РОССИЙСКАЯ АКАДЕМИЯ НАУК

ВСЕРОССИЙСКИЙ ИНСТИТУТ НАУЧНОЙ И ТЕХНИЧЕСКОЙ ИНФОРМАЦИИ
(ВИНИТИ)

НАУЧНО • ТЕХНИЧЕСКАЯ
ИНФОРМАЦИЯ

Серия 2. ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ И СИСТЕМЫ

ЕЖЕМЕСЯЧНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ СБОРНИК

Издается с 1961 г. № 1 Москва 2013

ИНФОРМАЦИОННЫЙ ПОИСК

УДК 004.78:025.4.036:[004.65:61(051.6)]

Б.А.Кузнецов

Опыт применения поисковой оболочки с автоматическим созданием терминологических комбинаций из текста запроса для поиска в реферативных массивах на примере БД «Медицина»

Рассматривается проблема качественного поиска в реферативных и библиографических базах данных. Дается анализ двух классов поисковых средств – с применением поисковых предписаний на основе булевых комбинаций терминов и средств с применением свободных предложений на естественном языке. Отмечается, что системы первого класса дают более четкое представление о полученном результате, но требуют от пользователя высокой подготовки, и с ними очень трудно достичь высоких показателей полноты поиска. Системы второго класса проще в работе, допускают обработку более многословных запросов, ориентированы на неподготовленного пользователя. Однако выдача в таких системах требует более длительного рутинного просмотра для отбора релевантных записей.

Описан опыт использования нетрадиционной поисковой оболочки с автоматическим созданием терминологических комбинаций из текста запроса. В качестве промежуточного результата пользователю выдается множество терминологических комбинаций из текста запроса, которые содержатся в документах, найденных при поиске. Имеется удобный механизм просмотра интересующих пользователя терминологических комбинаций и собственно найденных записей, а также механизм поиска по рубрикационным индексам с выходом на них через текст запроса. Даны наглядные примеры применения системы при поиске в реферативной БД ВИНИТИ «Медицина».

Ключевые слова: *поиск на естественном языке, реферативные базы данных, медицина, автоматическая генерация терминологических комбинаций, ранжирование тематических аспектов, многотерминовые запросы, поиск по рубрикам*

1. ОЦЕНКА ПРИМЕНЯЕМЫХ ПОИСКОВЫХ СРЕДСТВ

Средства, применяемые для поиска в реферативных БД, еще далеки от совершенства и хороший качественный поиск - это скорее искусство, чем рутинная операция - чтобы найти информацию с высокой полнотой и малыми потерями, требуется высокая квалификация пользователя.

Причины становятся ясными, если проанализировать основные виды поисковых средств, применяемых на практике. Они упрощенно делятся на два класса.

1-й класс – это поиск по булевым комбинациям, когда задаются слова запроса, связанные между собой строгими связями и только такие терминологические комбинации считаются допустимыми в тексте искомого документа. Это могут быть требования обязательной совместной встречаемости в документе, абзаце, предложении, непосредственной близости слов в тексте или на определенном расстоянии друг от друга (например, чтобы между заданными основами в тексте было не более N слов) и т.д. Для выражения таких зависимостей имеется целый набор поисковых операторов, которые предлагается применять пользователю при формулировке запроса.

Если документ структурирован, например, разделен на отдельные самостоятельные поля, то существуют специальные средства поиска по конкретным выбранным полям. Традиционные библиографические и реферативные БД относятся именно к таким документам. Например, если искать документы конкретного автора, то поиск по полю автора позволяет, по мнению создателей таких поисковых средств, избежать ошибок, когда фамилия автора упоминается еще и в текстах собственно документов.

Кроме того, реферативные и библиографические БД традиционно сопровождаются развитой системой классификационных индексов, помогающих пользователю вести поиск по знакомым рубрикам. Такие поиски дают возможность простыми средствами отобрать записи соответствующей узкой предметной области.

В ряде систем предлагаются механизмы усечения слов. По мнению создателей таких систем пользователь самостоятельно может отслеживать тончайшие нюансы словоизменения и словообразования и учитывать их с помощью соответствующих операторов. Считается, что опытный пользователь с помощью простых приемов может эффективнее автоматических средств задать огромное множество вариантов словоформ: например, записать нефт$ (здесь $ - знак усечения слова), что позволяет отыскивать такие термины, как нефть, нефтяной, нефтедобыча, нефтепереработка, нефтеналивное (например, судно) с учетом всевозможных падежных словоформ и т.д.

Сторонники таких систем утверждают, что употребление на практике всего предлагаемого разнообразия средств не представляет особых сложностей – надо прилагать усилия не столько к упрощению поисковых систем, сколько к интенсивному обучению всем этим тонким механизмам поиска рядового пользователя.

2-й класс – это поиск по свободным наборам терминов, вплоть до текстовых запросов. В этом случае от пользователя не требуется знания булевой логики, поисковых операторов и правил составления запросов. Достаточно в поле запроса написать некий текст, или набор слов с использованием знаков препинания или без них. Это и объявляется собственно поисковым предписанием. Наиболее характерным примером таких систем являются практически все поисковики Интернета: Яндекс, Гугл, Бинго и т.д.

В качестве результата поиска выдается множество документов, которые ранжируются, как правило, по степени релевантности (если не предложен иной принцип – например, по дате создания документа). Принцип ранжирования документов в таких системах – это некоторая суммарная взвешенная функция, которая представляет собой сумму весов слов, включенных в текст запроса. Вес каждого слова чаще всего отражает частоту его встречаемости в базе данных, где происходит поиск.

Предполагается, что чем выше номер документа в выдаче, тем более он соответствует запросу. Пользователь, к сожалению, более точно не может, в отличие от поисковых систем 1-го класса, ответить на вопрос – какой именно набор слов из текста запроса вошел в текст документа, который он смотрит. Допустим, если запрос состоит из 8 слов, то в 1-й документ может войти 1-е, 5-е и 8-е слово, а 2-й документ включает 2-е, 4-е и 5-е и т.д. Поэтому, нередко пользователь оказывается в непростом положении – сколько надо просмотреть документов, чтобы они включали наиболее интересный с его точки зрения набор терминов запроса. В результате приходится смотреть выдачу документ за документом в надежде, что вот-вот выйдут интересующие его записи. И если «хорошие» документы перестают выходить, скажем, на 3-м десятке выдачи, то пользователь, который не может слишком много времени тратить на листание записей, решает, что больше ждать нечего и прекращает просмотр. А зря, на самом деле может оказаться, что, скажем, документ под № 42 содержит неплохую комбинацию из 1-го, 5-го и 7-го слова. Но поздно, пользователь уже прекратил просматривать выдачу.

Если дать краткую оценку применяемым средствам информационного поиска в реферативных БД, то следует обратить внимание на следующее.

Надо признать, что поиск средствами систем 1-го класса - это фактически метод проб и ошибок со значительными затратами времени пользователя на ведение сеанса, но в большинстве случаев с довольно слабыми гарантиями достижения качественного результата. Основная трудность – это выбор конкретных терминов и их количества. Традиционные запросы – это, как правило, формулировки из двух-трех слов. Как только составляется более многословная булева комбинация, выдача стремительно падает, постепенно приближаясь к нулю. Это и понятно. Слишком жесткое многословное выражение ограничивает выдачу только теми документами, которые в точности содержат термины запроса. Даже если пользователь получил в результате такого запроса несколько «хороших» документов, не надо спешить

радоваться. Часто оказывается, что «хороших» документов на самом деле несколько десятков, а то и сотен, но чтобы выйти на них, надо строить все новые и новые поисковые предписания совсем с другими наборами близких по содержанию слов. Но в процессе поиска невозможно заранее знать о допустимых потерях. Поэтому, для того чтобы избежать больших потерь, пользователь на всякий случай и ограничивает запрос двумя – тремя словами. Но здесь его ожидает другая неприятность – подавляющее число документов в выдаче может оказаться шумовыми – двух слов редко оказывается достаточно, чтобы выразить специфику содержания документа.

Теперь несколько слов о механизме усечения слов. Рассуждения о якобы практической эффективности применения механизма усечения слов вместо автоматического объединения общих словоформ по результатам морфологического анализа не выдерживают серьезной критики. Относительно удачные примеры применения усечения типа «нефт$» не могут оправдать множество других примеров, когда усечение приводит к громадному числу «ложных» основ. Возьмем для примера слова: *бедро, боец, деньги, друг, заяц, кашель, кольцо, метла, мойка, окно, парень, угол* и т.п. Еще хуже с примерами слов: *вошь, день, дно, лев, лед, мох, пес, рожь, сон, ухо, яйцо* и т.п. Здесь «тупое» усечение до «основы» дает в остатке только одну букву, за которой вместо нужного семейства близких словоформ получим целый словарь шумовых слов. В этих ситуациях останется только задавать все словоформы перечислительно без всяких усечений. Так что привычно рекомендуемый аппарат усечений сразу рассыпается на большом множестве «неудобных», но вполне реальных слов. Кроме того, это требует от пользователя совсем нетривиальных познаний в области лингвистики.

Поиск средствами системами 2-го класса большинству пользователей кажется лучше прежде всего потому, что он гораздо проще – формулировка запроса – это обычный текст. Словоизменительные, а иногда и словообразовательные варианты слов (для русского языка это важно) учитываются встроенными средствами морфологического анализа. Да и ранжированная выдача часто представляется пользователю так, что при желании он может получить выдачу вроде бы с достаточной полнотой. Для этого нужно как бы только время: смотри документ за документом, и в какой-нибудь десятке или сотне увидишь все, что нужно. На самом деле здесь есть элемент лукавства.

Пользователь практически никогда долго не просматривает выдачу, если «хорошие» с его точки зрения документы перестают попадаться. Он решает на каком-то этапе – все, дальше смотреть бесполезно. Больше ничего нужного, скорее всего, нет. И здесь практически не имеет значения, что среди череды шумовых документов где-то в конце есть и релевантные. Причина такого отношения к просмотру понятна. В поиске системами 1-го класса, по крайней мере, пользователь знает, какому конкретно терминологическому набору соответствует то, что он просматривает. А здесь такой информации нет – он смотрит что-то, может быть, релевантное с какими-то сопровождающими выдачу терминами.

Не следует также думать, что в системах 2-го класса можно писать запрос без всяких ограничений. Ограничения, по крайней мере по числу и составу слов, на самом деле есть (хотя надо признать, что они не такие жесткие, как в системах с булевым поиском). Реально может оказаться все – и что допускается много слов (5-10) и что не допускается - многое зависит от конкретного терминологического состава запроса, баз данных и т.п. Плохо то, что ситуация в целом неустойчива - пользователь заранее этого не знает – в один прекрасный момент ему, вдруг, высылается рекомендация сократить число слов в запросе, хотя ему кажется, что все слова «хорошие». Простая замена одних «хороших» слов на другие похожие по смыслу может в корне поменять всю последовательность выдаваемых документов – документы только что бывшие в лидерах сразу же могут оказаться, скажем, в третьем десятке, хотя пользователь не видит для этого совершенно никаких оснований.

С этим сталкиваются и пользователи Яндекса, Гугла и т.п.

Приведем для иллюстрации обескураживающий пример, с которым коллеги столкнулись в свое время, когда искали публикации по интересующей их теме в Интернете. В связи с всплеском информационных сообщений, связанных с присуждением Нобелевской премии по физике российским ученым Александру Гейму и Константину Новоселову, решено было поискать новые материалы по теме. Был задан следующий запрос (полагали, что формулировка по общим меркам не слишком многословна и вполне заурядна, запрос уже обрабатывали с переменным успехом в некоторых массивах другими поисковыми оболочками – казалось: что-нибудь, да найдется): «**Графен, флюорографен, Фторографен, одноатомные слои, нанослои, наночипы, наноэлектроника, двумерный тефлон**.».

В Яндексе был получен нулевой результат без всяких комментариев – выходи из ситуации как хочешь. В Гугле ответ был такой же нулевой результат, но его сопроводили назидательными рекомендациями:

*«Не найдено ни одного документа, соответствующего запросу **Графен, флюорографен, Фторографен, одноатомные слои, нанослои, наночипы, наноэлектроника, двумерный тефлон**.*

Рекомендации:

- *Убедитесь, что все слова написаны без ошибок.*
- *Попробуйте использовать другие ключевые слова.*
- *Попробуйте использовать более популярные ключевые слова.*
- *Попробуйте уменьшить количество слов в запросе».*

Ясно, что ничего, кроме раздражения, эти советы не вызывают. Тем более, что «игра» с последовательными сокращениями исходного запроса дала удручающий результат. Сокращенный до 4-х слов запрос: **Графен нанослои, одноатомные наночипы»** дал тот же нуль в выдаче. И только сокращение до трех слов вывело ситуацию из нулевого тупика.

2. ПОИСК В БД ВИНИТИ «МЕДИЦИНА» СРЕДСТВАМИ ОБОЛОЧКИ С АВТОМАТИЧЕСКИМ СОЗДАНИЕМ ТЕРМИНОЛОГИЧЕСКИХ КОМБИНАЦИЙ ИЗ ТЕКСТА ЗАПРОСА

В свое время ВИНИТИ РАН вел исследования, связанные с созданием ИПС нового класса (условно названного ДИАНА), в которых был сделан шаг к устранению некоторых из упомянутых выше недостатков. Один из вариантов такой ИПС получил название «Ариадна» Она была использована для поиска на CD-ROM в созданной ВИНИТИ базе данных «Медицина».

Запросы в этой системе принимаются в виде набора слов, свободных предложений или классификационных индексов. Запросы допускается формулировать либо на русском, либо на английском языке. Двуязычие необходимо, так как в БД имеется большое число названий или наименований на английском языке, кроме того попадаются и англоязычные рефераты. В запросе допускается смешение как русскоязычных, так и англоязычных слов. Встроенные средства морфологического анализа обрабатывают как русскоязычные, так и англоязычные слова. Таким образом, пользователю нет необходимости приводить слова к какой-либо канонической форме (например, именительному падежу, мужскому роду, единственному числу и т.п.).

Считается, что не стоит выходить за пределы двух десятков слов – этого обычно достаточно для практических целей. Если это ограничение соблюдается, то гарантируется, что система обязательно примет запрос к обработке и не выдаст никаких рекомендаций по необходимости, например, сократить запрос. На самом деле общее количество слов в запросе может быть и больше (достигать, скажем, трех десятков), но на момент запуска оболочки с БД «Медицина» достаточных испытаний с такими запросами не проводилось, хотя и отказов не наблюдалось. Система морфологического анализа построена таким образом, что время от времени появляющиеся в медицинских массивах неологизмы обрабатываются должным образом и словоизменительные варианты учитываются. Устойчивость системы с этой точки зрения подтвердилась реальным опытом многолетней эксплуатации БД «Медицина».

Чтобы традиционные потребители БД ВИНИТИ были обеспечены привычными средствами поиска по классификационным индексам, поисковые файлы БД включают 2 части: собственно реферативная БД «Медицина» и дополнительная – «рубрикатор БД «Медицина».

Массив «рубрикатор БД «Медицина» строится автоматически при загрузке исходного массива записей БД «Медицина» в поисковую оболочку. Дело в том, что этот массив не является раз и навсегда заданным. Он строится по рубрикам, которые содержатся в загружаемых записях. Каждое рубрикационное поле каждой записи обрабатывается и включается в массив «рубрикатор БД «Медицина».

Особенность использования функции – поиск по рубрикам - следующая. Поиск в массиве «рубрикатор БД «Медицина» происходит по той же схеме, как и поиск собственно в БД Медицина, но с той разницей, что в качестве результата выдаются наименования (и коды рубрик), которые содержат те или иные сочетания терминов поискового предписания. Строго говоря, выдаются в качестве результата не только те рубрики, которые явно включают термины запроса, но и те рубрики, в текстах подрубрик которых есть эти термины. Поэтому не следует удивляться, когда выдается рубрика, в которой вроде бы и нет ни одного слова запроса. Это просто означает, что слова попали в какие-то подчиненные рубрики.

Сначала опишем общую схему обработки запроса, а затем, на реальных примерах коснемся деталей.

Отличительной особенностью метода поиска является то, что в качестве промежуточного результата выдаются не сами документы, а перечень терминологических комбинаций запроса, по которым имеется один или много документов. Анализируя эту терминологическую матрицу, пользователь видит, по каким комбинациям имеются реальные записи и сколько их. Это очень важная для пользователя информация. Он просматривает какую-то терминологическую комбинацию и видит, что документов, которые ее содержат, столько-то (1,5, 20 и т.д.). После этого, нажав на соответствующий терминологический набор, можно просмотреть все записи, которые ему соответствуют.

Что здесь особенно ценно для пользователя. Он видит не «что-то релевантное», а документы, которые точно включают вполне конкретный набор терминов (например: «инфаркт», «миокарда», «ишемическая», «сердца», «гипертензия»). Все найденные терминологические наборы легко просматривать. Где-то на другой строке можно увидеть, например: «ишемическая», «болезнь», «сердца», «давление». При развернутом запросе (скажем в десяток слов) общее число рангов (терминологических сочетаний) может быть больше сотни. Совершенно очевидно, что при использовании традиционных средств булевого поиска «угадать» все такие реально встречающиеся в БД комбинации абсолютно немыслимо, не говоря уже о безумных затратах времени.

Все найденные терминологические комбинации ранжированы. Механизм ранжирования довольно сложен. Но упрощенно можно сказать, что в верхних рангах число терминов и их уникальность выше (т.е. более редкие термины имеют приоритет), таким образом, чем ниже ранг, тем ниже его оценка. При выдаче это обстоятельство учитывается. Например, можно заказать выдать все записи во верхним рангам – с 1-го по заданный, т.е. выдать, не помечая каждый из них. Тогда по умолчанию будет выдана наиболее «важная» часть найденной коллекции.

Возможности простого поиска по многотерминовым запросам меняет традиционный взгляд на поисковые системы. Обычно считается, что по мере роста числа слов запроса точность повышается, а полнота падает. У нас возникает парадоксальная, но благоприятная для поиска ситуация. Чем больше слов в запросе, чем более подробно составлено поисковое предписание, тем выше одновременно и точность, и полнота. Этому обстоятельству есть объяснение. Увеличение числа терминов запроса приводит лишь к росту разнообразия актуальных терминологических комбинаций, оказавшихся в документах БД, но все они доступны для обозрения и вывода документов. Поэтому, чем больше терминологических входов в документы БД, тем полнее представлено содержательное описание. В традиционных системах с булевым поиском ситуация иная. В качестве поискового предписания представлена фактически только одна из строго заданных формул – остальные возможные, (которые у нас автоматически сгенерированы из исходного текста запроса и представлены в виде самостоятельных рангов) выброшены из рассмотрения. Высокая точность достигается за счет того, что каждый ранг (терминологический набор) максимально насыщен словами из исходной текстовой формулировки – каждый ранг это фактически строгая конъюнктивная нормальная форма и чем больше в ней элементов, тем выше точность.

