АВТОМАТИЗАЦИЯ ФОРМИРОВАНИЯ ИНДИКАТОРНЫХ СЛОВАРЕЙ И ВОЗМОЖНОСТИ ИХ ИСПОЛЬЗОВАНИЯ[1]

AUTOMATION OF CUE DICTIONARIES FORMATION AND THEIR APPLICATIONS

Н.В. Саломатина *(nataly@math.nsc.ru)*
Институт математики СО РАН, г. Новосибирск

В.Д. Гусев *(gusev@math.nsc.ru)*
Институт математики СО РАН, г. Новосибирск

Идея индикаторного метода извлечения информации о различных аспектах содержания научного текста (цель, новизна и т.п.) была сформулирована еще в 70-е годы прошлого века. Узким местом в этой методике является высокая трудоемкость составления индикаторных словарей. Предлагается схема автоматизации этого процесса, существенно уменьшающая затраты ручного труда.

Введение

Автоматическое извлечение информации о различных аспектах содержания текста является актуальной задачей информационного поиска. Применительно к научным текстам интерес, в частности, представляют такие аспекты как «цель исследования», «элементы новизны», «метод решения», «полученные результаты» и др. Одним из методов извлечения подобной информации является обнаружение в тексте специфических подсказок в виде различного рода словесных клише (или образцов, маркеров) типа: «в настоящей работе», «в работе рассматривается», «целью… является», «новый подход к», «предлагается использовать», «проведенное исследование» и т.п. Эти клише являются индикаторами того или иного аспекта содержания текста. Основы индикаторного подхода к извлечению информации из текста были заложены еще в 70-е годы прошлого столетия (см. обзор [1] и цикл работ [2], отражающих историю вопроса). Достаточно детальное описание подхода представлено в [3], а возможности развития обсуждаются в [4].

Одним из препятствий к широкому использованию данного подхода на практике является необходимость составления достаточно полных *«индикаторных словарей»* для каждого аспекта содержания текста. Этот этап является весьма трудоемким, поскольку обычно такие словари составляются вручную путем просмотра значительного числа текстов квалифицированными специалистами в области информационного поиска. В данной работе рассматривается возможность частичной автоматизации составления индикаторных словарей. При этом просмотр текстов и отбор потенциально возможных аспектных индикаторов осуществляется с помощью компьютера, а окончательное решение по поводу конкретного претендента принимает эксперт путем анализа предоставляемых ему компьютером коротких контекстов (одно- два предложения), поясняющих функционирование потенциального индикатора в различных текстах.

Обоснование подхода

При отборе потенциально возможных индикаторов будем исходить из следующих предпосылок.

1). Как показывают приведенные выше примеры, в качестве аспектных индикаторов могут выступать цепочки из L *подряд следующих* слов текста (L = 1, 2, 3…), либо разрывные цепочки типа «целью … является», где вместо троеточия может стоять какое-либо допустимое слово или комбинация слов: «целью (работы, доклада, настоящего исследования и т.п.) является». Следуя терминологии из области формальных языков, разрывные цепочки можно трактовать как образцы с константными и переменными параметрами (в рассматриваемом примере имеем образец с одной переменной: «целью X является»). На данном этапе мы ориентируемся на выделение слитных «константных цепочек», а возможность появления в них замен, вставок будем отражать введением переменных, что соответствует усложнению поискового запроса.

2). Сколь ни велико многообразие возможных вариантов отражения конкретного аспекта содержания, эти варианты будут повторяться при наличии достаточно представительной обучающей подборки текстов. Поэтому элементы индикаторных словарей – это, в первую очередь, межтекстовые повторы, т.е. L-словные цепочки, встречающиеся в разных текстах.

[1] Работа выполнена в рамках проекта № 06-06-80467, поддержанного грантом РФФИ.

