа: {основная, центральная, фундаментальная, сложная, актуальная, важная и т.п.} (задача, проблема). Для указания на *прикладной характер* исследований используется другой синонимический ряд: {практическая, прикладная, конкретная, содержательная и т.п.} (задача). Отсылки на *рассматриваемую проблему* встречаются в виде: {эта, наша, данная, решаемая, настоящая и т.п.} (проблема, задача). Варьирует и *форма публикации*: {статья, работа, доклад, сообщение, проект и т.п.}. Наиболее длинный ряд связан с ответом на вопрос о том, *что делается* в данной работе, статье и т.п. Он включат в себя глаголы {исследовать(ся), обсуждать(ся), предлагать(ся), рассматривать(ся), описывать(ся), анализировать(ся), излагать(ся), приводить(ся), показывать(ся), формулировать(ся) и др.}. Эти примеры показывают, что индикаторная цепочка типа «в настоящей работе рассматривается» может быть проварьирована множеством способов путем замены второго, третьего и четвертого слова любым элементом из соответствующих им синонимичных рядов. При этом с большой вероятностью будут получены даже такие комбинации, которые *отсутствовали* в обучающей выборке.

2). *Варьирование на уровне словообразования* – другой способ пополнения индикаторного словаря, не требующий увеличения обучающей подборки. Его можно проиллюстрировать следующими наборами однокоренных слов: {важный, важность, важнейший}; {основа, основание, основанный, основной, основываться}; {описать, описывать, описываться, описываемый, описанный, описание} и др. Очевидно, что не все элементы словообразовательных гнезд могут включаться в индикаторные словари. Например, глагол «важничать» не будет включен в первый набор, поскольку не несет информации об интересующих нас аспектах содержания. Отметим также, что не все элементы словообразовательного гнезда будут отнесены к одному и тому же аспекту. Так, глагол «описываться» чаще фигурирует среди маркеров аспекта А3, тогда как «описать» естественнее отнести к аспекту А12.

3). *Использование простейших форм отрицания* иллюстрируется парами типа: {рассматриваться – не рассматриваться}, {претендовать – не претендовать}, {требующий – не требующий}, {решенный – нерешенный} и др. Этот тип варьирования почти всегда сопровождается изменением аспекта. Например, маркер «претендовать» ассоциируется с аспектом А4, тогда как «не претендовать» – с А9; «рассматриваться» служит маркером для А3, а «не рассматриваться» – для А8.

4). *Перестановка элементов длинных маркеров* ($L \geq 2$) возможна, но встречается нечасто: {проблема является – является проблема}, {несомненный интерес представляет – представляет несомненный интерес}, {речь идет о – идет речь о} и др. Принадлежность к конкретному аспекту при такого рода преобразованиях обычно не меняется, однако, учет порядка следования элементов в маркере важен для его обнаружения. Допустимое, но не зафиксированное в словаре изменение порядка приводит к пропуску маркера.

5). *Варьирование индикаторных цепочек путем ограниченных по длине вставок* – эффективный инструмент обогащения словаря и повышения показателя полноты поиска. Допуская вставки, мы переходим от полностью или частично (с точностью до нормализации) специфицированных индикаторных цепочек к образцам с переменными (см. для обзора [6]). Ориентируясь на двух- и трехэлементные маркеры, мы рассматривали обобщающие их образцы двух типов: $p_1 = c_1\ x\ c_2$ и $p_2 = c_1\ x\ c_2\ y\ c_3$. Здесь, p_1 – образец, соответствующий произвольному двухэлементному маркеру $c_1\ c_2$, допускающему вставку x между элементами c_1 и c_2; аналогично, p_2 – это обобщение трехэлементного маркера $c_1\ c_2\ c_3$ на случай, когда разрешены вставки x и y между элементами $c_1\ c_2$ и $c_2\ c_3$ соответственно. Мы ввели ограничения лишь на длины вставок (не более двух слов), но не на их лексический состав. Приведем примеры поиска образцов со вставками в той же самой подборке, по которой производился первоначальный отбор аспектных маркеров (напомним, что тексты и маркеры представлены в нормализованной форме).