Для иллюстрации приведем пример обработки одного из запросов в БД «Медицина» средствами ИПС «Ариадна». Запрос 1 состоит из 14 слов: «*ишемическая болезнь сердца инфаркт миокарда лекарственная терапия лечебные препараты хирургическое вмешательство аортокоронарное шунтирование АКШ*». Последнее слово АКШ – это принятое в описаниях сокращение термина «аортокоронарное шунтирование».

На рис. 1 показан результат обработки запроса (часть текста запроса можно видеть в поле «Запрос». Поиск проведен в БД ВИНИТИ «Медицина» за 2010 г. Общее число автоматически сгенерированных самостоятельных терминологических сочетаний - 244

На рис. 2 показана часть страницы результата поиска. Приведено содержание терминологических сочетаний верхних 8-ми рангов. Слева в каждой строке с текстом терминологического сочетания указано количество документов, которые его включают (в данном случае – по каждому из 8-ми сочетаний имеется 1 документ).

На рис. 3 приведена экранная выдача документа 1-го ранга. В первой строке указан терминологический состав ранга найденного документа. В верхнем поле приведено библиографическое описание. В среднем поле указаны ключевые слова документа. Ниже следует текст реферата. Курсивом в тексте выделены слова документа из терминологической строки. В нижней строке экрана слева указано общее число найденных по запросу записей – 1019 по всем рангам.

На рис. 4 для иллюстрации приведены терминологические сочетания нижних рангов (начиная примерно со 120-го ранга – местонахождение можно видеть по ползунку положения ранга справа).

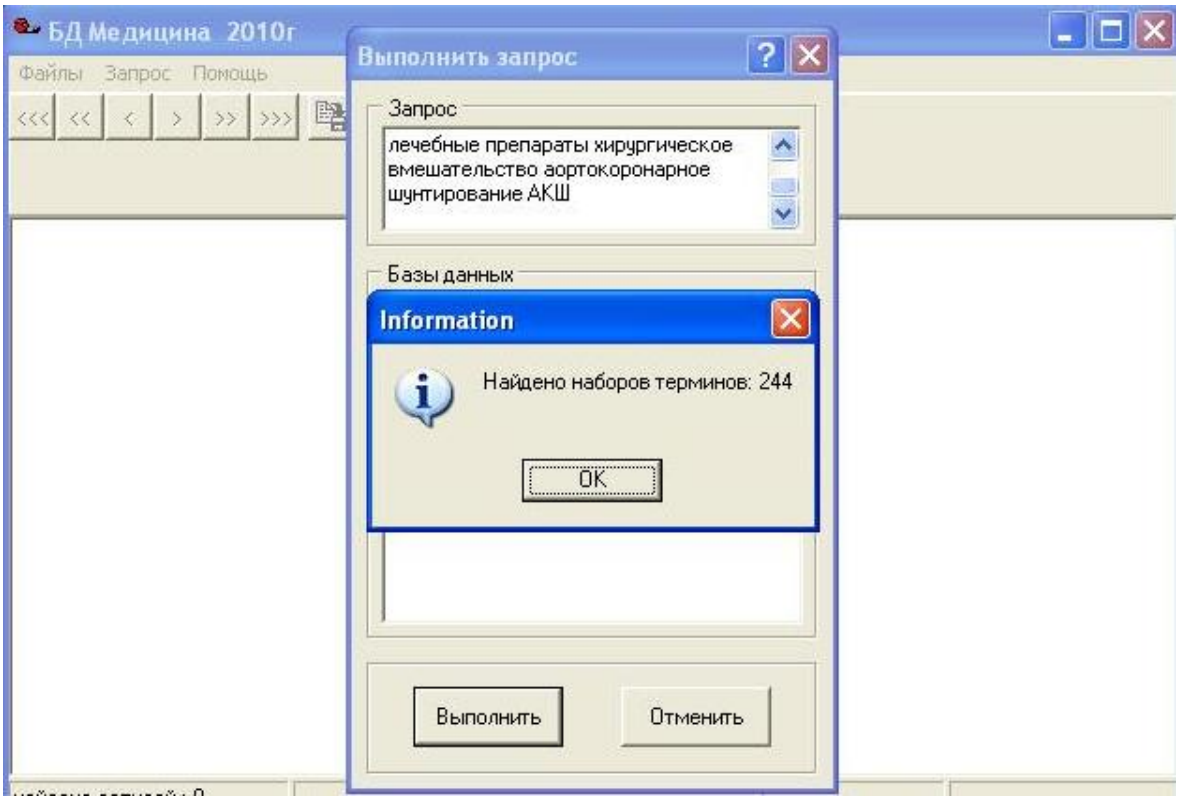

Рис.1. Поиск по запросу 1

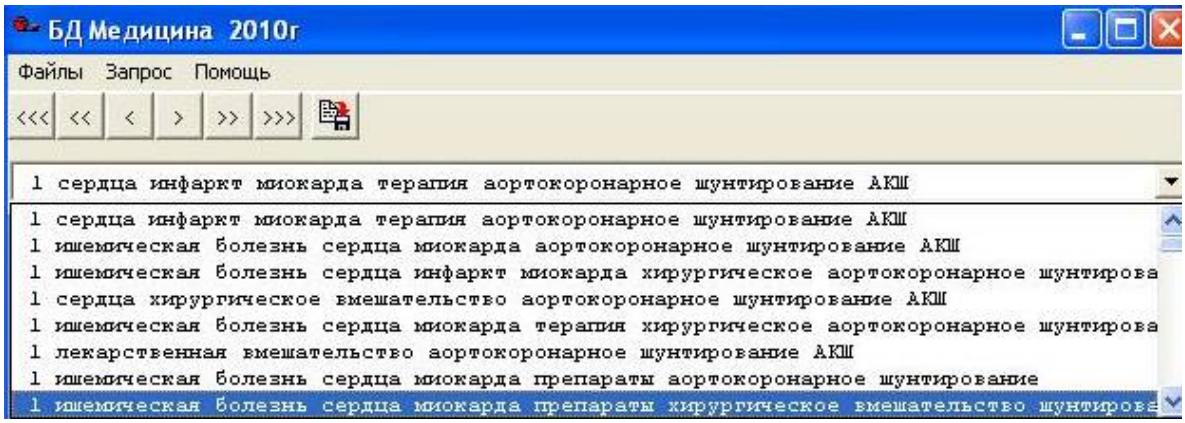

Рис.2. Результат поиска по запросу 1 – терминологические комбинации

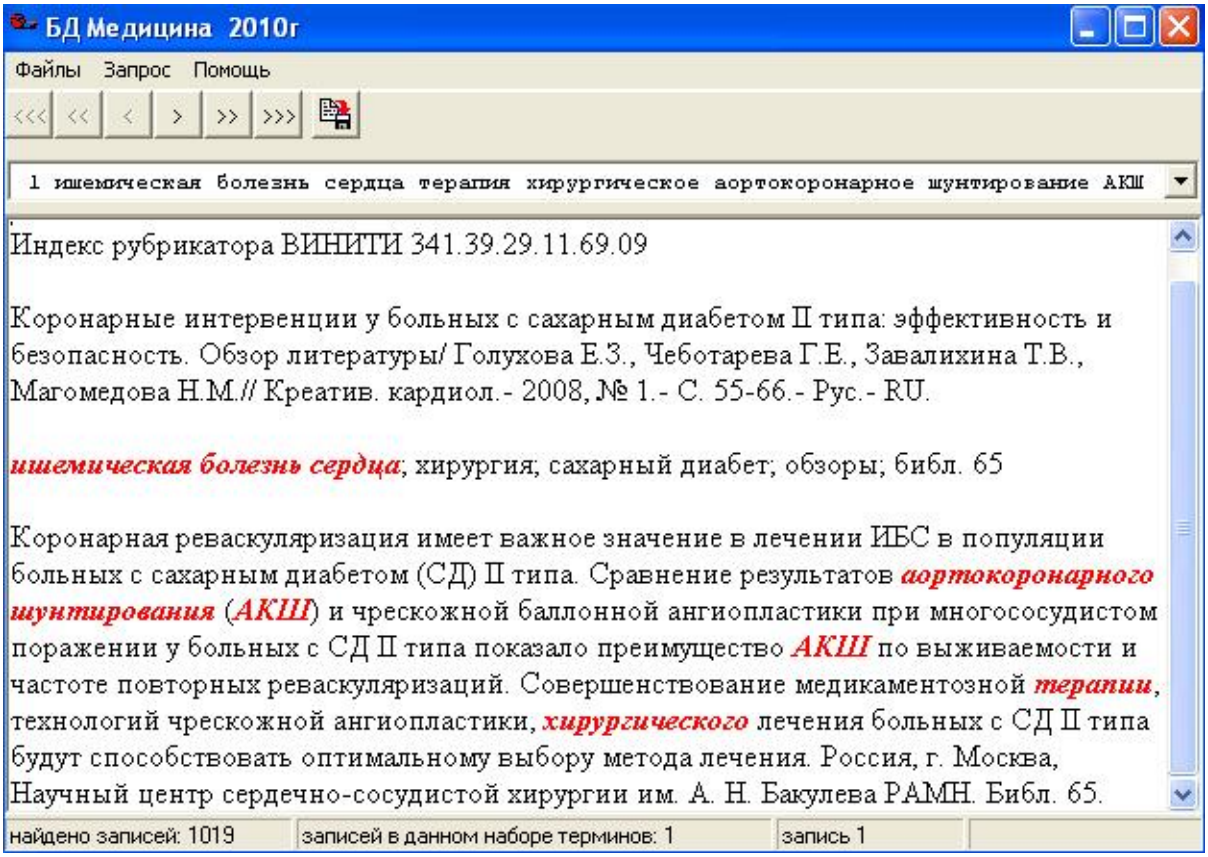

Рис. 3. Результат поиска – документ, соответствующий 1-му рангу

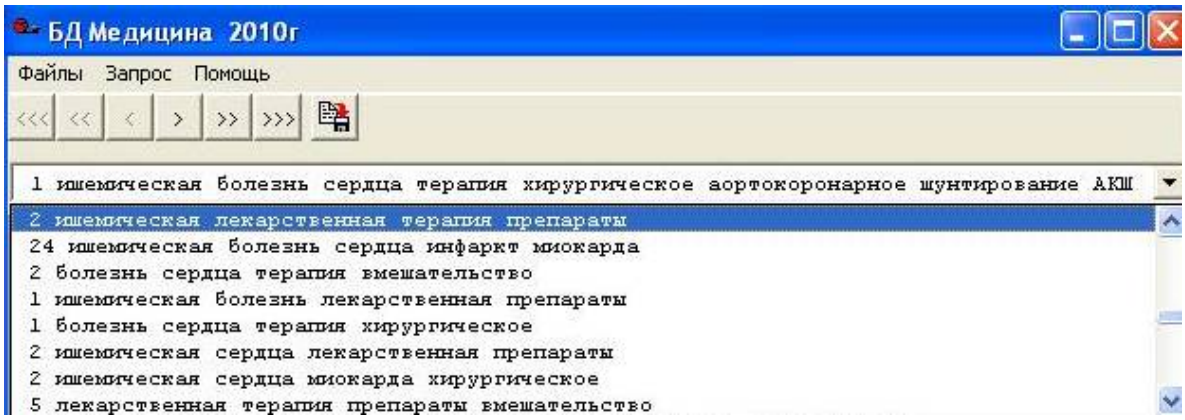

Рис 4. Терминологические сочетания нижних рангов

На рисунке показана выборка из 8 терминологических групп от 4 до 5 терминов в каждой. Количество документов в группах колеблется от 1 до 24. Следует обратить внимание, что в нижних рангах можно встретить набор терминов, который входит составной частью в какие-то наборы верхних рангов (например, 24 документа по терминологическому набору: «ишемическая болезнь сердца инфаркт миокарда». Как это понимать? Это означает, что если указанный набор слов дополняется еще какими-то другими терминами, то документ с таким терминологическим составом попадает в более высокий ранг, а здесь этот документ исключен.

Легкость составления запроса из любого текста нередко приводит к смене стратегии поиска, к быстрому переходу нахождения документов по новым аспектам проблемы в процессе просмотра выдачи. Пользователь, ведущий поиск по запросу 1, обратил внимание на один из документов, в котором в поле ключевых слов обнаружил новый интересующий его аспект. В результате он выгрузил это поле, и с некоторыми собственными дополнениями получился новый запрос 2 - «ишемическая болезнь сердца; инфаркт миокарда; постинфарктная аневризма; лечение аневризмы дуги аорты; стентирование» (пунктуация сохранена из поля ключевых слов и не редактировалась). Общее число слов – 12.

На рис. 5 приведены результаты поиска новых терминологических композиций.

Общее число рангов по запросу 4 – 128. На рис. 5 показано содержание 8-ми верхних рангов. Число слов в этих терминологических комбинациях – от 5 до 7. На экране в первом ранге можно увидеть фактически термины отобранного пользователем для запроса поля ключевых слов, которое и было взято за основу запроса 2. Документ, откуда взяты ключевые слова выдается первым рангом. Нетрудно видеть, что следующие по степени релевантности документы содержат термины, сильно ассоциированные по содержанию с этим документом.

На рис. 6 приведен для примера документ 9-го ранга, содержащий терминологическую цепочку: «сердца инфаркт миокарда постинфарктная аневризма лечение аневризмы». Из этого примера видно, что наблюдается высокая степень терминологической концентрации запросных слов даже в документе конца первой десятки.

Теперь перейдем к процедуре поиска по рубрикам.

На рис. 7 показаны результаты обработки запроса 1 в БД «Рубрикатор БД «Медицина». Показаны 9 верхних рангов, которые охватывают 164 наименований и кодов рубрик, в той или иной степени связанных с текстом запроса. Дадим краткий комментарий результата. По первому рангу, набору терминов: «ишемическая болезнь сердца инфаркт миокарда лечение» имеется одна рубрика. По набору терминов: «ишемическая болезнь сердца лечение» имеется 7 рубрик и т.д.

Приведем конкретные коды и наименования найденных рубрик для верхних рангов:

По рангу -Ишемическая болезнь сердца инфаркт миокарда лечение

*761.31.29.11.11.11 Лекарственное **лечение** острого **инфаркта миокарда**.*

По рангу - Ишемическая болезнь сердца лечение

761.31.29.11.11.02 Общие проблемы

761.31.29.11.11.05 Методы клинических исследований.

761.31.29.11.11.07 Фармакокинетика.

*761.31.29.11.11.13 Лекарственное **лечение** предынфарктного синдрома.*

*761.31.29.11.11.21 Лекарственное **лечение** хронической ИБС.*

*761.31.29.11.11.25 Лекарственное **лечение** стенокардии.*

*761.31.29.11.11.29 Лекарственное **лечение** других форм хронической ИБС.*

По рангу - Ишемическая болезнь сердца

*341.39.29.11.63 Патологическая физиология **ишемической болезни сердца.***

341.39.29.11.63.02 Общие проблемы.

341.39.29.11.63.05 Методы исследования.

*341.39.29.11.63.11 Обмен веществ при **ишемической болезни сердца***

*341.39.29.11.63.19 Факторы риска **ишемической болезни сердца***

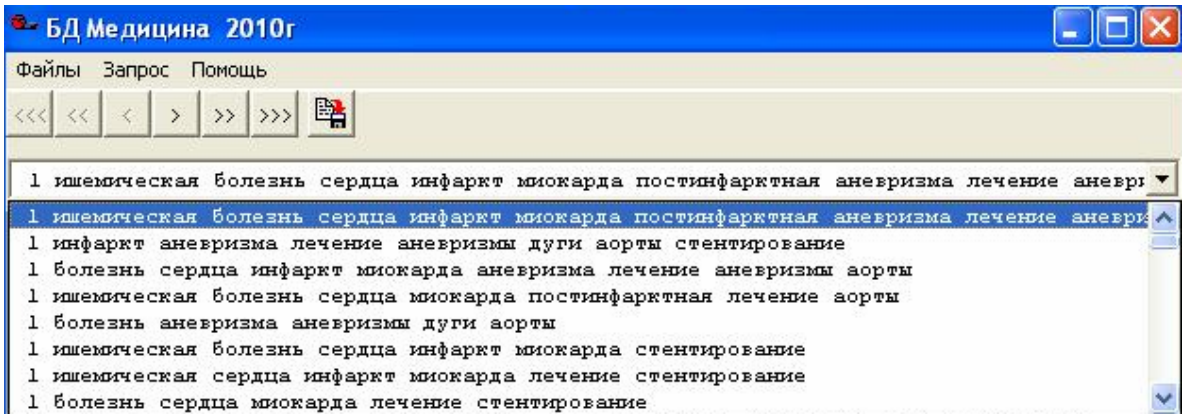

Рис.5. Поиск по модифицированному запросу 2

Рис. 6. Документ, выданный по запросу 2

Рис. 7. Обработка запроса в БД «Рубрикатор БД»

Для иллюстрации был проведен поиск по коду рубрики самого высокого ранга. Для этого при обращении в БД «Медицина» был в качестве запроса введен не набор терминов, а код 761.31.29.11.11.21. В результате по нему было выдано 33 записи, первая из которых следующая:

Запрос: 761.31.29.11.11.21
Запись № 1
Термины: 33 761.31.29.11.11.21

Индекс рубрикатора ВИНИТИ 761.31.29.11.11.21

Современные подходы к коррекции электрической нестабильности миокарда и улучшению прогноза у больных ишемической болезнью сердца/ Радзевич А.Э., Попов В.В.// Кардиоваскуляр. терапия и профилакт.– 2007.– 6, № 3.– С. 106–115.– Рус.; рез. англ.– ISSN 1728-8800.– RU.

ишемическая болезнь сердца; внезапная смерть; адреноблокаторы×β-; амиодарон; имплантируемые дефибрилляторы; больные

Внезапная сердечная смерть (ВСС) остается наиболее частой причиной смерти пациентов с ишемической болезнью сердца (ИБС). В связи с этим изучение механизмов ВСС, внедрение новых методов прогнозирования, и путей предотвращения являются важнейшими задачами современной

кардиологии. В обзоре проанализированы современные подходы к коррекции электрической нестабильности миокарда и улучшению прогноза у больных ИБС. В настоящее время доказанным является факт, что среди всех антиаритмических средств только β-адреноблокаторы и, возможно, амиодарон способны снижать частоту ВСС. Снижение общей смертности и частоты ВСС продемонстрировано на фоне приема ингибиторов ангиотензинпревращающего фермента, аспирина, статинов и антагонистов альдостерона. Недавно опубликованные результаты рандомизированных исследований показали, что имплантируемые кардиовертеры-дефибрилляторы превосходят по эффективности медикаментозную терапию в профилактике ВСС у пациентов с высоким риском. Россия, МГ МСУ, Москва. Библ. 75.