3). В отдельно взятом тексте конкретный маркер не должен встречаться более одного- двух раз; поскольку основные аспекты содержания научного документа (цель, актуальность, новизна, метод, результат), как правило, формулируются однократно. Повторения возможны из-за переноса отдельных (ключевых) фраз из основного текста в аннотацию, наличия корреляции между введением и заключением, а также вследствие «неудачной стилистики».

4). Чем длиннее маркерная цепочка, тем однозначнее она отражает тот или иной аспект содержания. Поэтому цепочку следует расширять до тех пор, пока не сформируется устойчивое словосочетание.

Итак, для отбора возможных маркеров мы должны иметь достаточно представительную подборку научных текстов разных авторов, уметь выделять в них межтекстовые повторы произвольной длины с учетом их морфологической вариативности и отбирать из них те, которые удовлетворяют критерию устойчивости и ограничению сверху на частоту встречаемости в одном тексте. Последнее ограничение позволяет отсеять весьма значительный пласт служебной, общенаучной и тема-рематической лексики.

Схема отбора потенциально возможных маркеров

Пусть $T = (T_1, T_2, ..., T_m)$ – анализируемая группа текстов, m – число текстов в подборке. Цепочку из L подряд следующих слов текста будем для краткости называть L-граммой[2]. *Частотной характеристикой L-го порядка группы текстов T* назовем совокупность всевозможных представленных в T L-грамм с указанием частот их встречаемости и распределения по отдельным текстам:

$\Phi_L(T) = \{\varphi_{L1}(T), \varphi_{L2}(T), ..., \varphi_{LM_L}(T)\}$,

где каждый элемент $\varphi_{Li}(T)$ $(1 \le i \le M_L)$ есть четверка $\langle$ i-я L-грамма x_i ; $F_T(x_i)$ – «текстовая частота», равная числу текстов из T, в которых представлена x_i ; $F_a(x_i)$ – абсолютная частота встречаемости x_i в T; $f(x_i) = (f_1(x_i) f_2(x_i) ... f_m(x_i))$ – вектор вхождения x_i в каждый из текстов подборки $T\rangle$. Очевидно, что параметры $F_T(x_i)$ и $F_a(x_i)$ извлекаются из $f(x_i)$: $F_T(x_i)$ – это число ненулевых компонентов в $f(x_i)$, а $F_a(x_i) = \sum_{k=1}^{m} f_k(x_i)$. Параметр M_L определяет число различных L-грамм в T (объем словаря L-грамм). Наряду с параметрами $F_T(x)$ и $F_a(x)$, где x – произвольная L-грамма из T, будем использовать еще два: $p(x) = F_T(x)/m$ – доля текстов, в которых представлена L-грамма x, и $q(x) = F_a(x)/F_T(x)$ – среднее число вхождений L-граммы x в тексты, где она фигурирует.

Совокупность частотных характеристик $\Phi(T) = \{\Phi_1(T), \Phi_2(T), ..., \Phi_{L_{\max}}(T)\}$, где $L_{\max}(T)$ – длина максимальной цепочки слов, общей хотя бы для пары текстов из T, будем называть *совместным частотным спектром* группы текстов T. Иными словами, это такой набор частотных характеристик, который содержит полную информацию о связях между текстами в виде общих цепочек с длинами $L \le L_{\max}(T)$.

Схема отбора потенциально возможных маркеров выглядит следующим образом.

Шаг 1. С помощью морфологического анализатора [5] проводим нормализацию текстов обучающей подборки T, т.е. представляем каждое слово в канонической форме. Тем самым устраняется вариативность, обусловленная словоизменением, и существенно сокращаются объемы L-граммных словарей.

Шаг 2. Последовательно для значений $L = 1,2,..., L_{\max}(T)$ вычисляем частотные характеристики $\Phi_L(T)$, формируя полный совместный частотный спектр подборки.

Полученный спектр L-грамм весьма объемен, поэтому последующие шаги связаны с устранением малоинформативных (в плане построения индикаторных словарей) L-грамм.