Пример 1. Маркер: решение\задача. Образец p = решение\x\задача. Поиск по тексту дал следующие варианты возможных *однословных* вставок: x{поставленный, прикладной, этот, данный, настоящий, подобный, общий, такой, указанный, соответствующий, четвертый, различный}. Список вставок из двух слов значительно

короче: *x* {лишь практический, некоторый другой, следующий круг, различный аналитический}. Отметим, что эксперт при ручном просмотре 3-грамм отобрал лишь три индикатора, содержащих слова «решение» и «задача» в первой и третьей позициях, соответственно: это – {решение\этот\задача; решение\ данный\задача; решение\настоящий\задача}. Нетрудно видеть, что потенциально возможных маркеров значительно больше, чем отобранных экспертом. Это можно объяснить, как минимум, двумя причинами: 1) предварительной многоплановой фильтрацией *L*-граммных характеристик с целью уменьшения объема материала, предъявляемого на просмотр эксперту; 2) желанием эксперта ограничиться лишь теми маркерами, которые, по его мнению, в большей степени имеют отношение не к «решению задач» вообще, а к решению конкретной (этой, данной, настоящей) задачи, рассматриваемой в работе. Как бы то ни было, *анализ вставок* дает богатый материал для формирования синонимических рядов, о которых шла речь в начале данного раздела. В частности, применительно к данному примеру, было бы естественно пополнить словарь маркерами «решение\поставленный\задача» и «решение\указанный\задача».

Анализ вставок позволяет разделить их на *усиливающие* аспектную ориентацию маркера, *ослабляющие* ее (или даже разрушающие маркер) и *нейтральные*.

Пример 2. Исходный маркер: использоваться\только. Образец *p* = использоваться\x\только. Поиск образца в тексте выявил единственный вариант вставки из одного слова: *x* = {не}. Это усиливающая вставка, исходный вариант носил ограничительный характер. Сходная вставка, но уже из двух слов, появляется и в маркере «использовать\в». Варианты подстановок имеют вид: {не только; текстовый документ; именно они}. Очевидно, в словарь нужно добавить маркеры «использоваться\не\только» и «использовать\не\толь-ко».

Пример 3. Исходный маркер: играть\роль. Образец *p* = играть\x\роль. Элемент *x* (выявленный вариант однословной вставки) принадлежит множеству {пассивный, ощутимый, центральный, большой, значительный}. Здесь первый элемент множества («пассивный») *ослабляет* маркер, второй («ощутимый»), скорее, можно отнести к разряду *нейтральных*, а оставшиеся варианты вставок *усиливают* маркер.

Пример 4. Исходный маркер: этот\идея. Образец *p* = этот*x*\идея. В тексте обнаружена вставка из двух слов: *x* = {связь\возникать}. Она разрушает маркер, поскольку его компоненты оказываются синтаксически не связанными (контекст выглядит следующим образом: «… в этой связи возникает идея…»).

Пример 5. Образцы нейтральных вставок.

Маркер	*Варианты вставок*
заметить\что	{однако, наконец, также}
сводиться\к	{такой\образ}
использоваться\в	{в\частность}

В заключение данного раздела приведем без особых пояснений примеры разнесенных вставок. Отметим лишь, что их количество невелико.

Пример 6. Результаты поиска образцов вида $c_1\ x\ c_2\ y\ c_3$.

Маркер	*Варианты вставок*
в\статья\рассматриваться	*x* = {настоящий}, *y* = {не}
в\работа\быть	*x* = {рамка\данный}, *y* = {я}
в\работа\предлагать	*x* = {данный}, *y* = {впервые}

4. Формирование и анализ квазирефератов с помощью индикаторных словарей

С помощью описанных выше методов обогащения индикаторного словаря мы довели его суммарный (по всем аспектам) объем примерно до 1000 маркеров и провели эксперимент по построению квазирефератов текста на материале, не использовавшемся для обучения (это труды конференций по компьютерной лингвистике «Диалог–2005» и распознаванию образов «PRIA–7–2004»). Использовались маркеры по всем аспектам содержания за исключением A2 (история вопроса), поскольку этот аспект весьма объемен по количеству релевантного ему материала. Список литературы не учитывался.

Квазиреферат формировался следующим образом. Определялись все вхождения индикаторных цепочек в анализируемый текст. При наличии вложенных цепочек учитывалась максимальная из них. Каждому предложению назначался вес равный сумме весов попавших в него маркеров. Маркеру длины *L* (*L* = 1, 2, …) назначался вес *L* (чем длиннее маркер, тем меньше неопределенность в идентификации аспекта). Все аспекты считались одинаково значимыми, т.е. вес маркера не зависел от номера аспекта. Позиционная привязка (внутри текста) учитывалась лишь в минимальной степени и только для некоторых аспектов (например, A3, A12). Элементы «явленной» структуры (заголовок, подзаголовки, ключевые слова) не использовались. В квазиреферат включались предложения с весом 2 и выше, которые сохраняли тот же порядок, что и в тексте. При выбранном пороге в квазире-

ферат попадало в среднем порядка 7÷8 % предложений текста, общее же количество предложений с ненулевым весом было примерно в 2,5 раза больше.