Справедливости ради следует сказать, что пользователи обращаются к поиску по рубрикам крайне редко и только в тех случаях, когда наполнение рубрик им хорошо известно и хочется просто просмотреть соответствующие тематические разделы. Надо признать, что полностью полагаться на классификационные индексы при поиске трудно, так как качество рубрицирования во многом зависит от индексатора, и очень часто многоаспектному документу присваиваются далеко не все рубрикационные индексы. Поэтому, как показала практика, в подавляющем числе случаев пользователь ведет содержательный поиск исключительно по текстовому запросу на естественном языке, игнорируя коды рубрик.

ЗАКЛЮЧЕНИЕ

БД ВИНИТИ «Медицина» распространяется в оболочке с автоматическим созданием терминологических комбинаций из текста запроса уже много лет. Никаких претензий к качеству поиска высказано не было. Пользователи по достоинству оценили предоставляемые поисковые возможности.

Можно сделать вывод, что поисковые системы такого класса на практике оправдали возлагаемые надежды и вполне могут занять соответствующую нишу среди поисковых систем для работы с реферативными базами данных.

СПИСОК ЛИТЕРАТУРЫ

1. Кузнецов Б.А., Солнцева Е.К., Закамская Д.В., Леонтьев А.А., Деревянкин М.В., Быховский Д.В., Ашкинадзе Б.Л. «Диана» – система поиска в текстовых базах данных по запросам на естественном языке // Информационные продукты и технологии. Материалы конференции НТИ-96 – М: ВИНИТИ, 1996.- С. 156-158.
2. Арский Ю.М., Леонтьева Т.М., Шогин А.Н. Создание инфраструктуры системы многоуровневой интеграции разнородных данных // НТИ. Сер. 2. - 1997. - № 2. – С. 18-20.
3. Кузнецов Б.А., Солнцева Е.К., Закамская Д.В., Леонтьев А.А., Деревянкин М.В., Быховский Д.В., Ашкинадзе Б.Л. Интеллектуальный поиск в текстовых БД с помощью системы «Диана» // Информационные ресурсы, интеграция, технологии. Материалы конференции НТИ-97. – М: ВИНИТИ, 1997. - С.131-135.
4. Борисова Л.Ф., Кузнецов Б.А., Солнцева Е.К. Практика применения интеллектуальной поисковой системы третьего поколения «Diana» для расширения круга пользователей БД ВИНИТИ по наукам о жизни // Материалы конференции НТИ-99 – М: ВИНИТИ. – 1999. - С. 64-66.
5. Кузнецов Б.А., Солнцева Е.К., Деревянкин М.В., Закамская Д.В. Обработка запросов на естественном языке – новое качество поиска в БД ВИНИТИ. // НТИ. Сер.2.- 2001.- №11. - С. 31-37

Материал поступил в редакцию 22.10.12.

Сведения об авторе

КУЗНЕЦОВ Борис Антонович – кандидат технических наук, ведущий научный сотрудник ВИНИТИ РАН, Москва

E-mail: ois@viniti.ru

УДК 001.8 : 81’32

Д.А. Ильвовский, М. А. Климушкин

Выявление дубликатов объектов в прикладных онтологиях с помощью методов анализа формальных понятий*

Описывается новый подход к поиску дублей среди объектов онтологии, построенной на избыточных реальных данных, основанный на преобразовании исходной онтологии в формальный контекст и исследовании контекста методами Анализа Формальных Понятий (АФП). Для выявления объектов-дублей вводится новый индекс оценки сходства объектов объема формальных понятий. Рассматриваются результаты работы предложенного подхода на реальных данных, представленных в виде онтологии, построенной на основе подборки информационно-аналитических материалов политической тематики. Производится сравнение введенного индекса с уже существующими индексами и методами выявления сходства объектов.

Ключевые слова: *анализ формальных понятий, прикладные онтологии, дубли, фильтрация понятий*

ВВЕДЕНИЕ

Прикладные онтологии - одна из наиболее универсальных и популярных моделей представления структурированных данных. Распространенный способ построения прикладной онтологии - её автоматическая или полуавтоматическая генерация из неструктурированных данных (как правило, текстов) на основе заранее подготовленного набора правил. Однако при таком способе построения онтологии возникает проблема избыточности данных, поскольку реальные источники информации могут существенно дублировать или перекрывать друг друга: например, во многих статьях может описываться одна и та же компания, человек, место и т.д.

При этом выявление и устранение избыточности непосредственно на этапе построения или дополнения онтологии (например, путем попарного сравнения новых объектов с уже существующими объектами) не слишком эффективно. Во-первых, такой подход существенно увеличивает нагрузку на эксперта, принимающего окончательное решение (особенно эта нагрузка возрастает при частом обновлении данных). Во-вторых, избыточные данные поступают неравномерно, и имеет смысл устранять избыточность не при каждом обновлении онтологии, а через более продолжительные промежутки времени, определяемые особенностями предметной области.

Предлагаемый подход позволяет эффективно выявлять объекты-дубли в исходных данных, представленных в виде онтологии. Разработанный метод может либо автоматически формировать списки объектов-дублей, либо работать в качестве рекомендательной системы для эксперта, одновременно минимизируя нагрузку на него и предоставляя ему четкие и интуитивно понятные рекомендации по определению объектов-дублей.

Выявление объектов-дублей в онтологии осуществляется на основе объединения замкнутых множеств объектов с помощью методов анализа формальных понятий [1].

Постановка задачи

Задача, решаемая в данной работе, заключается в поиске среди объектов онтологии дубликатов, т. е. объектов, описывающих один и тот же объект реального мира. Исходная задача была поставлена аналитиками компании Avicomp. Основное направление - поиск дублей среди объектов, описывающих людей и компании, в онтологиях, строящихся путем автоматической семантической обработки потока новостных текстов. В настоящий момент в компании задача решается методами на основе расстояния Хэмминга и различными дополнительными эвристиками.

На вход алгоритм принимает прикладную онтологию, построенную по новостным текстам. Онтология содержит объекты разных классов, объекты могут быть связаны отношениями, соответствующими их классам. Количество выявленных признаков и связей объекта может сильно варьироваться. Некоторые объекты описывают один и тот же объект реального мира.

На выходе алгоритм должен выдавать списки объектов, которые были идентифицированы им как дубли. При этом алгоритм должен обладать высокой точностью, так как объявление двух различных объектов дублями считается более грубой ошибкой, чем невыявленные дубликаты одного объекта.

* Исследование осуществлено в рамках Программы фундаментальных исследований НИУ-ВШЭ в 2012 г.

Специфика исходной онтологии

Прикладные онтологии, описывающие различные предметные области, в особенности, социальные сети, имеют специфические свойства, которые необходимо учитывать при разработке алгоритма:

1. Онтологии содержат достаточно большое количество объектов (десятки тысяч). Многие объекты имеют редкие или даже уникальные значения признаков, поэтому в онтологии содержится большое количество различных значений признаков.

2. Объекты содержат различное число выявленных признаков и связей (горизонтальных отношений) с другими объектами. Распределение этих чисел не линейное, а имеет "гиперболическую" форму (распределение Ципфа).

3. Также специальным требованием к алгоритму является "неравносильность" ошибок первого и второго рода. Ошибка первого рода (принятие двух дубликатов одного объекта за разные объекты) приводит к тому, что объекты онтологии содержат неполную информацию об объектах реального мира. Ошибка второго рода (объявление двух различных объектов дубликатами одного объекта) приводит к более серьезным последствиям - введению в онтологию неверной информации об объекте.

ОСНОВНЫЕ ОПРЕДЕЛЕНИЯ

Онтологии

Исходные данные в разработанном алгоритме представлены в виде прикладной онтологии предметной области. Введем формальное определение онтологии[2].

Определение 1. *Структура онтологии* - это шестерка вида $O=(C,P,A,H^C,prop,attr)$. C - это множество всех классов онтологии, P - множество всех отношений, заданных на онтологии, A - множество всех атрибутов, H^C - это упорядоченное транзитивное бинарное отношение, называемое таксономией или иерархией классов, $H^C \subseteq C\times C$, запись $(D,B)\in H^C$ означает, что класс D является подклассом B. Функция $prop: P\to C\times C$ задает множество всех отношений между классами, не относящихся к таксономии (их обычно называют горизонтальными). Функция $attr: A\to C$ определяет, какими атрибутами наделен каждый класс.

Определение 2. *Экземпляр онтологии* (или просто онтология) – это 6-ка вида $MD=(O,I,L,inst,instr,instl)$. O – это структура онтологии, I - множество всех идентификаторов, используемых в данном экземпляре (множества I, P, C не пересекаются), L - множество значений атрибутов. Функция $inst: C\to 2^I$ задает множество экземпляров классов, называемых объектами онтологии. Функция $instr: P\to 2^{I\times I}$ задает множество экземпляров отношений. Функция $instl: A\to 2^{I\times L}$ задает значения атрибутов для каждого объекта.

В простейшем случае можно считать, что все атрибуты - это литералы. В более сложном случае каждый атрибут принадлежит определенному домену, а вместо одного множества L рассматривается набор используемых доменов.

Анализ формальных понятий

Анализ формальных понятий (АФП)[1] - это прикладная ветвь теории решеток. Основные сущности АФП были формально описаны Рудольфом Вилле в 1982 г. С точки зрения анализа данных, методы, основанные на анализе формальных понятий, относятся к методам бикластеризации (объектно-признаковой кластеризации). В АФП рассматриваются не кластеры объектов, оторванных от исходного описания, а группы объектов и признаков, сильно связанных друг с другом.

Определение 3. *Формальный контекст* K есть тройка (G,M,I), где G - множество, называемое множеством *объектов*, M - множество, называемое множеством *признаков*, $I\subseteq G\times M$ - бинарное отношение.

Отношение I интерпретируется следующим образом: для $g\in G$, $m\in M$ gIm выполнено тогда и только тогда, когда объект g обладает признаком m.

Определение 4. Для формального контекста $K=(G,M,I)$ и произвольных $A\subseteq G, B\subseteq M$ определена пара отображений:

$A'=\{m\in M \mid gIm\ \forall g\in A\}$, $B'=\{g\in G \mid gIm\ \forall m\in B\}$.

Эти отображения задают *соответствие Галуа* между частично упорядоченными множествами $(2^G,\subseteq)$ и $(2^M,\subseteq)$, а оператор $(\cdot)''$ является *оператором замыкания* на $G\cup M$ - дизъюнктном объединении G и M, т. е. для произвольного $A\subseteq G$ *или* $A\subseteq M$ имеют место следующие соотношения [3]:

1. $A\subseteq A''$ (экстенсивность),
2. $A'''=A'$ (идемпотентность),
3. если $A\subseteq C$, то $A''\subseteq C''$ (изотонность).

Определение 5. *Формальное понятие* формального контекста $K=(G,M,I)$ есть пара *(A,B)*, где $A\subseteq G, B\subseteq M$, $A''=B, B''=A$. Множество A называется *объёмом*, а B - *содержанием* понятия *(A,B)*.

Для двух формальных понятий *(A,B)* и *(C,D)* некоторого контекста *(A,B)* называется \ *подпонятием (C,D)*, если $A\subseteq C$ (эквивалентно $D\subseteq B$). В этом случае *(C,D)* является *надпонятием (A,B)*.

Множество формальных понятий контекста K, упорядоченных по вложению объемов (содержаний), образуют *решетку формальных понятий* $\mathrm{B}(K)$.

ПОИСК ДУБЛЕЙ В ОНТОЛОГИИ

Опишем алгоритм поиска дублей в прикладной онтологии, основанный на методах анализа формальных понятий. Алгоритм делится на два этапа. Первый - преобразование онтологии в формальный контекст, второй - построение множества формальных понятий контекста онтологии и порождение списков дублей, производимое на основе отобранных по специальному критерию формальных понятий.

Преобразование онтологии в формальный контекст

Преобразование онтологии в многозначный контекст

Исходные данные, представленные в виде (экземпляра) онтологии, преобразуются в многозначный контекст, задаваемый следующим образом:

1. Множество **объектов контекста** - это множество O объектов исходной онтологии.

2. Множество **атрибутов контекста** - это множество $M = L \cup C \cup R$, где:

- L - множество атрибутов исходной онтологии,
- C - множество бинарных атрибутов, совпадающее с множеством классов из структуры онтологии,
- R - множество бинарных атрибутов, описывающих связи между объектами онтологии. Каждая связь $(x, y) \in instr(P)(p \in P)$ в онтологии порождает два бинарных признака в контексте: $p(x,\_)$ и $p(\_, y)$. Они соответствуют связи p, идущей от объекта x, и связи p, идущей к объекту y. Таким образом, объект x будет обладать признаком $p(\_, y)$, объект y – признаком $p(x,\_)$.

3. Каждый объект g получает следующие **значения атрибутов**:

- Для атрибута из исходной онтологии $l \in L$:

$$l(g) = \begin{cases} instl(L), если\ attr(l) = inst(g) \\ null\ в\ противном\ случае \end{cases}$$

- Для атрибута $c \in C$:

$$c(g) = \begin{cases} True, если\ (inst(g), c) \in \left(H^C\right)^* \\ False\ в\ противном\ случае \end{cases},$$

где $\left(H^C\right)^*$ - транзитивное рефлексивное замыкание отношения H^C.

- Для атрибута $r \in R$ вида $p(x,\_)$:

$$r(g) = \begin{cases} True, если\ (x, g) \in instr(p) \\ False\ в\ противном\ случае \end{cases}$$

- Для атрибута $r \in R$ вида $p(\_, x)$:

$$r(g) = \begin{cases} True, если\ (g, y) \in instr(p) \\ False\ в\ противном\ случае \end{cases}$$

Иными словами, каждый объект получает значения своих исходных атрибутов, специальное значение *null* для атрибутов, которыми он либо не обладает, либо значение которых для него не известно. Остальные бинарные признаки соответствуют классу объекта (и всем его надклассам) и его связям с другими объектами. Такой подход к преобразованию позволяет учесть всю информацию об объекте, содержащуюся в исходной онтологии.

Преобразование многозначного контекста в формальный контекст

После получения многозначного контекста по онтологии необходимо построить по нему бинарный (формальный) контекст. Каждый признак многозначного контекста преобразовывается в несколько бинарных признаков. Этот процесс называется *шкалированием* [4]. Признаки многозначного контекста из множеств C и R изначально имеют бинарный вид, поэтому в преобразовании не участвуют. Признаки из множества L шкалируются в зависимости от типа признака. Как правило, большая часть признаков описывает неколичественные свойства объекта (например, имя человека, название компании и т.д.). К тому же многие из количественных или просто числовых признаков таковы, что приближенное сходство по этим признакам не говорит о сходстве объектов. К примеру, если два объекта-компании имеют значения признака «Год создания» 2005 и 2006, то близость (но не совпадение) значений этого признака не повышает уверенность в том, что объекты описывают одну и ту же компанию, а скорее дает обратный эффект. Для таких признаков имеет смысл только совпадение значений признака, если же значения различны, то расстояние между ними не имеет значения. Такие признаки шкалируются *номинальной шкалой*, т. е. каждому значению признака соответствует свой бинарный признак. К остальным количественным признакам могут применяться другие типы шкалирования, такие как:

- *Интервальное*: преобразование признака A во множество бинарных признаков вида «$a \le A < b$». При этом интервалы $[a,b)$ могут быть как непересекающимися, так и с перекрытием.
- *Порядковое*: признак A преобразовывается во множество бинарных признаков вида «$A > b$».
- Другие виды шкалирования, которые, по мнению эксперта, могут лучшим образом характеризовать сходство объектов как дублей.

В описанных ниже экспериментах на сгенерированных данных и реальной онтологии использовалось только номинальное шкалирование, однако это не ограничивает общности предложенного подхода.

Построение множества формальных понятий

По построенному формальному контексту строится множество формальных понятий. Существует несколько эффективных методов нахождения формальных понятий. В настоящей работе использовался алгоритм AddIntent [5].

Время работы алгоритма асимптотически равно $O(|L| * |G|^2 * \max(|\{g'\}|, g \in G))$, где $|L|$ - количество формальных понятий контекста, G - множество объектов контекста, $|\{g'\}|$ - число признаков, которыми обладает объект.

Алгоритм довольно эффективен для работы с контекстами, полученными из онтологий, так как такие контексты содержат относительно небольшое число формальных понятий и большая часть объектов имеет всего несколько признаков.

Один из альтернативных подходов построения формальных понятий основан на построении надпонятий уже найденных понятий. Этот подход реализован, например, в алгоритме Замыкай-по-Одному [6]. Его преимущество - возможность остановки алгоритма при достижении определенного размера понятий. Это свойство позволяет порождать не все понятия контекста, а только понятия с небольшим

объемом, так как большие группы объектов скорее всего не являются дубликатами одного и того же реального объекта.

Критерии фильтрации формальных понятий

После построения множества формальных понятий необходимо выделить понятия, объем которых содержит только дубликаты одного объекта.

При подборе критериев были учтены основные свойства, которыми должны обладать эти понятия. Во-первых, **критерий должен принимать большее значение, если, при прочих равных, число признаков, которыми отличаются объекты понятия, будет меньше**. В качестве критерия, характеризующего "разброс" признаков, был использован следующий индекс:

$$I_1(A,B) = \frac{|A||B|}{\sum_{g\in A}\left|\{g\}'\right|}$$

Максимальное значение индекса ($I_1 = 1$) достигается в случае, если ни один из объектов понятия не обладает признаками, не входящими в содержание понятия. Значение индекса стремится к 0 при уменьшении содержания понятия и увеличения у объектов понятия числа признаков вне содержания понятия.

Второе свойство, которым должен обладать критерий - **увеличение значение индекса при увеличении числа общих признаков при прочих равных**. При этом необходимо учитывать частоту признака. Распространенный признак должен делать меньший вклад в значение критерия, чем редкий, так как чем признак более распространен, тем больше шансов, что понятие с данным признаком возникло из-за случайного пересечения признаков.

В результате был разработан индекс, обладающий этим свойством:

$$I_2(A,B) = \sum_{m\in B}\frac{|A|}{\left|\{m\}'\right|}$$

Легко заметить, что появление нового признака в содержании формального понятия (при прочих равных) увеличивает значение индекса. При этом, чем больше объектов в контексте обладают этим признаком, тем меньше изменится значение индекса.

Итоговый критерий *DII* представляет собой комбинацию описанных выше индексов. В настоящей работе использовались следующие способы комбинации.

1. Линейная комбинация описанных индексов:

$$DII_+ = k_1 I_1 + k_2 I_2$$

2. Произведение индексов со степенными коэффициентами:

$$DII_* = I_1^{k_1} * I_2^{k_2}$$

Так как абсолютные значения коэффициентов влияют только на значение порога, а качество фильтрации будет определяться соотношением коэффициентов в формуле критерия, можно сузить семейство критериев без потери оптимального, взяв 1 в качестве значения одного из коэффициентов. Тогда семейства критериев будут представлены в виде формул с одной степенью свободы:

$$DII_+ = I_1 + kI_2,$$

$$DII_* = I_1 * I_2^{\,k}.$$

Следует отметить, что при ранжировании с помощью произведения, формальные понятия, для которых значения одного из индексов равны или близки к нулю, окажутся в конце списка, т. е. этот способ делает обязательным наличие у понятия обоих свойств, описанных выше. Неизвестный коэффициент может определяться с помощью экспертной оценки, так как интерпретация обоих индексов, использованных в критерии, достаточно легка для понимания.