Шаг 3. Устраняем «уникальные» L-граммы, встретившиеся лишь в одном тексте ($F_T(x) = 1$). Среди них много таких, которые важны в плане отражения содержания конкретного текста (например, ключевые слова), но они, как правило, не являются аспектными индикаторами.

Шаг 4. Оставшиеся L-граммы проверяем на «устойчивость». На содержательном уровне «устойчивыми» считаются цепочки с длиной $L \ge 2$, встречающиеся в большом числе разнообразных контекстов. И, наоборот, максимально неустойчивыми считаются цепочки, которые могут быть лишь единственным образом продолжены вправо или влево. Это эквивалентно тому, что они не имеют самостоятельного значения и функционируют лишь в составе более длинных цепочек. На формальном уровне процедура выделения устойчивых цепочек (и соответственно, отсеивания неустойчивых) описана в [6].

Шаг 5. Отфильтровываем цепочки, заканчивающиеся (а иногда и начинающиеся) служебными частями речи. По большей части они не являются синтаксически связными.

[2] Термин впервые был использован К. Шенноном применительно к цепочкам из L подряд следующих букв, а затем перенесен на цепочки слов.

Шаг 6. Упорядочиваем оставшиеся цепочки по убыванию параметра q. Предъявляем на просмотр эксперту лишь цепочки со значениями $1 \le q \le 2$ (см. п.2 из предыдущего раздела). Исходя из списка интересующих пользователя аспектов содержания и руководствуясь собственной интуицией, эксперт формирует начальные версии словарей по каждому аспекту.

Шаг 7. С помощью специальной программы поиска образцов фиксируются для каждого аспекта все вхождения в тексты обучающей выборки цепочек, являющихся (предположительно) индикаторами данного аспекта. Каждое вхождение конкретной цепочки иллюстрируется контекстом из одного- двух[**] ненормализованных предложений. Анализ контекстов позволяет эксперту оценить точность идентификации аспекта с помощью конкретной маркерной цепочки. Если точность невысока, эксперт может удалить цепочку из словаря, перевести ее в словарь, соответствующий другому аспекту, или разместить ее, к примеру, в двух словарях, формально соответствующих разным аспектам. Например, цепочка «отличительная особенность» может служить индикатором двух аспектов одновременно: «новизна» и «особенности предлагаемого решения».

Шаг 8. После коррекции словаря на *Шаг*е 7 проводится его «обогащение» путем варьирования представленных в нем маркерных цепочек. Допустимые схемы варьирования обычно легко просматриваются путем анализа лексического состава и структуры отобранных цепочек. Чаще всего используются синонимичные или условно синонимичные подстановки ({работа, статья, доклад, сообщение, исследование}; {рассматриваться, анализироваться, исследоваться, …}), а также варьирование на уровне словообразования({рассмотреть, рассматривать, рассматриваться, …}; {предлагаемый, предложенный}; {важный, важнейший, особо важный, особенно важный}).

Описание эксперимента

В качестве исходного материала была использована подборка трудов конференции «Компьютерная лингвистика и интеллектуальные технологии» (Диалог-2002. М.: Наука, 2002. Т. 1, 2). Число текстов в подборке (параметр m) равнялось 146, суммарный объем N – порядка 442 тыс. словоупотреблений. Стандартная (но не всегда выдерживавшаяся) структура статьи включала в себя заголовок, ключевые слова, аннотацию на русском и английском языке, развернутое изложение материала и список литературы. Этапу нормализации (*Шаг* 1) предшествовала не совсем тривиальная (в силу наличия значительного числа аббревиатур, числовых данных, иноязычных вкраплений, структур типа «текст в тексте» и т.п.) процедура выделения отдельных словоформ и L-грамм, описание которой ради краткости изложения опускается. Формально в частотных характеристиках фиксировались L-граммы как на русском, так и на английском языке (из аннотаций и списков литературы), однако последние игнорировались, равно как и большинство алфавитно-цифровых L-грамм.