Получить объективную оценку качества квазирефератов достаточно трудно. Нам неизвестны подборки, в которых рефераты были бы построены по принципу детального отражения всех аспектов работы (обычно рассматриваются лишь доминирующие: цель, полученные результаты, элемент новизны). Подготовка таких данных *в минимально необходимом объеме* требует привлечения квалифицированных экспертов. При этом результат, скорее всего, будет прогнозируемым: мнения экспертов сильно разойдутся (в качестве иллюстрации можно привести пример из [2], стр. 57: в рефератах, построенных 4-мя экспертами, отмечено только 25%-е перекрытие на уровне предложений). Поэтому мы ограничились лишь проверкой того, насколько аспекты содержания, отраженные в авторской аннотации, учтены в компьютерном квазиреферате, при построении которого *авторская аннотация игнорировалась*.

Анализ соответствий «авторская аннотация – компьютерный квазиреферат» проводился вручную на 30 первых текстах из «Диалога–2005» и «PRIA–7–2004». Лишь в двух случаях из 30 для «Диалога» и 3 случаях из 30 для «PRIA» содержание квазиреферата, с нашей точки зрения, не соответствовало авторской аннотации. Наиболее частой причиной этого являлось использование аннотации авторами не по «прямому назначению» (краткое изложение содержания текста), а в качестве элемента самого текста, содержащего определения, постановку задачи, обобщающие выводы и т.п., без дальнейшего дублирования этих аспектов в основном тексте, который является исходным при построении квазиреферата. Отмечены случаи, когда авторская аннотация просто неудачна, и основные аспекты содержания извлекаются из квазиреферата.

Наши соображения по качеству получаемых индикаторным методом квазирефератов (они же потенциал для дальнейшего развития) можно суммировать следующим образом.

1). Наблюдается некоторая зависимость индикаторных словарей не только от жанра (научные, политические и другие тексты), но и от специфики конкретной предметной области. Это проявилось в некоторых моментах при переходе от обучения, где фигурировали тексты по компьютерной лингвистике, к контролю, где рассматривались и тексты по распознаванию образов. Желательно, по-видимому, уже на этапе обучения формировать смешанную (из разных проблемных областей) подборку текстов.

2). Процедура взвешивания должна быть более гибкой и распространяться как на аспекты, среди которых есть и «факультативные», так и на маркеры внутри аспекта (см. [5]).

3). Предложения с нераскрытыми референциями встречаются почти в каждом реферате и требуют специального рассмотрения применительно к каждому аспекту (игнорировать; раскрывать; не раскрывать, но снижать вес и т.д.).

4) Из-за того, что вводная и заключительная части статей «перекликаются» друг с другом в квазиреферат попадают сходные по смыслу и лексическому составу предложения, дублирующие друг друга. Такого рода дублирование желательно автоматически обнаруживать и устранять. Одних лишь процедур выравнивания предложений для этого недостаточно (порой лишь добавление частицы «не» радикально меняет ситуацию).

5). Процедура *автоматического членения* текста на «предложения» и «фразы», считающаяся «технической», при не слишком «тщательной» ее реализации может существенно повлиять как на качество, так и на объем квазиреферата.

Заключение

По оценкам экспертов [2] и самих авторов *индикаторный подход* является весьма перспективным для *многоаспектного* извлечения информации из научных текстов. Наиболее ответственным и трудоемким этапом методики является формирование индикаторных словарей по обучающей подборке. В плане автоматизации этого процесса исследованы возможности уточнения и пополнения индикаторных словарей без привлечения дополнительного обучающего материала. Приводятся субъективные оценки качества квазирефератов, построенных с использованием сформированного авторами индикаторного словаря, включающего в себя порядка 1000 маркеров по различным аспектам содержания.

Список литературы

1. Пащенко Н.А., Кнорина Л.В., Молчанова Т.В. и др. Проблемы автоматизации индексирования и реферирования // Итоги науки и техники. Информатика, 1983. Т. 7, С. 7–164.
2. Advance in Automatic Text Summarization // Ed.by Inderjeet Mani and Mark T. Maybury. Section 1: Classical Approaches. Section 2: Corpus-based Approaches. The MIT Press. Cambridge, Massachusetts. 1999. P. 15–98.

3. Блюменау Д.И., Гендина Н.И. и др. Формализованное реферирование с использованием словесных клише (маркеров) // НТИ, 1981. Сер. 2, № 2, С. 16–20.
4. Блюменау Д.И., Афанасова Л.Н. Развитие индикаторного метода компьютерного свертывания текстов // НТИ, 2002. Сер. 2, № 5, С. 29–36.
5. Саломатина Н.В., Гусев В.Д. Автоматизация формирования индикаторных словарей и возможности их использования // Труды международной конференции Диалог-2006 “Компьютерная лингвистика и интеллектуальные технологии”, Бекасово, 31 мая–4 июня 2006. М.: Изд-во РГГУ, 2006. С. 459-463.
6. Handbook of Formal Languages // G. Rosenberg, A. Salomaa (Eds). Springer–Verlag, 1996. Vol.1. P.230–242.