Другой подход к подбору коэффициента заключается в оптимизации критерия качества ранжирования формальных понятий. Алгоритм подбора коэффициента получает на вход множество формальных понятий одного контекста с разметкой: '1' - все объекты понятия - дубли, '0' - понятие содержит различные объекты. Далее рассматривается сетка на положительной вещественной оси, и на ней максимизируется метрика качества ранжирования. Одна из таких метрик - Mean Average Precision (MAP) - более подробно описана далее.

Формирование списка дублей

Списки объектов, которые алгоритм будет выдавать в качестве дублей, формируются на основе объемов формальных понятий с высоким значением критерия. Алгоритм предусматривает два режима работы: автоматическое принятие решения и полуавтоматический режим с привлечением эксперта-аналитика.

В автоматическом режиме подразумевается, что аналитик не участвует в принятии решения о том, являются ли объекты формального контекста дубликатами одного объекта. Алгоритм состоит из двух этапов.

Первый - заключается в фильтрации формальных понятий по порогу на разработанный критерий. На этом шаге могут добавляться различные эвристики, которые трудно учесть с помощью критерия.

В результате фильтрации формируется список формальных понятий с высоким значением критерия.

Второй этап заключается в формировании списков объектов-дублей. Поскольку предполагается, что отношение «быть дублем» (здесь и далее - на множестве объектов) транзитивное, а объекты отобранного формального понятия связаны этим отношением друг с другом, задача формирования списков дублей сводится к поиску компонент связности этого отношения на множестве объектов.

Построение списков дублей осуществляется следующим образом. Сначала строится симметричное отношение *R* «быть дублем». Для каждого формального понятия с объемом $\{g_1,\ldots,g_n\}$ в отношение *R* добавляются пары $(g_1, g_i), g \in 2\ldots n$. Затем по построенному отношению находятся компоненты связности алгоритмом на основе обхода в ширину. Полученные компоненты связности будут соответствовать спискам объектов, выделенных как дубли.

Альтернативный режим работы алгоритма подразумевает, что решение о том, являются ли объекты понятия дублями, принимает аналитик. В этом случае алгоритм предлагает понятия аналитику в порядке уменьшения «уверенности», что его объекты - дубли.

Алгоритм последовательно предлагает аналитику оценить понятия, упорядоченные по убыванию значений критерия *DII*. При этом списки дублей формируются по мере получения ответов аналитика.

Перед тем как предлагать понятие аналитику для оценки, находятся все списки объектов-дублей, имеющие пересечение с объемом понятия. Если объем уже вкладывается в один из списков дублей, то понятие не предлагается аналитику для оценки, В противном случае, аналитику предлагается оценить понятие. Если аналитик дает положительный ответ по формальному понятию (считает, что объекты этого понятия - дубли), то алгоритм редактирует списки дублей:

- Если объем понятия не имеет ни одного пересечения со списками дублей, он добавляется как новый список.
- Если объем понятия имеет одно пересечение, то объекты из объема добавляются к списку, с которым есть пересечение.
- Если объем понятия имеет пересечения с несколькими списками, то эти списки объединяются в один и пополняются объектами из объема.

Таким образом, алгоритм получает список дублей, соответствующий текущей разметке аналитика. Соответственно, аналитик может на каждом шаге остановить процесс оценки понятий и получить сформированные списки дублей.

Ранжирование формальных понятий по индексу *DII* позволяет давать эксперту в первую очередь понятия, которые с большей вероятностью содержат дубли, что значительно упрощает работу эксперту по сравнению с оценкой понятий в произвольном порядке.

ЭКСПЕРИМЕНТАЛЬНЫЕ ИССЛЕДОВАНИЯ

Эксперименты на случайных контекстах

Чтобы получить статистические оценки качества разработанного алгоритма, основные эксперименты проводились на искусственно сгенерированных данных с заранее известными дублями. Это позволило оценить качество метода на большом количестве входных данных и провести количественное сравнение разработанного метода с наиболее распространенными альтернативными подходами. Наряду с этим, при генерации данных также учитывались особенности прикладной онтологии, что позволяет экстраполировать полученные результаты на реальные данные.

Генерация входных данных

Для оценки качества метода использовались различные метрики качества на искусственно сгенерированных контекстах. При этом генерируемые формальные контексты обладали свойствами контекстов, получаемых из прикладных онтологий.

Во-первых, генерируемые контексты должны содержать большое количество объектов и признаков. Подразумевается, что количество объектов будет измеряться десятками тысяч. При этом количество бинарных признаков сравнимо с количеством объектов, так как многие объекты содержат уникальные или редкие признаки. При этом каждый объект обладает относительно небольшим количеством признаков. Их число обычно не превышает нескольких десятков. Поэтому контекст сильно разрежен, и, несмотря на большой размер контекста, число формальных понятий в нем относительно небольшое.

Во-вторых, количество признаков у объектов достаточно сильно варьируется и, как правило, удовлетворяет закону Мандельброта, т. е. количество признаков примерно обратно пропорционально рангу объекта среди объектов, упорядоченных по количеству признаков у них.

Третье свойство, которое учтено при генерации контекста, - это неравномерное распределение частот признаков. Как правило, частота признака обратно пропорциональна его рангу в последовательности, упорядоченной по частоте появления признака у объектов контекста.

После генерации уникальных объектов генерировался входной контекст. Для каждого объекта создавался объект в контексте следующим образом: каждый признак объекта с определенной вероятностью добавлялся во множество признаков объекта в контексте. Для некоторых исходных объектов создавалось несколько объектов подобным образом. Полученные объекты считались дублями одного объекта.

Сравнительный анализ методов поиска дублей

В ходе исследования был проведен сравнительный анализ разработанного метода с несколькими из наиболее распространенных методов, способных решать данную задачу. Использовавшиеся методы основаны на анализе формальных понятий и попарном сравнении объектов контекста. В качестве методов попарного сравнения объектов были рассмотрены методы на основе *расстояния Хэмминга* и *абсолютного сходства*. Эти методы могут быть применены к многозначному контексту или напрямую к онтологии, но для простоты сравнения они будут описаны в применении к бинарному контексту.

Метод на основе экстенсиональной устойчивости понятия

Устойчивость формального понятия была впервые введена С. О. Кузнецовым в 1990 г. [7]. Позднее в работе [8] было предложено различать два типа устойчивости: экстенсиональную и интенсиональную. В этой работе была выбрана экстенсиональная устойчивость, так как предполагается, что объекты, являющиеся дублями, должны быть сильно связаны большим количеством признаков и иметь небольшое количество отдельных признаков, соответственно, формальное понятие, которое они образуют, должно быть устойчиво к удалению отдельных признаков.

Алгоритм поиска дублей аналогичен основному методу: из множества формальных понятий выделяются наиболее (экстенсионально) устойчивые понятия. Затем предполагается, что объекты из объема устойчивого формального понятия являются дубликатами одного и того же объекта. По множеству выбранных формальных понятий строится отношение «быть дублями» *R*.

Затем находятся компоненты связности данного отношения. Полученные компоненты выдаются на вход в качестве списков объектов-дублей.

Метод на основе меры абсолютного сходства

Данный метод основан на попарном сравнении объектов. Предполагается, что объекты онтологии, являющиеся дублями одного и того же объекта имеют большое количество общих признаков. Поэтому в качестве критерия близости объектов используется количество их общих признаков. Индикатор, основанный на данной мере, представляет собой порог на количество общих признаков.

Алгоритм получает на вход квадратную матрицу близости $A: A[i][j] = k \Leftrightarrow$ *i*-й и *j*-й объекты имеют *k* общих бинарных признаков, а также порог $t(N)$.

По матрице *A* и порогу строится матрица смежности $B: A[i][j] > t \Rightarrow B[i][j] = 1$. Матрица смежности (аналогично входной матрице) является симметричной и описывает некоторое отношение близости *R*. Исходя из того, что отношение «быть дублем» является отношением эквивалентности и обладает свойством транзитивности, по полученному отношению *R* строится его транзитивное замыкание R^*. Классы эквивалентности в R^* соответствуют группам объектов, являющихся дублями одного объекта. Тот же результат можно получить, выделив все компоненты связности отношения *R*.

Асимптотическая сложность алгоритма по времени - $O\left(n^2 * m\right)$, где *n* - количество объектов в формальном контексте, *m* - количество признаков.

Метод на основе расстояния Хэмминга

Алгоритм поиска дублей основан на попарном сравнении объектов. В качестве метрики близости используется расстояние Хэмминга. Вначале составляется квадратная матрица расстояний между объектами. Затем, по построенной матрице *A* и заданному порогу $t(N)$, строится матрица *B* отношения «быть дублями» $R: A[i][j] > t \Rightarrow B[i][j] = 1, (x_i, x_j) \in R$. Полученное отношение будет симметричным и рефлексивным. По данному отношению находятся компоненты связности. Объекты, попавшие в одну компоненту связности, считаются дублями одного объекта.

Асимптотическая сложность алгоритма аналогична сложности алгоритма на основе абсолютного сходства - $O\left(n^2 * m\right)$, где *n* - количество объектов в формальном контексте, *m* - количество признаков.

Метрики для сравнения методов выделения дублей

Для проведения сравнительного анализа использовалось несколько метрик качества метода: полнота, точность, среднее значение полноты алгоритма при 100% значении точности, MAP. В качестве основных метрик использовались **полнота** и **точность** алгоритма.

Для того чтобы корректно определить полноту и точность, рассмотрим задачу поиска дублей как задачу удаления из множества объектов онтологии объектов-дублей. Тогда выделенную алгоритмом группу объектов будем интерпретировать как удаление из онтологии всех объектов группы за исключением одного. Таким образом, мы определяем полноту и точность алгоритма:

$$Precision = \frac{\left|D_{dub} \cap D_{del}\right|}{\left|D_{del}\right|},$$

$$Recall = \frac{\left|D_{dub} \cap D_{del}\right|}{\left|D_{dub}\right|}.$$

Здесь D_{dub} - количество дублей (если есть *n* объектов онтологии, которые по факту являются одним и тем же объектом, считается, что среди них есть $n-1$ дубль), D_{del} - количество удаляемых объектов (если алгоритм выдал группу из *n* объектов, считается, что мы удаляем $n-1$ объект; причем если среди них есть *k* различных по построению объектов, то считается что $k-1$ объект мы удалили неправильно).

Так как качество характеризуется комбинацией этих показателей, а все сравниваемые алгоритмы имели дополнительные параметры (пороги), то рассматривались зависимости полноты алгоритма от точности, путем прогона алгоритмов с различными входными параметрами.

Также для оценки использовалась метрика качества ранжирования MAP (Mean Average Precision):

$$Map(K) = \frac{\sum_{i=1}^{|K|} AveP(K_i)}{|K|},$$

$$AveP(k) = \frac{\sum_{c \in C_k} (P(c))}{|C_k|},$$

где *K* - множество контекстов, C_k - множество релевантных формальных понятий контекста *k*, $P(c)$ - доля релевантных понятий среди всех понятий, имеющих ранг не ниже, чем у понятия *c*.

Результаты

Для оценки нового метода сначала были подобраны оптимальные коэффициенты для индекса. Коэффициент подбирался по одному из сгенерированных контекстов. Бралась сетка на положительной вещественной оси, и на ней максимизировался индекс MAP.

Таким образом, были получены коэффициенты для использовавшихся вариантов индекса *DII*:

$DII_+ = I_1 + 0.25 I_2$

$DII_* = I_1 * I_2^{0.18}$

Алгоритм с данным индексом сравнивался с альтернативными методами поиска дублей. Для построения зависимости точности алгоритма от его полноты для каждого метода задавалось несколько десятков различных порогов, затем рассчитывались полнота и точность алгоритма при каждом пороге. Эти показатели рассчитывались для нескольких сгенерированных контекстов, далее определялось среднее значение полноты и точности для каждого порога. Полученные соотношения позволяют сравнить использовавшиеся алгоритмы (рис.1, 2).

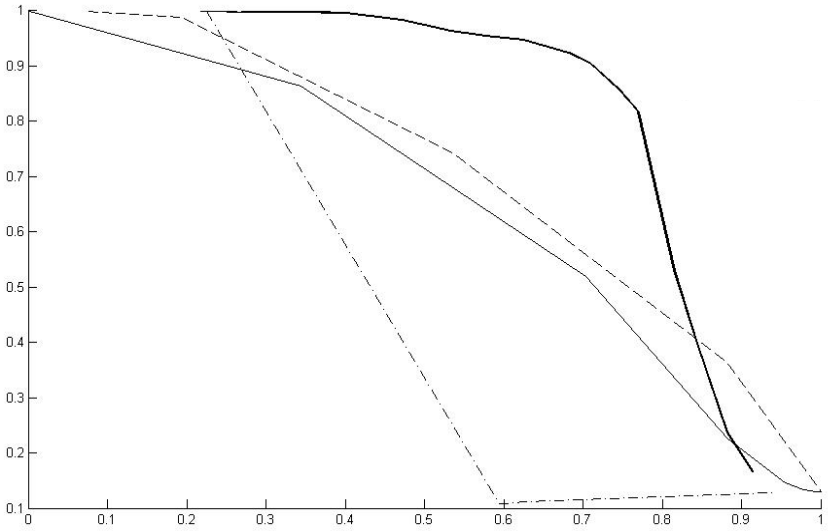

Рис. 1. Зависимость точности алгоритмов от полноты:

Разработанный индекс
Расстояние Хэлминга
Экстенсиональная устойчивость
Абсолютное сходство

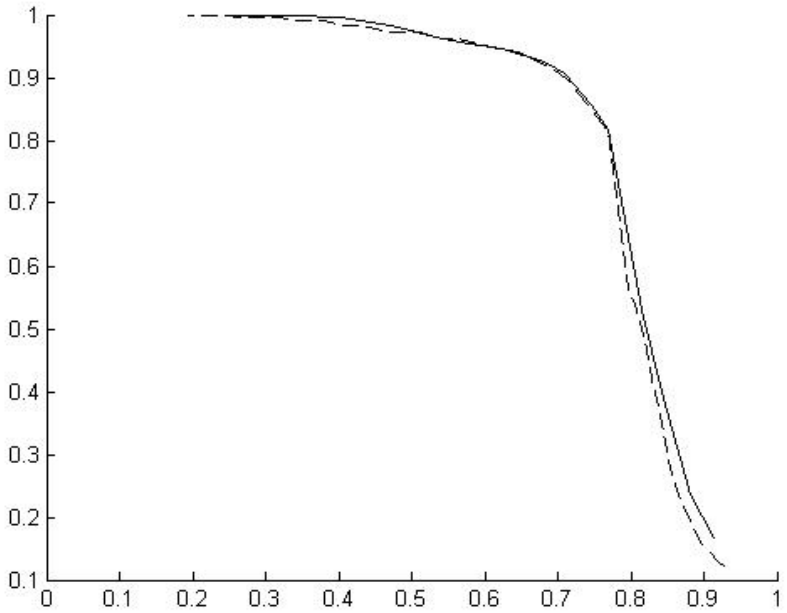

Рис. 2. Зависимость точности от полноты для двух вариантов нового индекса *DII:*

DII+
DII*

Метод на основе экстенсиональной устойчивости показывает хорошие результаты при высоком пороге на индекс. При пороге больше 0.5 отбираются только формальные понятия, содержащие дубли. При пороге ниже 0.5 точность алгоритма падает в среднем до 10%, так как большое количество формальных понятий с устойчивостью 0.5 - однопризнаковые понятия, которые не характеризуют объекты-дубли.

Алгоритм поиска дублей с использованием расстояния Хэмминга показал сравнительно низкие результаты. Так как расстояние может быть только целым положительным числом, снижение порога на 1 добавляет группу новых связей. При достаточно низком пороге точность близка к 100%, но даже среди объектов, имеющих одинаковый набор признаков, могут быть пары, не являющиеся дублями. Как пра-

вило, это объекты с одним-двумя распространенными признаками. Но расстояние Хэмминга не учитывает количество общих признаков, а только различия в признаках.

Алгоритм на основе абсолютного сходства объектов оказался наиболее эффективным среди рассмотренных альтернативных алгоритмов. В большинстве случаев большое количество общих признаков у пары объектов говорит о том, что объекты являются дублями. Недостаток индекса в том, что он не учитывает различия объектов. К тому же некоторые признаки встречаются у большого количества объектов, и наличие их среди общих признаков не дает большого вклада в уверенность, что объекты являются дублями.

Алгоритм на основе нового индекса (с использованием как одного, так и другого варианта комбинации) показал более высокие результаты, чем рассмотренные альтернативы. Основной отличительной особенностью метода является небольшое падение точности алгоритма (до 90%) при росте полноты вплоть до 70%. По остальным метрикам данный метод показал высокие результаты. Результаты для DII_+ и DII_* оказались весьма схожими. Отличием DII_* стало менее стабильное поведение: иногда, ошибаясь при большом пороге, в ряде случаев алгоритм не делал ошибок при малых порогах, выделяя при этом 42% дублей.

По показателю максимальной полноты без потери точности наиболее эффективным оказался метод на основе индекса устойчивости, который позволяет, поставив порог на индекс равным 0,5, выделять в среднем 22,44% дублей. При этом индекс DII_+ «отстал» по этому показателю незначительно, в отличие от методов попарного сравнения. Методы на основе попарного сравнения показали значительно более низкие результаты по данной метрике (табл. 1).

Таблица 1

Максимальная полнота алгоритмов при максимальной точности

Алгоритм	Максимальная полнота при точности 100% (на экспериментальных данных)
Абсолютное расстояние	6,22%
Расстояние Хэмминга	0,56%
Индекс устойчивости	22,44%
Новый индекс DII_+	21,78%
Новый индекс DII_*	9,49%

При сравнении методов на основе индекса экстенсиональной устойчивости и вариантов нового индекса DII_+ и DII_* по мере *MAP* очевидное преимущество имеет новый индекс (табл. 2).

Таблица 2

Mean Average Precision

Алгоритм	MAP
Индекс устойчивости	0,4992
Новый индекс DII_+	0,9352
Новый индекс DII_*	0,9382

Для каждого метода был подобран оптимальный порог, при котором алгоритм имеет оптимальную полноту при минимальных потерях точности (табл. 3).

Таблица 3

Оптимальные пороги для методов поиска дублей и качество поиска

Алгоритм	Порог в алгоритме	Полнота	Точность
Абсолютное расстояние	3,5	19,35%	98,82%
Расстояние Хэмминга	0,5	34,37%	86,32%
Индекс устойчивости	0,5	22,44%	100%
Новый индекс DII_+	1,15	40,09%	99,58%
Новый индекс DII_*	0,9	31,8%	99,55%

Эксперименты на прикладной онтологии

Прикладная онтология

Онтология, на которой был апробирован предложенный алгоритм, была построена компанией Avicomp. Онтология строилась и расширялась автоматически путем семантической обработки потока новостных сайтов программным средством OntosMiner[9].