Предварительно были намечены следующие аспекты содержания: A1: Рассматриваемая проблема (тема, задача); A2: Введение в проблему. Состояние дел; 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. Список возможных аспектов далеко не исчерпывается приведенным перечислением, но по мере увеличения числа аспектов они становятся более «расплывчатыми» и коррелированными. С другой стороны, среди выписанных аспектов уже можно выделить группы коррелированных, например, A1, A2, A3 или A6, A7, A8, так что число аспектов можно было бы и уменьшить. Оценка «качества» отбираемых индикаторов (насколько точно они характеризуют тот или иной аспект содержания) зависит от степени взаимосвязи выделяемых аспектов.

Анализ L-граммных цепочек со значениями q в диапазоне от 1 до 2 показал, что наиболее весомый вклад в аспектные словари дают цепочки длины 2 и 3. Среди цепочек длины 4 и выше аспектных маркеров уже мало. Цепочек длины 1 (отдельные словоформы) также относительно немного, но они встречаются в значительном числе текстов. Общая закономерность такова: при близких значениях q цепочки с меньшим значением L имеют в среднем более высокое значение p, т.е. обеспечивают более высокую покрываемость подборки. Придерживаясь терминологии, используемой в [3], можно отметить, что цепочки длины 1 часто играют роль лексического сопровождения к ядерному слову (или словосочетанию).

Всего по 12 аспектам было выделено путем просмотра L-граммных характеристик около 700 потенциально возможных индикаторов. Распределение их по аспектам неравномерное. Наименьшие по объему словари характеризуют аспекты A11 (18 индикаторов) и A5 (28 индикаторов), что довольно естественно, поскольку возможности дальнейшего развития (A11) рассматриваются не в каждой работе, а элементы новизны (A5) часто не формулируются в явном виде, а как бы подразумеваются «по умолчанию». Наибольший по объему словарь характеризует аспект A6 (101 индикатор), что тоже объяснимо, поскольку обоснованию подхода обычно уделяется значительное внимание, к тому же данный аспект коррелирован со многими другими (в частности, с A2, A5, A7) и, как следствие, включает в себя некоторые индикаторы «совместного пользования». Ниже в нормализованной форме приведены примеры отобранных на *Шаге* 6 потенциально возможных индикаторов по разным аспектом с указанием их текстовой и абсолютной частоты:

[**] Два предложения обычно используются, когда маркерная цепочка содержит отсылку на предыдущую или последующую часть текста («эта задача»,…, «следующие вопросы» и т.п.)

A1: *постановка* ($F_T = 29$, $F_a = 39$);
важная\проблема ($F_T = 6$, $F_a = 7$);
проблема\заключаться ($F_T = 2$, $F_a = 2$);
A2: *существующий* (43, 72); *обсуждаться* (26, 41);
не\учитываться (10, 11);
в\настоящее\время (44, 80);
A3: *предлагаться* (48, 80); *исследоваться* (19, 21);
ставиться\задача (5, 5);
данная\работа\посвящать (5, 5);
A4: *актуальный* (40, 56); *представляться* (77, 117);
чрезвычайно\важный (8, 9);
особый\интерес\представлять (5, 6);
A5: *уникальный* (26, 37); *впервые* (16, 19);
новая\возможность (8, 9);
новый\алгоритм (6, 7);
принципиально\отличаться\от (3, 3);
A6: *исходить* (37, 49); *основываться* (20, 26);
мочь\предложить (7, 11);
предлагаемый\подход (10, 12);
строиться\на\основе (3, 3);
A7: *оптимальный* (21, 30); *применять* (23, 30);
решаться (26, 41); *быть\разработать* (14, 17);
решение\данной\задачи (4, 5);
A8: *отметить* (69, 131); *трудность* (36, 55);
не\требоваться (5, 7);
отличительная\особенность (6, 6);
следовать\обратить\внимание (4, 4);
для\повышения\эффективности (3, 3);
A9: *разработанный* (45, 79); *предложенный* (34, 50);
полученный\результат (4, 5);
быть\установлено\что (4, 4);
A10: *обеспечивать* (42, 68); *использовать\в* (10, 11);
область\применения (7, 10); *мочь\помочь* (8, 8);
позволять\работать\с (3, 3);
пользователь\иметь\возможность (3, 5);
A11: *дальнейший* (71, 121); *перспектива* (21, 26);
позволить\бы (9, 9); *в\развитие* (8, 9);
в\обозримом\будущем (3, 3);
A12: *заключение* (70, 80); *итог* (19, 27);
показать (60, 101); *сделать\вывод* (7, 13);
проведенное\исследование (4, 5);
исследование\показать\что (3, 3).