По обработанному документу строится небольшая онтология с объектами и связями, выделенными в тексте. Затем онтология документа сливается с основной онтологией. Во время слияния происходит поиск дублей среди объектов основной онтологии и онтологии документа методом на основе расстояния Хэмминга с дополнительными эвристиками. При этом часто объекты, являющиеся дублями, не идентифицируются как один объект, и в результате в онтологии возникает большое количество дублирующих друг друга объектов.

Анализируемая онтология была построена по новостным документам политической направленности. Она содержит 12006 объектов различных классов. Объекты имеют различное количество признаков и связей с другими объектами. Количество признаков и связей с другими объектами распределено по закону Ципфа.

В анализируемой онтологии был проведен поиск дублей среди объектов классов «Персона» и «Компания». Таких объектов в онтологии 9821. Признаки формального контекста строились с использованием всех объектов и связей в онтологии.

Обсуждение результатов

Для получения аккуратных оценок полноты и точности алгоритмов необходимо иметь информацию о том, какие объекты являются дубликатами одного и того же объекта. Данную информацию можно получить лишь с помощью экспертной оценки коллекции обработанных документов.

К сожалению, из-за отсутствия документов, на основе которых строилась онтология, точную оценку качества метрики получить не представляется возможным. Но ее можно оценить примерно с помощью экспертной оценки сформированных списков дублей.

Алгоритм на основе индекса *DII* (использовался вариант DII_+) выделил около 900 групп объектов. В результате экспертной оценки было выявлено несколько ошибок в результате. Алгоритм объединил объекты с разными именами/фамилиями, которые имели большое количество общих связей и признаков (партнеры, коллеги). Ошибка возникает из-за того, что алгоритм не учитывает, что разные значения некоторых признаков свидетельствуют о том, что объекты не являются дублями. Поэтому в алгоритме было добавлено довольно простое дополнительное ограничение – отбрасывать понятия с объектами, у которых разные имена или фамилии. Стоит отметить, что подобное ограничение не распространяется на все признаки, так как они могут меняться со временем.

Далее метод использовался с дополнительными условиями. Алгоритм выделил 905 групп объектов. Размеры групп варьируются от 2 до 41 объекта. Наиболее крупные группы, выделенные алгоритмом, описывают Нетаньяху Биньямина (41 объект), Юлию Тимошенко (35 объектов), Владимира Путина (34 объекта), Дмитрия Медведева (33 объекта), Стива Джобса (31 объект) и др. Но основная часть выделенных групп состоит из 2-3 объектов.

В результате оценки результатов работы алгоритма были получены примерные оценки точности алгоритма. В 98% групп с высокой вероятностью можно утверждать, что объединенные в них объекты являются дублями. Часто это следует из наличия у объектов таких общих признаков, как *фамилия* и *имя*. Также нередко встречаются группы, где данные признаки не являются общими, но по другим признакам и связям объекты объединяются в одну группу. Например, в онтологии было выявлено семь объектов, описывающих Ксению Собчак. При этом часть объектов имели признаки “Фамилия:Собчак”, “Имя:Ксения”, другая часть имели признаки “Имя:Ксения”, “Отчество:Анатольевна”. Несмотря на то что у объектов всего один общий признак (имя), за счет общих связей было выявлено, что это один и тот же объект. Аналогичная ситуация с объединением объекта с признаком “Имя:Усама” и объекта с признаком “Фамилия:Ладен”.

Стоит также отметить, что наличие весов у признаков в индексе I_2 позволяет выделять большие группы объектов, описывающие Путина, Тимошенко, Медведева и т.д. Особенности данных объектов в том, что каждый и них имеет большое количество собственных признаков и связей, поэтому расстояние Хэмминга между этими объектами значительное, а число общих признаков - небольшое. Поэтому рассмотренные альтернативы, основанные на попарном сравнении объектов, плохо работают на данных объектах. При этом формальное понятие, содержание которого состоит из имени и фамилии персоны, имеет высокое значение индекса *DII*, так как объекты понятия составляют значительную часть объектов, обладающих данными признаками. При этом его подпонятия имеют более низкое значение индекса *DII*.

ЗАКЛЮЧЕНИЕ

В настоящей работе предложен алгоритм поиска дублей в онтологии, основанный на методах анализа формальных понятий. Метод состоит из двух основных этапов: преобразование онтологии в формальный контекст и формирование списков дублей объектов с помощью отбора формальных понятий. Помимо метода решения задачи, разработан также индекс, позволяющий отбирать формальные понятия, описывающие объекты-дубли.

Рассмотрены альтернативные методы решения поставленной задачи, основанные на попарном сравнении объектов. Также был рассмотрен альтернативный критерий отбора формальных понятий, основанный на применении индекса экстенсиональной устойчивости.

Был произведен сравнительный анализ разработанного метода с его альтернативами и выявлены основные свойства всех методов. Сравнение методов производилось на искусственно сгенерированных данных. При генерации были учтены все выявленные свойства реальной онтологии, что позволяет результаты, полученные на сгенерированных данных, перенести на реальные онтологии. Для сравнения использовались основные метрики качества классификаторов (полнота, точность) и методы ранжирования (*MAP*).

Эксперименты на случайных данных продемонстрировали преимущества нового метода. Эксперименты на реальных данных показали, что разработанные метод и критерий для фильтрации понятий довольно эффективны. На реальной онтологии алгоритм допустил всего несколько грубых ошибок, но при добавлении простейших дополнительных условий при отборе понятий алгоритм показывает высокую точность. Экспертная оценка сформированных групп объектов не выявила явных ошибок.

Возможным вариантом развития исследования является получение численных оценок полноты метода на реальных данных. Для этого необходимо наличие исходных текстов и экспертного анализа. Необходимо провести выявление объектов из разных текстов, описывающих один и тот же объект реального мира, так как в некоторых случаях дубли в онтологии невозможно с абсолютной уверенностью выделить даже с помощью экспертной оценки.

СПИСОК ЛИТЕРАТУРЫ

1. Ganter B., Wille R. Formal Concept Analysis: Mathematical Foundations. - Berlin: Springer, 1999.
2. Maedche A., Zacharias V. Clustering Ontology-based Metadata in the Semantic Web // Proc. of 6th European Conference on Principles of Data Mining and Knowledge Discovery. - 2002. - P. 348 – 360.
3. Биркгоф Г. Теория решеток. — М.: Наука, 1989.
4. Prediger S. Logical scaling in formal concept analysis // ICCS, Lecture Notes in Computer Science. – 1997. - Vol. 1257. Springer. - P. 332-341.
5. Merwe D., Obiedkov S., Kourie D. AddIntent: a new incremental algorithm for constructing concept lattices // Lecture Notes in Computer Science. - 2004. – Vol. 2961. Springer. — P. 205 – 206.
6. Кузнецов С.О. Быстрый алгоритм построения всех пересечений объектов из конечной полурешетки. // НТИ. Сер. 2. – 1993. - № 1. - С.17-20.
7. Кузнецов С.О. Устойчивость как оценка обоснованности гипотез, получаемых на основе операционального сходства // НТИ. Сер.2 - 1990. - № 12. - С.21-29.
8. Roth C., Obiedkov S., Kourie D. On Succinct Representation of Knowledge Community Taxonomies with Formal Concept Analysis // Intl Journal of Foundations of Computer Science. – 2008. – Vol. 19. - P. 383-404.
9. Программное средство Ontos Miner: URL - http://www.ontos.com/?page_id=630.
10. Rudolf Wille. Restructuring lattice theory: an approach based on hierarchies of concepts. - Ordered Sets: Dordrecht/Boston, Reidel. – 1982. - P. 445-470.
11. Евтушенко С.А. Система анализа данных «Concept Explorer» // Труды 7-ой Национальной Конференции по Искусственному Интеллекту (КИИ-2000). – М., 2000. - С.127-134.
12. Formal Concept Analysis / eds. R. Medina, S.A. Obiedkov // 6th International Conference, ICFCA 2008. - Montreal, Canada: Springer. – 2008.
13. Newman M.E.J., Strogatz S., Watts D. Random graphs with arbitrary degree distributions and their applications // Phys. Rev. E 64. - 2001.
14. Kuznetsov S., Obiedkov S., Roth C. Reducing the representation complexity of lattice-based taxonomies // 15th Intl Conf on Conceptual Structures, ICCS 2007. - Sheffield, UK. - LNCS/LNAI. - Springer.- 2007. - Vol. 4604.
15. Климушкин М., Четвериков Д. Исследование американских политических блогов на основе анализа формальных понятий // ЗОНТ-09. – Новосибирск: РИЦ прайс-курьер, 2009.
16. Klimushkin M., Chetverikov D., Novokreshchenova A. Formal Concept Analysis of the US Blogosphere during the 2008 Presidential Campaign // 9th international session of the HSE "Baltic Practice". – Belgium, 2009.
17. Klimushkin M.A., Obiedkov S.A., Roth C. Approaches to the selection of relevant concepts in the case of noisy data. // 8th International Conference, ICFCA2010. – Morocco: Springer, 2010.

Материал поступил в редакцию 17.10.12.

Сведения об авторах

ИЛЬВОВСКИЙ Дмитрий Алексеевич – аспирант Национального исследовательского университета – Высшая школа экономики (НИУ ВШЭ), Москва.

E-mail: dilv_ru@yahoo.com

КЛИМУШКИН Михаил Александрович – студент магистратуры Национального исследовательского университета – Высшая школа экономики (НИУ ВШЭ), Москва.

E-mail: klim.mikhail@gmail.com

ИНФОРМАЦИОННЫЙ АНАЛИЗ

УДК 004.934.5.8:811.161.1'322.6

А. А. Карпов

Машинный синтез русской дактильной речи по тексту*

Представлен анализ особенностей и характеристик дактильной речи, которая является неотъемлемым компонентом разговорного жестового языка глухих людей. Приводится описание реализации многомодальной компьютерной системы для аудиовизуального синтеза русской звучащей и дактильной жестовой речи по произвольному русскоязычному тексту, которая ориентируется на применение в ассистивных технологиях и универсальных диалоговых человеко-машинных системах, предназначенных как для слышащих людей, так и для лиц с ограниченными возможностями по слуху.

Ключевые слова: *русский жестовый язык, дактильная речь, машинный синтез, многомодальные системы, аудиовизуальная речь, объединение информации*

1. ВВЕДЕНИЕ

Русской дактильной речью пользуются в основном неслышащие люди и имеющие проблемы со слухом, которых только на территории России насчитывается несколько сотен тысяч человек. Дактильная речь всегда сопутствует разговорному жестовому языку, который является основным средством коммуникации в среде глухих людей. Дактильной речью (дактилологией, англ.: *«fingerspelling»*) называется система жестов рук для передачи графем (букв) естественных языков, применяющаяся в тех случаях, когда общение посредством голоса или письма затруднено или невозможно, а также когда жестовый язык не позволяет передать некоторое слово или высказывание, например, когда глухой человек или его собеседник не знает правильного жеста для соответствующего слова, либо его попросту не существует. Дактильная речь по своей структуре является довольно близким аналогом письменности с тем отличием, что графические символы национальных азбук воспроизводятся при помощи ручных жестов.

Под дактилологией может пониматься как одна из пространственных визуально-кинетических систем общения при помощи жестов, так и непосредственно само общение с использованием этой системы. Во избежание недоразумений в настоящей статье в первом значении используется термин «дактилология» или «дактильная азбука», а во втором — «дактильная речь».

Практически в каждой стране мира существует собственный язык жестов глухих, всего жестовых языков насчитывается более 130 среди 6900 разговорных языков всех типов [1]. Среди наиболее авторитетных современных исследователей русского жестового языка необходимо отметить Г.Л. Зайцеву [2], И.Ф. Гейльмана [3], Л.С. Димскис [4], Р.Н. Фрадкину [5], А.Л. Воскресенского [6], М.Г. Грифа [7], А.А. Кибрика [8], Р.М. Фрумкину [9] и ряд других. В отличие от исследований по жестовым языкам мира, библиография по которым насчитывает нескольких десятков тысяч, исследований непосредственно по дактильной речи совсем немного. В первую очередь, это связано с вторичностью дактильной азбуки по отношению к письменности национальных языков. Лингвистическая ценность дактилологии не настолько велика по сравнению с жестовыми языками, которые рассматриваются практически на равных основаниях со звучащими языками, выступая объектом лингвистических исследований [10–12]. Наиболее детальным описанием русской дактилологии является монография Г.Л. Зайцевой [2] и многотомный словарь И.Ф. Гейльмана [13], в котором один из томов посвящен исключительно дактилологии. Зачастую в учебных пособиях, предназначенных специально для неслышащих людей и сурдопедагогов, раздел по дактилологии ограничивается только алфавитом и краткими примерами применения. Тем не менее, дактилология является неотъемлемой составляющей специфических средств общения слабослышащих людей и должна рассматриваться как языковая система, дополняющая систему разговорного жестового языка, со своими специфическими особенностями структуры и функционирования.

* Данное исследование поддержано Советом по грантам Президента РФ (проект № МК-1880.2012.8), Министерством образования и науки РФ в рамках ФЦП «Исследования и разработки» (госконтракт № 11.519.11.4025) и Комитет по науке и высшей школе Правительства Санкт-Петербурга.

2. ОСНОВНЫЕ ПРИНЦИПЫ И ОСОБЕННОСТИ ПРИМЕНЕНИЯ ДАКТИЛЬНОЙ РЕЧИ

2.1. Типы дактильных алфавитов

Доподлинно неизвестно, кем впервые была применена или описана дактильная азбука, но в XVII в. в Европе она уже использовалась глухими людьми для межличностного общения. С уверенностью можно говорить лишь о том, что дактильная азбука появилась в одном из сообществ, обладающих развитой системой письменности, подразумевающей умение выделять сравнительно небольшое количество элементов словесной речи и, как следствие, довольно высокий уровень лингвистического абстрагирования [14].

С точки зрения лексической структуры не существует большой разницы между дактильной азбукой и жестовым языком, в котором один жест обозначает лексему. В обоих случаях речь идет о знаковой системе, состоящей из лингвистических знаков. Однако как жестовый язык, так и дактильная речь значительно отличаются от словесной речи в плане выражения своих знаков. Лингвистический знак обладает свойством конвенциональности, иными словами, он достаточно произволен по своей форме, не имея какой-либо ассоциативной связи с означаемым. Так, орфографическое слово СТОЛ не имеет ничего общего ни с понятием стола, ни с множеством столов, существующих в окружающей действительности. В жестовых же языках, наоборот, очень многие жесты мотивированы и ассоциированы с реальным объектом. Например, изображая жест ДЕЖУРНЫЙ (в школе) движения руки имитируют позванивание колокольчиком, т.е. одним из атрибутов дежурного по классу в школе, а жест ДОМ показывают сложенные «крышечкой» руки, изображающие крышу дома. Точно такой же принцип часто соблюдается и в дактильной азбуке, когда жесты воспроизводят форму букв соответствующего письменного алфавита русского языка (например, З, Л, М, О, П) [4].

Однако не все дактильные алфавиты идут по пути «копирования» графем. В некоторых ручных азбуках мира дактилемы не имеют ничего общего по своей форме с графемами. Проблема состоит в том, что графемы некоторых письменностей чрезвычайно сложно или даже невозможно изобразить при помощи пальцев. Согласно общепринятой классификации [15], различают следующие дактильные алфавиты:

Классификация основных типов дактильных алфавитов

По составу	По способу образования	По принципу обозначения
Одноручные	Копирующие	Буквенные
Двуручные	Вариантные	Слоговые
-	-	Совмещенные

Одноручные азбуки отличаются от двуручных, как это понятно из названия, тем что, одна или две руки используются при образовании жестов азбуки. Признак копирующий/вариантный определяет, воспроизводит ли форма кисти руки графему, или же положение руки и пальцев достаточно условно. Наконец, признак буквенные/слоговые/совмещенные тесно связан с принципами конкретной системы письменности. Помимо прочего, принято выделять фонетические письменности и слоговые. Если в письменностях первого типа соблюдается (хотя и не всегда строго) принцип «одна графема — одна фонема», то в слоговых одной графеме соответствует один слог. Русская дактильная азбука относится, соответственно, к одноручной, копирующей и буквенной.

Дактильная речь, несмотря на специфический характер передаваемой информации (графемы, а не лексемы), активно применяется в общении глухих. В то время как произнесение слов по фонемам делало бы голосовое общение неудобным [16], дактильная речь успешно справляется с коммуникативной функцией естественного языка. Можно выделить следующие аспекты функционирования и применения дактильной речи: базовое средство общения, вспомогательное средство общения и педагогический аспект. Рассмотрим эти аспекты и сферы использования дактилологии более подробно.

2.2. Аспекты применения дактильной речи

Дактильная речь может использоваться в коммуникативной ситуации, когда хотя бы один из ее участников является глухим. Иными словами, дактильная речь употребляется как слабослышащими в межличностном общении, так и при общении глухих со слышащими и наоборот. Степень привлечения дактилологии напрямую зависит от уровня владения ею и жестовым языком коммуникантами, как правило, ею передаются отдельные слова или словосочетания. Общение посредством дактильной речи требует соблюдения определенных правил [3]:

1. Дактилируют в соответствии с нормами правописания (правилами орфографии естественного языка).

2. Дактильная речь сопровождается четкой артикуляцией (устным проговариванием сообщения, возможно беззвучно).

3. В России, как и в большинстве стран мира, дактилировать общепринято одной ведущей (правой) рукой, хотя в ряде других стран дактилирование ведется обеими руками (Великобритания, Чехия, Турция и т.д.).

4. Дактильные знаки показывают ведущей рукой точно и четко.

5. Дактилирование ведут слитно и плавно в невысоком темпе.

6. Слова разделяются короткой паузой, фразы — остановкой (прекращением дактилирования на некоторое время).

7. При дактилировании рука согнута в локте, кисть руки находится на уровне плеча (недалеко ото рта диктора), слегка вынесена вперед и обращена ладонью от себя, к собеседнику.

8. В случае ошибки или непонимания повторно дактилируется все слово целиком.

9. Дактилирующий человек смотрит прямо на своего собеседника.

10. Предпочтительно, чтобы дактилирующий был в однотонной темной или черной одежде с длинными

рукавами, в этом случае собеседнику отчетливее видны кисти его рук.

11. Дактилирующий может также постепенно сдвигать (незначительно) ведущую руку справа налево по мере дактилирования слова (человек как бы пишет слово жестами по воздуху).