В следующем разделе на примере аспекта A3 (Цель работы) проиллюстрируем более подробно шаги 7 и 8 описываемой методики.

4. Детализация аспекта «Цель исследования»

По итогам анализа *L*-граммных характеристик (*Шаг* 6) в словарь по данному аспекту было отобрано 60 индикаторов (13 + 16 + 31 для L = 1, 2, 3, соответственно). При оценке качества сформированного словаря принимались во внимание два параметра: его полнота и точность идентификации аспекта в двух режимах: на обучающем материале и на контрольном. Под полнотой мы понимали степень покрываемости элементами словаря текстов анализируемой подборки в предположении, что абсолютно бесполезных индикаторов в словаре нет, т.е. хотя бы в какой-то доле случаев индикатор срабатывает правильно (в противном случае легко можно было бы добиться 100% полноты при нулевой точности). Под точностью мы понимали долю случаев, когда индикатор, обнаруженный в тексте, действительно указывал на интересующий нас аспект. Этот факт устанавливался экспертом (см. *Шаг* 7) путем анализа контекста каждого вхождения (одно- два предложения). Если один и тот же индикатор входил в текст более чем один раз, то идентификация осуществлялась по первому от начала текста вхождению (для аспекта A12 правило выбора было бы противоположным). Анализ покрываемости обучающей подборки индикаторами длины 2 и 3 (для них случаи неоднократного появления в тексте относительно редки) показал наличие индикаторов аспекта A3 примерно в 82% текстов. Из них более чем в 80% случаев аспект был идентифицирован правильно. Примерами наиболее сильных индикаторов, продемонстрировавшими 100%-ю точность, являются: в\статье\рассматриваться ($F_T = 10$, $F_a = 11$); в\работе\рассматриваться (7, 7); в\работе\обсужда-ться (5, 6); цель\этой\работы (3, 3) и др. Из слабых индикаторов укажем на такие как: наше\исследова-ние ($F_T = 10$, $F_a = 15$, точность идентификации –10 %, т.е.

лишь в одном из 10 текстов этот индикатор соответствовал аспекту А3); в\своей\работе ($F_T = 4$, $F_a = 5$, точность – 25%); идти\речь\о ($F_T = 4$, $F_a = 4$, точность – 25%); в\рамках\проекта ($F_T = 8$, $F_a = 12$, точность – 50%).

С целью повышения степени покрываемости подборки и точности идентификации аспекта исходный словарь был расширен путем варьирования его элементов.

Анализ лексики и структуры отобранных цепочек позволяет выявить группы «условной синонимии»: X = (статья, доклад, работа, сообщение, исследование), Y = (этот, данный, предлагаемый, представленный, настоящий,…), Z = (рассматриваться, анализироваться, обсуждаться, излагаться, описываться,…) и др. Многие из отобранных («исходных») цепочек могут быть проварьированы путем синонимичных подстановок из указанных групп. Это эквивалентно тому, что исходные цепочки заменяются классами «условно синонимичных» цепочек, представимых «образцами с переменными» вида: цель\$x$; в\$y$\статье; в\работе\$z$; x\посвящать \описанию и т.п., где переменные x, y, z могут принимать любые значения из множеств X, Y, Z соответственно. Таким образом осуществляется значительное расширение аспектного словаря. В нашем конкретном случае его объем увеличился с 60 до 207 цепочек. Другие возможные схемы варьирования на данном этапе не рассматривались.