Важным признаком правильного дактилирования является слитность жестов, которая достигается за счет того, что при движении руки пропускаются практически все нейтральные, переходные движения. Отдельные дактилемы в потоке речи могут вести себя практически так же, как и фонемы, подвергаясь своеобразным редукциям и ассимиляциям. Однако дактильная речь привязана к достаточно консервативной коммуникативной системе — письменности, и поэтому позиционные изменения формы дактилем, равно как и редукции не так сильны, как в словесной речи.

В русской дактильной азбуке, как и в большинстве других дактильных алфавитов мира, дактилемы воспроизводят буквы. Поэтому, общаясь при помощи дактильной речи, разговаривающие следуют правилам письменной формы речи, т.е. нормам орфографии русского языка. В то же время дактилирование обязательно сопровождается устной речью. Произношение дактилируемых слов и словосочетаний должно соответствовать орфоэпическим нормам. Дактильная речь обращена, как правило, к глухому собеседнику, и обязательно сопровождается устной речью (артикуляцией губ) во избежание плохого взаимопонимания, в трудных местах адресат речи может помогать себе, читая не только жесты рук, но и артикуляцию и мимику говорящего. Дактилирующая рука должна быть расположена таким образом, чтобы неслышащему человеку было хорошо видно лицо партнера — в первую очередь, его губы. Русская дактильная речь воспроизводится одной правой рукой; в [2] отмечается, что в разных сообществах глухих существуют различные мнения относительно того, должна ли рука смещаться справа налево, как при зеркальном отображении письма, или же оставаться неподвижной. Движения производятся пальцами, кистью и предплечьем руки, локоть же преимущественно остается неподвижным.

Дактильная речь обычно применяется не как основное средство общения глухих людей, а как вспомогательное — в первую очередь тогда, когда требуется передать точный орфографический состав слова или словосочетание. Для наиболее точной передачи словесной речи, например, в тех случаях, когда употребляются имена собственные, разговорная жестовая речь дополняется дактильной. Таким же способом передаются и некоторые грамматические показатели и служебные слова (см. пример).

В данном примере “Ж” означает, что слово показывается жестом русского жестового языка, а “Д” значит, что слово сообщается собеседнику дактильно.

Многие слова, переданные дактилем, являются частью лексики национального жестового языка. В условиях бытового разговора неслышащие вставляют в свою речь специальные термины или редкие слова, показанные дактилем, по той причине, что они редко употребляются в речи, и не все знают для них соответствующие жесты. Многие дактилируемые слова учатся глухими детьми в школе и потом «по привычке» включаются в жестовый язык [2]. Еще чаще встречается дактильная речь в калькирующей жестовой речи (понятие предложено Г.Л. Зайцевой, но не считается общепринятым), которая применяется на официальных мероприятиях и встречах. В этом случае дактилемы могут дополнять калькирующие жесты для лексем, передавая аффиксы, грамматические показатели и модификаторы слов (например, дактилема+жест: З+СУХО = ЗАСУХА).

Значительна роль дактилологии и в обучении глухих людей, особенно слабослышащих детей. Сейчас существуют специальные книги и мультимедийные словари жестов, ориентированные на глухих детей [17]. В основе большинства обучающих методик лежит идея о том, что дактильная речь помогает глухим освоить устную речь и, как следствие, прививает им навыки словесного мышления и аналитического чтения, чрезвычайно важные для интеграции людей с инвалидностью в наше сообщество [18]. Поскольку устная речь чрезвычайно трудна для восприятия слабослышащих, то на начальном этапе письменная и, как следствие, дактильная речь являются самыми эффективными средствами обучения. Дактильная речь всегда сопровождается устной речью, что позволяет установить тесную взаимосвязь между дактилированием и устной речью, принятой в соответствующем сообществе. Слабослышащие дети соотносят дактилемы с нарисованными буквами, постепенно овладевая дактильной и жестовой речью, они воспринимают речевой материал с жестов рук и читают речь с губ, обучаясь самостоятельно строить дактильные слова и фразы. В конце процесса обучения учитель обращается к детям только устно, и учащиеся переходят от устно-дактильной к устной речи [2]. Основными аргументами за использование дактилологии в процессе обучения являются следующие [19]:

1. Дактильная речь легко воспринимается, адресат видит каждый элемент слова;

2. Она полностью контролируется самим говорящим; неслышащий человек может проверить себя, сопоставляя свою речь со словом, данным учителем;

3. При дактилировании формируются пальцевые кинестезии (мышечное чувство руки), благодаря которым структура слова запоминается быстрее и прочнее;

Пример

Ж	*Ж*	*Д*	*Д*	*Д*	*Д*	*Ж*	*Ж*	*Д*	*Д*
МОЯ	ПОДРУГА	ИЗ	ПСКОВА	ТАНЯ	АБРАМОВА	ПОСТУПИЛА	УЧИТЬСЯ	В	МИФИ

4. Между пальцевыми кинестезиями и кинестезиями артикуляционного аппарата устанавливаются прочные нейродинамические связи, благодаря которым дактильная речь становится опорой для устной речи;

5. Дактильная речь помогает овладевать членораздельной речью, ее грамматическим строем, словарным составом;

6. Эта форма речи обеспечивает глухому ребенку на ранних этапах обучения словесное общение.

Согласно одной из наиболее популярных методик обучения [20], изучение дактилологии целесообразно начинать с воспроизведения наиболее легких в отображении и запоминании дактилем: Г, З, Л, М, О, С. Эти дактилемы визуально очень похожи на свои буквы и резко контрастны по отношению друг к другу, форма руки достаточно характерна в каждом случае. Затем в три этапа вводятся и остальные дактилемы, уже в алфавитной последовательности.

Способ воспроизведения дактилемой буквы положен в основу метода, рекомендуемого И.Ф. Гейльманом [15], в котором обосновывается, что целесообразно учиться дактилировать буквы по группам, где дактилемы объединяются с учетом особенностей их образования и конфигурации. Работа начинается с изучения I группы — из пяти дактилем, при исполнении которых пальцы руки постепенно все более раскрываются: А, Е, Ё, С, В. Потом изучается II группа (пальцы поочередно соединяются): О, Р, Н, Ш, Щ. Далее порядок обучения дактилемам такой: III группа (К, И, Й, Н, У) — кисть руки поднята, прямые пальцы раскрыты; IV группа (З, Д, Ц, Я, Б) — кисть руки поднята, прямые пальцы соприкасаются; V группа (Г, П, Л, М, Т) — кисть руки опущена, пальцы прямые; VI группа (Ч, Ж, Ф, Ю) — кисть руки поднята, пальцы выпрямляются кончиками от себя; VII группа (Х, Э, Ь, Ъ) — кисть руки поднята, большой и указательный пальцы поочередно выпрямляются.

Тем не менее, существуют и противники использования дактильной речи при обучении глухих детей. Эта точка зрения мотивируется тем, что дактильная речь не вполне соответствует фонетической, и это различие может серьезно нарушить связь между понятием и фонетическим словом у глухих [21, 22]. В начале усвоения языка глухой ребенок должен овладевать им на основе целостного, глобального восприятия речи.

Необходимо также отметить, что в России в последние годы стали достаточно широко за счет государства проводить дорогую медицинскую операцию по кохлеарной имплантации (внедрение аппаратного речевого процессора в ухо) детям, страдающим глухотой [23]. Однако у этой операции есть свои трудности, связанные с тем, что кроме «вживления» чипа, необходим весьма длительный и сложный процесс реабилитации и обучения слуху (например, разработана реабилитационная программа-проект «Я слышу мир!»), без чего кохлеарный имплант сам по себе создает только значительную когнитивную нагрузку на человека.

2.3. Дактильная и манoральная речь

Помимо дактильной речи (“fingerspelling”), существует еще схожая с ней по принципу образования так называемая “cued speech” [24], применяемая в некоторых странах и школах для глухих людей. Пока что нет четко устоявшегося перевода этого термина на русский язык, однако иногда можно встретить термин «манoральная речь» (лат. *manus* — рука, *oris* — рот) или манoральная система [25], которую условно можно назвать фонетической дактильной речью. Если дактилология жестко привязана к соответствующему национальному алфавиту, то манoральная система речи передает фонологический состав речи; иными словами, разница между дактильной и манoральной речью сродни разнице между орфографическим написанием и фонематической транскрипцией слова.

Кроме того, в отличие от дактильной, манoральная речь базируется в основном не на жестах рук, а на чтении речи по губам; роль ручных жестов при этом сводится к «подсказкам», облегчающим собеседнику чтение с губ. Дело в том, что многие звуки речи (фонемы) образуются при схожей конфигурации формы губ (визем [26]), и жесты являются дополнительными визуальными «дифференциальными признаками» произносимых звуков. Таким образом, это средство коммуникации имеет в своей основе довольно мало общего с дактилологией, несмотря на то, что оба средства коммуникации имеют одну и ту же задачу — передавать фонемы или их графическое представление.

2.4. Роль артикуляции и мимики в дактильной и жестовой речи

Язык жестов и, в частности, дактилология способствуют восприятию и усвоению слабослышащими словесной речи. Активное изучение дактилологии укрепляет психолингвистическую связь между значениями слов и жестов. Существуют доказательства того, что формирование адекватных связей между значениями слов и жестов способствует более прочному запоминанию тех и других, облегчает их воспроизведение. Очевидно, что далеко не все люди с ограниченными возможностями по слуху научаются произносить устную речь так же хорошо, как и слышащие люди, однако они способны знать фонологическую форму слова и произносить слова «без голоса». Таким образом, устная речь является неотъемлемым сопровождающим элементом жестового языка (дактилологии и калькирующей жестовой речи [2]).

Кроме собственно проговаривания слов, дактильная речь может сопровождаться и изменением мимики лица, однако мимическая составляющая характерна преимущественно для разговорного жестового языка. Так, повелительная интонация может передаваться при помощи соответствующей утвердительной мимики лица и даже применения элементов пантомимы. Абсолютно грамматикализованной оказывается мимика и в случае со степенями сравнения прилагательных («количество качества»), в которых степень «эмоциональности» показа мимики лица диктора напрямую связана со степенью присутствия признака.

Дактилирование сопровождается устной речью в обязательном порядке. Такой метод является дополнительной «страховкой» на тот случай, если адресат не разберет какой-либо жест или слово говорящего.

Устное произношение дактилируемых слов в норме соответствует орфоэпическим нормам, иначе говоря, слова произносятся слитно и полностью, а не по буквам, например: СЛУЧАЙ – /'slučъj/ (фонетическая транскрипция), а не /s-l-u-č-a-j/ (отмечены паузы между элементами речи) или «эс-эл-у-че-а-и краткое».

3. АРХИТЕКТУРА СИСТЕМЫ МАШИННОГО АУДИОВИЗУАЛЬНОГО СИНТЕЗА

Мультимедийные компьютерные программы, способные синтезировать и распознавать элементы языка жестов, являются наиболее перспективными средствами обучения и реабилитации людей с инвалидностью по слуху. Возможно, наилучшим и универсальным вариантом реализации системы машинного синтеза жестового языка является использование виртуального трехмерного человека (аватара), который управляется посредством специального программного обеспечения компьютера (специальных кодов жестовой нотации), описывая требуемые конфигурации рук и различные типы движений и преобразуя их в действия трехмерного аватара.

Элементы жестового языка (динамические и статические жесты рук человека), в том числе калькирующей жестовой речи и дактильной речи, могут быть формализованы при помощи символов - кодов компьютерной нотации, отражающей основные дифференциальные признаки каждого жеста: форму кисти, ориентацию руки, место и характер движения. В этом смысле дактильная речь принципиально не отличается от жестовой, так как во всех случаях речь идет о жестах, выполняемых при помощи рук. Обзор и сравнение существующих систем жестовой нотации представлен в [27, 28]. Следует отметить, что для автоматических систем синтеза наиболее адекватной является система Гамбургской нотации (HamNoSys) [29], позволяющая моделировать практически любую конфигурацию и движения рук при наличии соответствующих визуальных средств. Альтернативной ей системой международной жестовой нотации является SignWriting [28].

Система машинного синтеза русской дактильной речи и калькирующей жестовой речи разработана в Санкт-Петербургском институте информатики и автоматизации РАН (СПИИРАН) совместно с Западночешским университетом (Západočeská univerzita) [30]. За основу взята модель синтеза чешского жестового языка [31], использующая международную жестовую нотацию HamNoSys. Необходимо отметить, что создана не просто система компьютерного синтеза жестов из предварительно записанных видеофрагментов, а универсальная многомодальная система для аудиовизуального синтеза русской звучащей и жестовой речи, предназначенная для диалоговых систем и систем человеко-машинного взаимодействия как для слышащих людей, так и для лиц с ограниченными возможностями по слуху. Общая архитектура системы машинного синтеза представлена на рисунке, основными компонентами многомодальной системы синтеза являются [32]: 1) текстовый процессор, осуществляющий анализ входного текста и разделение его на фразы, слова и буквы, а также формирующий фонематическую транскрипцию для последующего аудиовизуального синтеза речи; 2) имитационная модель головы/лица человека, в которой настраиваются управляющие параметры для передачи движений губ, мимики и выражений лица при говорении; 3) компьютерная система акустического синтеза разговорной русской речи, осуществляющая преобразование текста в звучащую речь по произвольному входному русскоязычному тексту; 4) русскоязычная бимодальная система аудиовизуального синтеза речи (так называемая «говорящая голова» [26]) на основе виртуальной трехмерной модели головы человека и компьютерного синтеза речи по тексту; 5) виртуальная трехмерная модель верхней части туловища и рук человека, в которой конфигурации и движения рук для показа жестов управляются на основе специальных кодов нотации HamNoSys; 6) многомодальная система объединения модальностей и синтеза аудиовизуальной звучащей речи и жестовой (дактильной и калькирующей жестовой) речи.

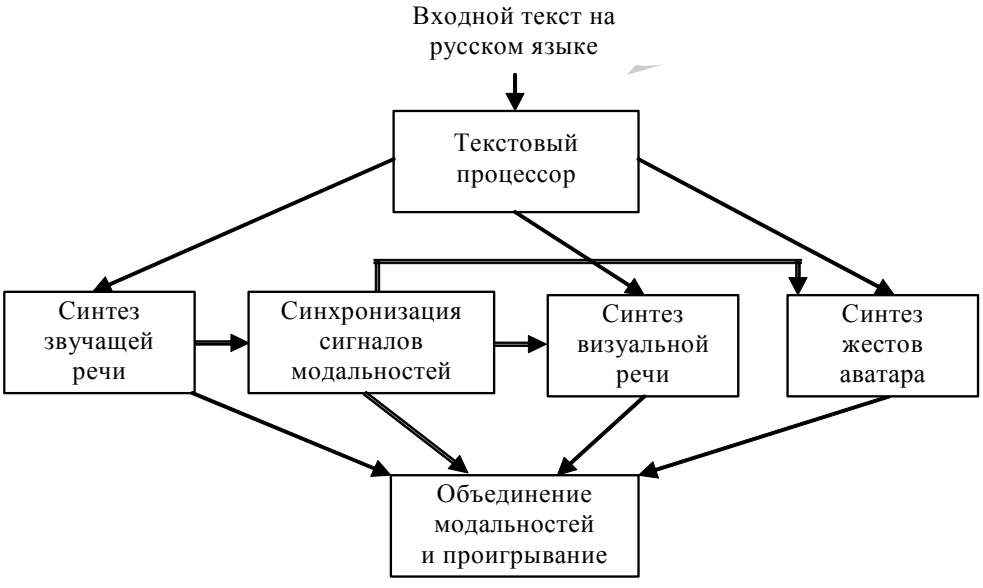

Рисунок - Общая архитектура многомодальной системы машинного синтеза русской звучащей и жестовой речи

На вход этой системы подается произвольный русскоязычный текст, который анализируется текстовым процессором; в нем выделяются предложения, слова (для аудиосинтеза речи и видеосинтеза артикуляции губ аватара) и буквы (для синтеза русской дактильной речи), которые автоматически преобразуется в символы жестовой нотации, на основе которой аватар воспроизводит мануальные жесты, декодируя символы нотации. Одной из основных задач системы является создание и экспертное наполнение словаря русского жестового языка с применением специального компьютерного редактора [33], анализирующего на входе символы HamNoSys и трансформирующего их в соответствующие движения аватара. Пока что система способна воспроизводить русскую дактильную и калькирующую жестовую речь (словарь системы составляет несколько сотен жестов для наиболее распространенных слов языка, цифр, букв и т.д.) по входному русскоязычному тексту.

Особо следует отметить, что виртуальный аватар максимально имитирует стиль жестикуляции людей. Так, все жесты (и дактилемы) следуют в речи без пауз, с соблюдением «плавности» и «текучести» жестов, что позволяет оформлять целые фразы и лексемы, а не набор изолированных друг от друга жестов. Естественность перехода от одного жеста к другому созданным аватаром можно отметить в мультимедийной демонстрации системы: www.spiiras.nw.ru/speech/demo/signlang.avi

Синхронизация звучащей и жестовой/дактильной речи в системе машинного аудиовизуального синтеза осуществляется на основе временных меток начала и конца слов звучащей речи, синтезируемой системой по тексту. Так как звучащая речь в среднем имеет более высокий темп воспроизведения, чем жестовая речь, то виртуальный аватар (говорящая голова) последовательно озвучивает и артикулирует с невысокой скоростью изолированные слова звучащей речи, дожидаясь окончания жестикуляции слова (может включать последовательность из нескольких жестов русской дактильной азбуки), плавно переходя к жестикуляции следующего слова.

Мультимедийная демонстрация работы разработанного многомодального машинного синтезатора доступна в Интернете: www.spiiras.nw.ru/speech/demo/daktilrus.avi. Необходимо отметить, что следующим этапом после создания системы машинного синтеза дактильной и калькирующей жестовой речи, будет разработка системы синтеза разговорного жестового языка и речи по тексту. Однако ее создание осложняется необходимостью машинного перевода текста в разговорный язык жестов, обладающий собственной структурой и грамматикой (отличной от русского письменного или устного языка), которые пока слабо изучены лингвистами и не формализованы, но в последние годы исследования в этом направлении ведутся [7, 12, 34, 35], что позволяет говорить о возможно скором решении данной задачи.

ЗАКЛЮЧЕНИЕ

В настоящей статье были рассмотрены основные аспекты, особенности и принципы функционирования дактильной речи, которая является неотъемлемым компонентом разговорного русского языка жестов. Дактильная речь, несмотря на свою дополнительность по отношению к письменности, обладает достаточно развитой структурой и обеспечивает простое и эффективное общение людей с инвалидностью по слуху. Приведено также описание многомодальной системы для аудиовизуального синтеза русской звучащей и жестовой речи, предназначенной для использования в универсальных диалоговых человеко-машинных системах (например, информационно-справочных автоматах самообслуживания, мобильных устройствах и т.д.), ориентированных как на слышащих людей, так и на лиц с ограниченными физическими возможностями.

Необходимо отметить, что 30 декабря 2012 г. Президент России подписал Федеральный закон, повышающий статус русского жестового языка. Благодаря поправкам в законы «Об образовании» и «О социальной защите инвалидов в РФ» русский жестовый язык определяется теперь как язык общения при наличии нарушений слуха или речи, в том числе в сферах устного использования государственного языка РФ. До этого русский жестовый язык определялся как «средство межличностного общения» и не являлся официально признанным языком России. В данном законопроекте отдельно предусматривается создание систем субтитрирования и сурдоперевода телевизионных программ и кинофильмов. Возможное внедрение в жизнь неслышащих людей автоматизированных компьютерных систем должно дополнительно привлечь к этой проблеме внимание общественности, а также обратить лингвистов и разработчиков к междисциплинарным исследованиям в этой сфере естественного языка.

СПИСОК ЛИТЕРАТУРЫ

1. Ethnologue: Languages of the World, Sixteenth Edition / ed. M. Paul Lewis. – 2009. - 1248 p.
2. Зайцева Г.Л. Жестовая речь. Дактилология: учеб. для студ. высш. учеб. заведений. – М.: ВЛАДОС, 2000. - 192 с.
3. Гейльман И.Ф. Знакомьтесь: Ручная речь. — М.: Загрей, 2001. - 172 с.
4. Димскис Л.С. Изучаем жестовый язык. — М., 2002.
5. Фрадкина Р.Н. Говорящие руки: Тематический словарь жестового языка глухих России. — М.: Изд-во «Сопричастность» ВОИ, 2001. - 598 с.
6. Воскресенский А.Л., Ильин С.Н., Железны М. О распознавании жестов языка глухих // Труды международной конференции «Диалог-2010», Бекасово, Россия, 2010. - С. 76-81.
7. Гриф М.Г., Тимофеева М.К. Интерлингва в системах машинного перевода для жестовых языков // Труды СПИИРАН. – 2012. - Вып. 20. - С. 116–137.

8. Кибрик А.А., Прозорова Е.В. Референциальный выбор в русском жестовом языке // Труды международной конференции «Диалог-2007», Бекасово, Россия, 2007. - С. 220–230.
9. Фрумкина Р.М., Браудо Т.Е. О знаковых системах, замещающих естественный язык // Культурно-историческая психология. – 2006. - № 3. - С. 28-37.
10. Aronoff M., Meir I., Sandler W. The Paradox of Sign Language Morphology // Language. – 2005. - № 81 (2).- P. 301-344.
11. Joahim G., Prillwitz S., Hanke T. International bibliography of sign language. - Hamburg, 2006.
12. Прозорова Е.В. Российский жестовый язык как предмет лингвистического исследования // Вопросы языкознания. – 2007. - № 1. - С. 44-61.
13. Гейльман И.Ф. Специфические средства общения глухих. Дактилология и мимика. В 4-х томах. - Л.: Ленинградский восстановительный центр ВОГ, 1975-1979 .
14. Friedrich J. Entzifferung verschollener Schriften und Sprachen. - Springer-Verlag: Berlin-Göttingen-New York, 1966.
15. Гейльман И.Ф. Дактилология. - Л, 1981.
16. Бондарко Л.В., Вербицкая Л.А., Гордина М.В. Основы общей фонетики. - Л., 1983.
17. Ватага С. Букварь русского жестового языка. - Архангельск, 2010.
18. Корсунская Б.Д. Методика обучения глухих дошкольников речи. - М., 1969.
19. Зыков С.А. Язык как учебный предмет в школе для глухих детей // Методика обучения глухих детей языку. - М., 1977. - С.5-51.
20. Геранкина А.Г. Практикум по дактильной речи. - М., 1972.
21. Пескова Л.П. Использование разных форм речи в обучении языку глухих дошкольников // Вопросы формирования речи аномальных детей дошкольного возраста. - М., 1982. - С.42-56.
22. Леонгард Э.И. Ранняя слухоречевая реабилитация детей с нарушениями слуха — основа их полноценного включения в общество слышащих // Проблемы младенчества. - М., 1999. - С.74-77.
23. Огородникова Е.А., Королева И.В., Пак С.П., Балякова А.А. Развитие и оценка восприятия временных характеристик звуковых сигналов у пациентов с кохлеарными имплантами с использованием инструментальных методик // Рос. оториноларингология.- 2010. - № 2. - С. 91-97.
24. Aboutabit N., Heracleous P., Beautemps D. Hand shape coding for HMM-based consonant recognition in cued speech for French // In Proc. 13th International Conference on Speech and Computer SPECOM-2009. - St. Petersburg, Russia, 2009. - P. 109-112.
25. Басова А.Г., Егоров С.Ф. История сурдопедагогики: учеб. пособие для студентов дефектол. фак. пед. ин-тов. - М.: Просвещение, 1984. - 295 с.
26. Карпов А.А., Цирульник Л.И., Железны М. Разработка компьютерной системы «говорящая голова» для аудиовизуального синтеза русской речи по тексту // Информационные технологии. – 2010. – Т. 9, № 8. - С. 13-18.
27. Карпов А.А., Кагиров И.А. Формализация лексикона системы компьютерного синтеза языка жестов // Труды СПИИРАН. – 2011. - Вып. 16. - С. 123-140.
28. Мясоедова М.А., Мясоедова З.П., Петухова Н.В., Фархадов М.П. Русский жестовый язык: банк жестов РЖЯ в письменной форме // Труды 6-го междисциплинарного семинара «Анализ разговорной русской речи» АР3-2012. - Санкт-Петербург, 2012. - С. 57-62.
29. Hanke T. HamNoSys - representing sign language data in language resources and language processing contexts // In Proc. International Conference on Language Resources and Evaluation LREC-2004. - Lisbon, Portugal, 2004. - P. 1-6.
30. Hruz M., Campr P., Dikici E., Kindirouglu A., Krňoul Z., Ronzhin Al., Sak H., Schorno D., Akarun L., Aran O., Karpov A., Saraclar M., Železný M. Automatic Fingersign to Speech Translation System // Journal on Multimodal User Interfaces, Springer-Verlag. – 2011. -Vol. 4, No. 2. - P. 61-79.
31. Krňoul Z., Kanis J., Železný M., Müller L. Czech Text-to-Sign Speech Synthesizer // In Proc. International Conference on Machine Learning for Multimodal Interaction MLMI-2007. LNCS 4892, 2008.- P. 180–191.
32. Karpov A., Železný M. Towards Russian Sign Language Synthesizer: Lexical Level // In Proc. 5th International Workshop on Representation and Processing of Sign Languages at LREC-2012. - Istanbul, Turkey, 2012. - P. 83-86.
33. Kanis J., Krňoul Z. Interactive HamNoSys Notation Editor for Signed Speech Annotation // In Proc. 6th Int. Conf. on Language Resources and Evaluation LREC-2008. – Paris, France, 2008. – P. 88–93.
34. Карпов А.А. Компьютерный анализ и синтез русского жестового языка // Вопросы языкознания. – 2011. – № 6. – С. 41-53.
35. Kimmelman V. Reflexive pronouns in Russian Sign Language and Sign Language of the Netherlands. Master thesis. - The Netherlands: University of Amsterdam , 2009.

Материал поступил в редакцию 13.09.12.

Сведения об авторе

КАРПОВ Алексей Анатольевич – кандидат технических наук, доцент, старший научный сотрудник лаборатории речевых и многомодальных интерфейсов Санкт-Петербургского института информатики и автоматизации Российской академии наук (СПИИРАН)

E-mail: karpov@iias.spb.su; karpov_a@mail.ru

АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

УДК 81’322.2 : [003.349.3. 035:033.344]

С.В. Гусев, А.М. Чеповский

Автоматическая идентификация текстов на славянских языках, пользующихся кириллицей, записанных латинским алфавитом

Рассматриваются задачи идентификации текстов на славянских языках, пользующихся кириллицей, записанных на латинице, и преобразования записи текста на алфавите языка оригинала. Описан метод определения языка и оценена его точность при применении к текстам на пяти славянских языках, записанных с использованием латинского алфавита. Предлагается методика уменьшения множественности вариантов при восстановлении написания записанных на латинице слов на алфавите языка оригинала.

Ключевые слова: *транслитерация, практическая транскрипция, идентификация естественного языка текста, автоматическая обработка текста*

ВВЕДЕНИЕ

Обработка неструктурированных данных из различных коммуникационных сервисов связана в некоторых случаях с анализом текстов, записанных в системе письменности другого языка.

В настоящей работе исследуется набор из пяти славянских языков: русского, украинского, белорусского, болгарского и македонского. Все эти языки используют кириллицу в качестве основного алфавита и могут быть записаны на латинице, для этого применяются различные правила (таблицы) транслитерации и транскрипции. Большое количество возможных правил в реальных текстах определяет главную проблему идентификации языка оригинала и восстановления написания текста на языке оригинала – проблему множественности интерпретаций символов одной системы письменности с помощью символов другой системы письменности.

Проблема анализа текстов, записанных на алфавите другого языка, разбивается на два этапа: определение оригинального языка текста и восстановление написания текста на алфавите языка оригинала. Два этапа решения рассматриваемой проблемы формулируют две задачи: идентификацию языка оригинала текста и восстановление написания текста на алфавите языка оригинала из его записи на латинице.

Проблема транслитерации/транскрипции подробно изучена лингвистами [1-3] и математиками [4]. Для практического применения в информационных технологиях наиболее разработаны вопросы практической транскрипции имен собственных [5, 6]. При развитии современных средств коммуникации становятся актуальными новые аспекты языка и его использования [7]. В частности, коммуникационные процессы связаны с использованием записей славянских текстов на латинском алфавите. При автоматическом анализе текстов возникает необходимость восстановить запись текстов на алфавите языка оригинала. В настоящей работе предлагается вероятностная модель для уменьшения множественности при восстановлении записи на кириллическом алфавите текста, записанного на латинице.

Определение языка текста можно решать, применяя принцип Байесовского классификатора к строке символов, считая, что нам известны статистические характеристики для символов в текстах на конкретном естественном языке, или текстах, относящихся к заданному классу. Модели, использующие частоты буквосочетаний, широко использовались для определения языка текста [8-10]. В работе [11] рассмотрена статистическая модель строки текста на естественном языке и описана методика ее реализации. Она была успешно применена к коротким текстам, записанным с использованием оригинальных алфавитов соответствующих языков. В настоящей работе мы рассматриваем применении той же методики для определения языка текста, написанного с использованием латинского алфавита, который в оригинале записывается кириллическим алфавитом.

СТАТИСТИЧЕСКАЯ МОДЕЛЬ СТРОКИ ТЕКСТА

Рассмотрим строку s, состоящую из N символов c^l_n $(n = 1, \ldots, N)$, принадлежащих алфавиту Σ_l одного из L языков $(l = 1,\ldots,L)$: $(l = 1,\ldots,L)$: $s = \langle c^l_1 c^l_2 \ldots c^l_N \rangle$. Предполагаем, что вероятность появления в строке s каждого n-го символа определяют $(n-1)$ символов перед ним.

Модель строки такого типа можно интерпретировать как автомат, состояниями которого являются последовательности символов длины (***n-1)***, а переходы между состояниями помечены вероятностями появления соответствующих символов. Таким образом, переход из состояния $s_1 = <c^l_1 c^l_2 \ldots c^l_{n-1}>$ в состояние $s_2 = <c^l_2 \ldots c^l_n>$ происходит по символу c^l_n и определяется условной вероятностью $P(c^l_n \mid c^l_1 c^l_2 \ldots c^l_{n-1})$.

Условную вероятность появления подстроки $s_n = < c^l_1 c^l_2 \ldots c^l_n>$ длиной ***n*** ***l***-го языка определяем через частоту $f(c^l_1 c^l_2 \ldots c^l_n)$ встречаемости данной подстроки в текстах на ***l***-м естественном языке (верхний индекс языка опущен для упрощения формул):

$$P(c_n \mid c_1 \ldots c_{n-1}) =$$

$$= \begin{cases} \dfrac{f(c_1 \ldots c_n)}{f(c_1 \ldots c_{n-1})} \cdot (1 - p_0), & f(c_1 \ldots c_n) \neq 0, f(c_1 \ldots c_{n-1}) \geq \theta \\ p_0, & f(c_1 \ldots c_n) = 0, f(c_1 \ldots c_{n-1}) \geq \theta \\ P(c_n \mid c_2 \ldots c_{n-1}), & f(c_1 \ldots c_{n-1}) < \theta \end{cases}, \quad (1)$$

где

$$P(c_n) = \begin{cases} f(c_n) \cdot (1 - p_0), & f(c_n) \neq 0 \\ p_0, & f(c_n) = 0 \end{cases}, \quad (2)$$

p_o – константа, задающая пороговый уровень вероятности для подстрок с малой частотой появления,

θ – константа, определяющая пороговое значение глубины учета длин подстрок при вычислении вероятностей для подстроки.

Константа глубины учета длин подстрок θ при вычислении вероятностей появления подстроки определяет порог для величины частоты $f(c^l_1 c^l_2 \ldots c^l_{n-1})$, при превышении которого используется оценка условной вероятности появления очередного ***n***-го символа по частотным характеристикам для строки из предыдущих (***n-1***) символов. Если же частота $f(c^l_1 c^l_2 \ldots c^l_{n-1})$ встречаемости подстроки из предыдущих (***n-1***) символов меньше порога θ, то рассматриваются частотные характеристики для подстроки из предыдущих (***n-2***) символов. Формула (1) для определения вероятности появления подстроки применяется рекурсивно для ***q= n-1, n-2, …, 1 (q>0)***.

Вероятность появления какой-либо строки длины ***N*** ***l***-го языка равна произведению вероятностей всех переходов между состояниями, которые произошли при чтении данной строки автоматом:

$$P(s^l_N) = \prod_{n=q}^{N} P(c^l_n \mid c^l_{n-q+1} \ldots c^l_{n-1}) . \quad (3)$$

Строка считается принадлежащей языку, для которого значение вероятности принимает наибольшую величину.

РЕАЛИЗАЦИЯ МОДЕЛИ ДЛЯ ИДЕНТИФИКАЦИИ ЯЗЫКА ОРИГИНАЛА

Каждый текст на естественном языке подвергается предварительной обработке, в результате которой из текста получается набор слов, состоящих только из символов алфавита соответствующего языка, приведённых к нижнему регистру. По полученному набору слов формируется частотный словарь буквосочетаний длиной от 1 до ***k+1***, с учётом количества вхождений слов в частотный словарь. Эта процедура выполняется при построении модели строки конкретного языка, которая строится на основе массива обучающих текстов и завершается заполнением базы данных профилей для каждого исследуемого языка. Модель представляется в виде конечного автомата, в котором состояния помечены последовательностями из предыдущих символов, а переходы помечены следующим символом и вероятностью осуществления перехода.

При определении языка текста вероятность появления каждого слова в данном языке определяется следующим образом. К каждому слову текста прибавляется один пробел справа, и оно подаётся на вход автомата. Обработка слова начинается из начального состояния автомата, соответствующего буквосочетанию, состоящему из (***n-1***) пробела. Для очередного символа вычисляется вероятность перехода по нему из текущего состояния в следующее состояние в соответствии с (1) – (3). Вероятностью появления данного слова считается произведение вероятностей всех переходов, произошедших во время обработки слова автоматом.

Язык оригинала текста определяется по методике, описанной в [11]. Для определения языка текста оценивается вероятность соответствия рассматриваемого текста моделям строк для каждого естественного языка, выбирается максимальная вероятность, соответствующая языку, на котором написан текст.

Эксперименты по определению языка текста оригинала проводились для пяти языков, использующих кириллическую письменность: русский, украинский, белорусский, болгарский, македонский. Для каждого из пяти определяемых языков были составлены массивы обучающих текстов, объёмом не менее 500 тысяч символов. Также были составлены проверочные массивы объёмом не менее 50 тысяч символов для каждого языка. Проверочные массивы состоят из реальных записанных на латинице текстов, взятых из различных интернет-источников.

Для оценки точности распознавания используются следующие характеристики: коэффициент релевантности, коэффициент полноты и F-мера. Когда мы определяем язык текста, взятого из проверочного массива, то знаем правильный ответ и можем понять, был ли язык определён корректно. Для массива текстов мы можем получить количество текстов с правильно определённым языком, а также количество текстов с ошибочно определённым языком. Коэффициент релевантности вычисляется как доля текстовых образцов в результирующем наборе, отнесённом к определенному языку, которые действительно являются текстами на соответствующем естественном языке. Коэффициент полноты вычисляется как доля текстовых образцов на соответствующем языке, которые правильно отнесены к данному языку. В качестве главной характеристики качества распознавания используется F-мера, которая определяется как взвешенное гармоническое среднее коэффициента релевантности и коэффициента полноты.

Экспериментальная проверка точности идентификации языка текста проводилась следующим обра-

зом: сначала по массиву обучающих текстов строился набор моделей текста, каждой из которых был приписан язык того текста, на котором происходило обучение. Затем по проверочным текстам строился набор текстовых образцов заданной длины – из каждого проверочного текста выделялось 1000 образцов, взятых на случайно выбранной позиции, соответствующей началу слова. Каждый образец из набора оценивался с помощью каждой из построенных моделей и ему приписывался язык, соответствующий модели с наибольшей оценкой. По собранной статистике правильно и неправильно определённых языков были вычислены коэффициенты полноты, точности и F-мера.

Набор моделей текста состоял из моделей для записанных на латинице текстов на пяти рассматриваемых языках и моделей для 31 языка, использующих латиницу в качестве родного алфавита. Для данных 36 языков были подобраны массивы проверочных текстов. Для определения языка оригинала текста было проверено два различных варианта создания набора моделей текста. В первом варианте для каждого языка создавалось несколько моделей – по одной для каждой из таблиц транслитерации и транскрипции (от 4 до 10 таблиц для каждого языка). Каждая модель обучалась на тексте, сгенерированном с помощью конкретной таблицы и ей приписывался соответствующий язык. Таким образом, в наборе присутствовало несколько моделей для одного языка, и если образец текста получал максимальную оценку для любой из этих моделей, то он относился к этому языку. Во втором варианте – для каждого языка создавалась ровно одна модель, которая обучалась на всех текстах для данного языка, независимо от вида таблиц транслитерации и транскрипции.

Вычисления показали, что использование нескольких моделей для одного языка вместо одной существенно не улучшает определение языка оригинала текста. Точность определения языка оригинала текстов, записанных на латинице, значительно ниже, чем точность определения языка текстов, написанных на тех же языках, но с использованием кириллицы, из-за проблем множественности интерпретации одинаковых букв и буквосочетаний в разных языках, использующих кириллическую письменность.

На рис. 1 и рис. 2 приведены результаты определения языка оригинала текста на основе описанной модели. На графиках показана усреднённая точность определения языка оригинала текстов, записанных на латинице, в зависимости от длины текста, измеренной в символах. На рис. 1 показана зависимость F-меры от длины текста для точности определения того, является ли текст записью на латинице текста, записанного на языке с кириллическим алфавитом, или одним из текстов на языке, использующих латинский алфавит. На рис. 2 приведены значения F-меры определения конкретного языка оригинала текста в зависимости от длины текста. Видно, что для всех пяти исследуемых языков, использующих в оригинале кириллическую письменность, высока точность определения языка оригинала для текстов длиной всего лишь в тридцать символов.