Добавление цепочек длины 1 и варьирование цепочек длины 2 и 3 повысило степень покрываемости (полноту) до 98%, а точность до 93%, т.е. в 144 текстах из 146 были обнаружены индикаторы аспекта А3, при этом в 134 текстах они правильно идентифицировали этот аспект. Следует добавить, что один текст из 146 был на английском языке и, естественно, не содержал в себе русскоязычных индикаторов, а еще в шести мы не сумели обнаружить явного указания на цель работы даже при прочтении полного текста.

Еще один эксперимент по идентификации аспекта А3 был проведен на контрольном материале из другой предметной области (33 доклада, опубликованные в трудах конференции PRIA-7-2004, секция 1: Математические методы в теории распознавания образов). Здесь индикаторы длины 2 и 3 были обнаружены лишь в 16 текстах. Добавление индикаторов длины 1 повысило покрываемость до 30 текстов ($\approx 90\%$). При этом в 29 текстах интересующий нас аспект был идентифицирован правильно. Существенный вклад индикаторов длины 1 в данном случае объясняется тем, что во многих работах формулировка цели извлекалась из аннотаций, где она была представлена конструкциями типа: «Рассматривается…», «Предлагается…» и т.п., содержащими индикатор длины 1 в первой позиции. В [3] индикаторы такого типа являлись элементами обязательного лексического сопровождения «главного маркера». В нашем случае они сами берут на себя функции этого маркера.

Заключение

Системы целенаправленного извлечения информации об отдельных аспектах содержания текста становятся все более востребованными продуктами на рынке информационных услуг. Использование индикаторных словарей для этой цели, содержащих наводящие подсказки в виде отдельных слов, словосочетаний и более общих структур типа «образцов с переменными», позволяет во многих случаях довольно точно выйти на локальный участок текста, содержащий полезную информацию. Узким местом таких разработок является формирование индикаторных словарей по каждому из интересующих нас аспектов, что обычно делается вручную квалифицированными экспертами путем анализа значительного числа текстов. Предложенная в работе человеко-машинная методика формирования индикаторных словарей и их пополнения путем синонимического и иных типов варьирования позволяет существенно сократить затраты ручного труда экспертов и обеспечивает быструю настройку на новый аспект. Предварительные ограниченные по объему эксперименты на научных текстах демонстрируют вполне приемлемую для практических целей точность идентификации отдельных аспектов по словарям индикаторов.

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: I. Mani, Inderjeet, Maybury, Mark T., The MIT Press Cambridge, Massachusetts, 1999. p. 433.
3. *Блюменау Д.И., Гендина Н.И. и др*. Формализованное реферирование с использованием словесных клише (маркеров) // НТИ, 2002. Сер. 2, № 5, С. 29–36.
4. *Блюменау Д.И., Афанасьева Л.Н.* Развитие индикаторного метода компьютерного свертывания текстов // НТИ, 1981. Сер. 2, № 2, С. 16–20.
5. *Саломатина Н.В*. Комбинированный алгоритм морфологического анализа для нормализации неизвестных системе слов // Анализ структурных закономерностей. Вып. 174: Вычислительные системы. Новосибирск: ИМ СО РАН, 2005. С. 61–75.
6. *Гусев В.Д., Саломатина Н.В.* Алгоритм выявления устойчивых словосочетаний с учетом их вариативности (морфологической и комбинаторной) // Труды международной конференции Диалог-2004 "Компьютерная лингвистика и интеллектуальные технологии", Верхневолжский, 2–7 июня 2004. М.: "Наука", 2004. С. 530-535.