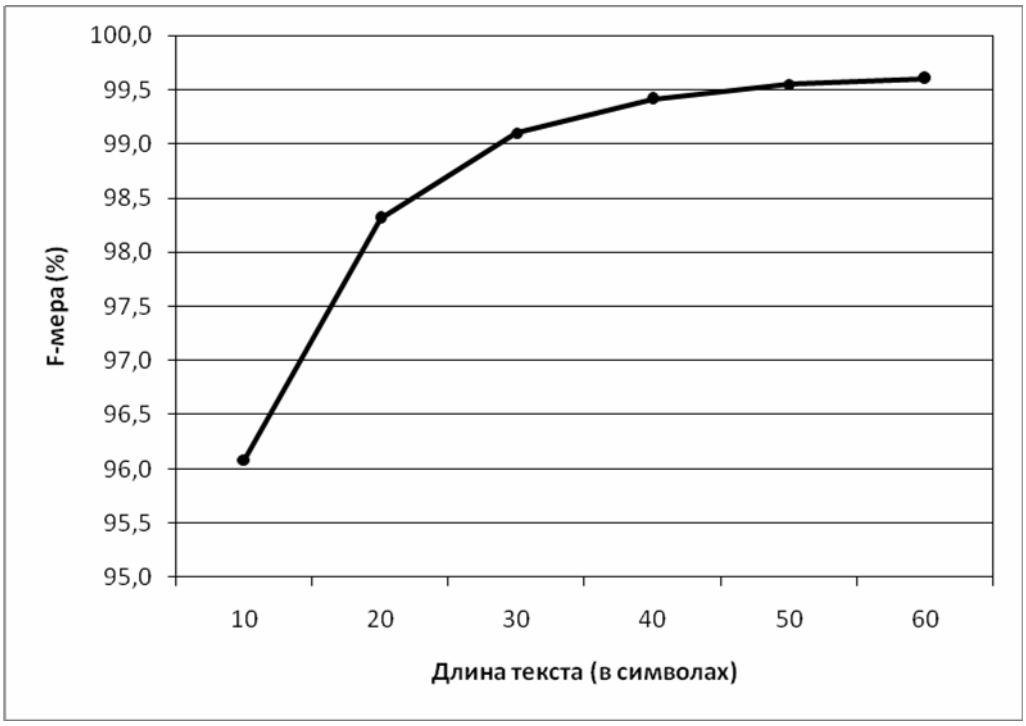

Рис. 1. Значение F-меры в зависимости от длины текста для определения, является ли текст записью на латинице текста на языке с кириллическим алфавитом

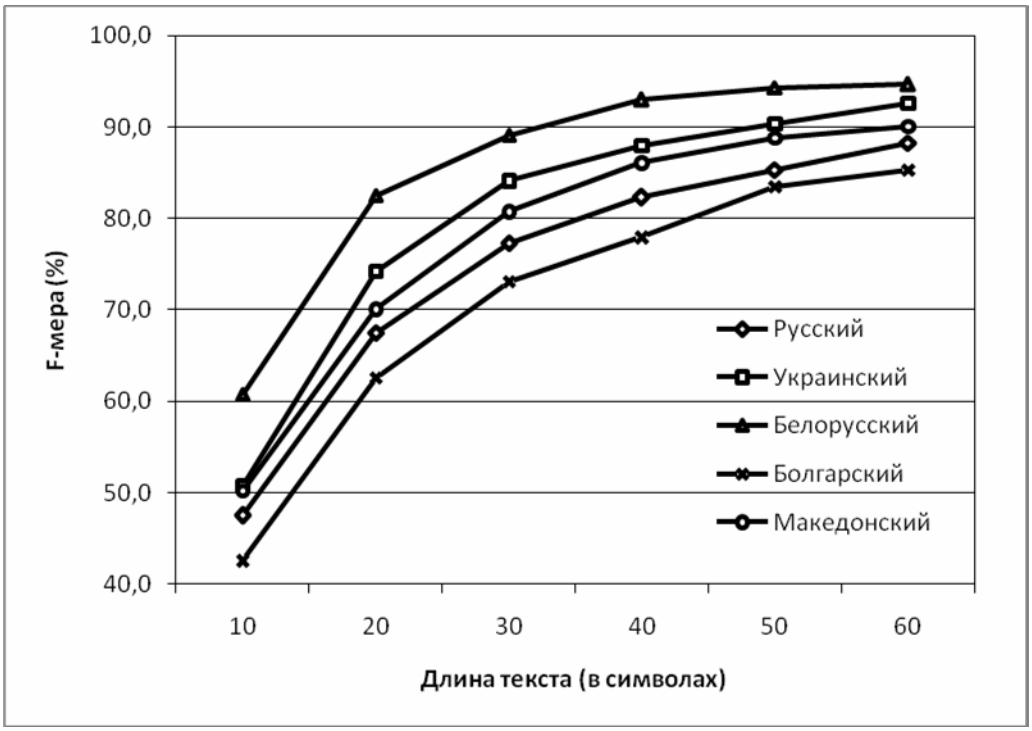

Рис. 2. Значение F-меры определения языка оригинала текста в зависимости от длины текста.

ВОССТАНОВЛЕНИЕ ЗАПИСИ ТЕКСТА НА АЛФАВИТЕ ЯЗЫКА ОРИГИНАЛА

Определив, что строка текста исходно является текстом на славянском языке, записанном латиницей, необходимо решать вторую задачу: восстановление записи текста на алфавите языка оригинала из его записи на латинице.

Пусть заданы правила вида $\boldsymbol{L \rightarrow R}$, где $\boldsymbol{L}$ – строка символов алфавита, на котором записан исходный текст (латиница); $\boldsymbol{R}$ – строка символов системы письменности языка оригинала (кириллица).

И пусть есть исходное слово $\boldsymbol{w}$, тогда строка $\boldsymbol{u}$ будет восстановлением слова $\boldsymbol{w}$ на алфавите языка оригинала, если существует последовательность правил $\boldsymbol{T_1, T_2,\ldots,T_n}$, такая, что:

$$\sum_{i=1}^{n} L_i = w \text{ и } \sum_{i=1}^{n} R_i = u \ , \qquad (4)$$

где $\boldsymbol{T_i = L \rightarrow R.}$

Последовательная конкатенация левых частей правил в (4) образует исходное слово на латинице, а конкатенация в том же порядке правых частей правил образует слово, восстановленное на алфавите языка оригинала. Для одного исходного слова может существовать более одного восстановленного варианта.

Для снижения множественности при восстановлении слов, записанных на алфавите языка оригинала, можно использовать морфологический словарь, который хранит словоформы русских слов в структуре данных типа «бора». Каждый узел бора, хранит информацию о морфологических характеристиках словоформ, читающихся на пути от корня бора до данного узла. Бор позволяет производить поиск по префиксу словоформ и имеет следующие полезные свойства:

- каждый префикс читается не более чем на одном пути, начинающемся из корня бора;
- если некоторый префикс не содержится в боре, то в словаре не содержится ни одной строки, начинающейся с данного префикса.

Частичным восстановлением написания слова $\boldsymbol{w}$ будем считать начало последовательности $\boldsymbol{T_1, T_2,\ldots,T_m}$, $\boldsymbol{m < n}$. Каждому частично восстановленному слову соответствует не более одного узла в боре, при этом на пути к этому узлу читается строка, полученная конкатенацией правых частей правил, входящих в частично восстановленное написание. Если частичной последовательности не соответствует ни один узел бора, то ни один из вариантов восстановления написания, содержащих данную частичную последовательность в качестве префикса, не содержится в боре.

Алгоритм получения восстановленного написания слова с проверкой по морфологическому словарю строится следующим образом: есть очередь состояний, каждое из которых описывается частично восстановленным написанием и соответствующим ему узлом бора. В начале работы алгоритма очередь содержит ровно одно состояние – частично восстановленное написание, соответствующее пустой последовательности правил, и корневой узел бора. Во время работы из очереди извлекается очередное состояние и для него строятся все состояния, которые достижимы из него при помощи добавления одного правила в последовательность правил преобразования символов. Состояние

считается достижимым, если частично восстановленное написание с добавленным правилом корректно – конкатенация левых частей правил является префиксом исходного слова, и существует узел в боре, на пути к которому читается конкатенация правых частей правил. Все достижимые состояния заносятся в очередь состояний. По ходу работы алгоритма состояния, у которых частично восстановленное написание совпадает с полностью восстановленным написанием слова, и узел бора содержит информацию о словоформах, заносятся в массив результатов. Алгоритм завершается, когда очередь состояний пуста.

Описанный выше алгоритм не решает полностью задачу однозначного восстановления написания слов на алфавите языка оригинала. Остаются две проблемы. Во-первых, после фильтрации по словарю для многих слов остаётся по несколько допустимых вариантов восстановления написания. Во-вторых, многих слов современной лексики может не быть в словаре. Поэтому необходимы дополнительные методы фильтрации восстановления написания слов, позволяющих выбрать наиболее подходящие варианты и уменьшить возникающую множественность.

ВЕРОЯТНОСТНАЯ МОДЕЛЬ ФИЛЬТРАЦИИ ВАРИАНТОВ ВОССТАНОВЛЕНИЯ ТЕКСТА

Модель фильтрации основана на определении вероятностей вариантов восстановления написания слов по вероятностям использования правил, составляющих каждый вариант. Пусть w – исходное слово на латинице, u – один из вариантов его написания на кириллице, $T_1, T_2,\ldots,T_n$ – последовательность правил преобразования, переводящих w в u.

Пусть: $L(w)$ – разбиение слова w на подстроки $L_1\ldots L_n$, такое, что $\sum_{i=1}^{n} L_i = w$, и для каждого L_i существует хотя бы одно правило с такой левой частью. Аналогично, $R(u)$ – разбиение слова u на подстроки $R_1\ldots R_n$, такое что $\sum_{i=1}^{n} R_i = u$, и для каждого R_i существует хотя бы одно правило с такой правой частью.

Рассмотрим два варианта построения модели. Первый вариант позволяет определить вероятность появления варианта восстановленного написания u для исходного слова w, этот вариант будем называть «***LR***». Второй вариант позволяет получить вероятность того исходного слова w для варианта восстановленного написания u, этот вариант обозначим как «***RL***».

Построение модели «***LR***». Пусть $P(L(w))$ – вероятность появления в тексте на латинице разбиения $L(w)$, а $P(R_i|L_i)$ – вероятность использования правила $L_i \to R_i$ для исходной подстроки L_i. Будем считать, что вероятности использования каждого из правил не зависят от выбранного разбиения слова и правил, использованных для других фрагментов исходного слова. Тогда вероятность появления варианта восстановленного написания u будет равна:

$$P(u) = P(L(w)) \cdot \prod_{i=1}^{n} P(R_i \mid L_i). \qquad (5)$$

Построение модели «***RL***». Пусть $P(R(u))$ – вероятность появления в восстановленном тексте на кириллице разбиения $R(u)$, а $P(L_i|R_i)$ – вероятность использования правила преобразования $L_i \to R_i$ для подстроки R_i. Также будем считать, что вероятности использования каждого из правил преобразования не зависят от выбранного разбиения слова и правил, использованных для других фрагментов исходного слова. Тогда вероятность исходного слова w для заданного варианта u будет равна:

$$P(w) = P(R(u)) \cdot \prod_{i=1}^{n} P(L_i \mid R_i). \qquad (6)$$

Для вычисления оценок вероятностей на первом шаге для каждого слова определяется, какие из его буквосочетаний восстановлены с использованием ровно одного правила. Например, рассмотрим слово *slezami*, для которого получено два проверенных по морфологическому словарю варианта восстановления написания слов на кириллическом алфавите: *шлицами, слезами*. В этом слове для букв *s, e* и *z* в разных вариантах восстановления написания слова используются разные правила, а для букв *l, a, m* и *i* используются одни и те же правила. Аналогично проверяются все остальные слова текста и подсчитываются количества однозначных использований каждого из правил. Так, слово *slezami* увеличивает на единицу количества использований правил «***l→л***», «***a→a***», «***m→м***» и «***i→и***».

Затем для каждого правила оцениваются вероятности их появления:

$$P(L \mid R) = \begin{cases} \dfrac{N(L \to R)}{\sum N(L \to X)} \cdot (1 - p_0), & N(L \to R) \neq 0 \\ p_0, & N(L \to R) = 0 \end{cases}, \qquad (7)$$

где $N(L \to R)$ – количество однозначных использований правила, переводящего строку L в строку R; $N(L \to X)$ – количество однозначных использований всех правил, переводящих строку L в какую-либо строку; p_0 – параметр модели, задающий вероятность появления неизвестных правил.

$$P(R \mid L) = \begin{cases} \dfrac{N(L \to R)}{\sum N(X \to R)} \cdot (1 - p_0), & N(L \to R) \neq 0 \\ p_0, & N(L \to R) = 0 \end{cases}, \qquad (8)$$

где $N(L \to R)$ – количество однозначных использований правила, переводящего строку L в строку R; $N(X \to R)$ – количество однозначных использований всех правил, переводящих какую-либо строку в строку R; p_0 – параметр модели, задающий вероятность появления неизвестных правил.

При реализации описанной фильтрации вариантов для каждого слова определяется вариант восстановления написания слов на кириллическом алфавите с максимальной вероятностью, и из вариантов восстановления написания слов удаляются все варианты, вероятность которых меньше максимальной.

ЭКСПЕРИМЕНТАЛЬНАЯ ПРОВЕРКА МЕТОДИКИ ВОССТАНОВЛЕНИЯ ТЕКСТА

Для анализа работы алгоритма был составлен корпус русскоязычных текстов, записанных на латинице, общим объёмом около 500 кБ. Из него были подготовлены три тестовых массива. Первые два массива составлены из выбранных случайным образом фраз из корпуса. Под фразой в данном случае понимается небольшое простое предложение или часть сложного предложения. Более короткие фразы были включены в первый массив (178 текстов средней длиной в 55 символов), более длинные – во второй (178 текстов средней длиной в 126 символов). Третий тестовый массив составлен из относительно длинных текстов (4 текста средней длиной в 1273 символа), каждый из которых полностью написан одним человеком с использованием одних и тех же правил преобразования (схемы транслитерации). Для всех текстов из тестовых массивов были составлены правильные варианты восстановления написания слов, чтобы можно было автоматически проверять корректность вариантов восстановленных написаний слов на кириллическом алфавите.

Проверялось использование алгоритма в двух режимах: с обучением на анализируемом тексте и с предварительным обучением на всём корпусе текстов. При обучении на анализируемом тексте для настройки весовых коэффициентов используется только текст, для которого восстанавливается написание слов на кириллице. При обучении на корпусе для настройки коэффициентов используется весь исходный корпус текстов, анализируемый текст в настройке коэффициентов не участвует.

При восстановлении написания слов для каждого слова, как правило, получается несколько вариантов с разными весовыми коэффициентами. Оставляется только вариант с наибольшим весом, все остальные варианты отбрасываются. Иногда максимальный вес имеют сразу несколько вариантов, в этом случае все они выдаются в качестве ответа.

В табл. 1 приведены количества слов (в процентах от общего количества слов в массиве) для которых найдены один, два или больше двух вариантов восстановленного написания слова на кириллице при использовании алгоритма в режиме обучения на анализируемом тексте. При этом среди вариантов восстановленного написания присутствует правильный вариант. Последняя строка содержит количество слов, для которых, среди вариантов восстановленного написания слова на кириллице, правильный вариант отсутствует. В табл. 2 приведены аналогичные результаты для алгоритма, использующего предварительное обучение на всём корпусе текстов.

Результаты, приведенные в табл. 1 и табл. 2, показывают, что предварительное обучение на большом корпусе текстов, записанных на латинице, позволяет получить точность восстановления написания слова на кириллице около 80%. Для достаточно больших текстов обучение непосредственно на восстанавливаемом тексте позволяет достигать хороших результатов восстановления написания слов на кириллице.

Таблица 1

Количество слов (в %), для которых найдены варианты восстановления написания при обучении только на анализируемом тексте

Количество вариантов	Массив 1		Массив 2		Массив 3	
	LR	RL	LR	RL	LR	RL
1	58,3 %	58,3 %	71,6 %	71,1 %	86 %	84,9 %
2	11,9 %	11,8 %	6,1 %	6,7 %	0,3 %	1,5 %
3 и более	7,2 %	7,4 %	3,0 %	3,3 %	0,1 %	0,3 %
нет вариантов	22,5 %	22,4 %	19,3 %	19 %	13,6 %	13,4 %

Таблица 2

Количество слов (в %), для которых найдены варианты восстановления написания при обучении на всём корпусе текстов

Количество вариантов	Массив 1		Массив 2		Массив 3	
	LR	RL	LR	RL	LR	RL
1	79,8 %	76,2 %	80,4 %	75,1 %	84,9 %	81,7 %
2	0,1 %	0,4 %	0 %	0,2 %	0,1 %	0,4 %
3 и более	0 %	0 %	0 %	0 %	0,1 %	0,1 %
нет вариантов	20,1 %	20,2 %	19,6 %	24,7 %	14,9 %	17,8 %

ЗАКЛЮЧЕНИЕ

Рассмотренная модель текста успешно разделяет тексты на естественных языках, для которых латиница является оригинальным алфавитом, и тексты, на славянских языках, записанных с использованием латинского алфавита. Точность разделения достигает 98% для строк длиной 20 символов.

Идентификация языков, использующих в оригинале кириллические алфавиты для текстов, записанных на латинице, достигает точности более 80% для строк, содержащих более 40 символов.

Разработанная методика позволяет в режиме обучения в реальном времени успешно восстанавливать написание на кириллице слов, записанных на латинице.

СПИСОК ЛИТЕРАТУРЫ

1. Реформатский А.А. Введение в языковедение (5-е издание). – М.: Аспект Пресс, 2008. — 536 с.
2. Реформатский А.А. Практическая транскрипция иноязычных собственных имен // Известия академии наук СССР. Отделение литературы и языка. – 1960. – Т. XIX, вып. 6. - С. 529—534.
3. Лингвистический энциклопедический словарь / гл. ред. В.Н.Ярцева. — 2-е изд., дополненное — М.: Большая Российская энциклопедия, 2002. — 709 с.
4. Успенский В.А. К проблеме транслитерации русских текстов латинскими буквами // Труды по нематематике. В 2 т.– М.: ОГИ, 2002. – Т. 1. - С. 390—412
5. Гиляревский Р.С., Старостин Б.А. Иностранные имена и названия в русском тексте: Справочник. - М.: Высшая школа, 1985. — 303 с.
6. Практическая транскрипция фамильно-именных групп / ред. Р.С. Гиляревский; Ин-т прикладной математики им. М.В.Келдыша – М.: Наука, 2006. – 526 с.
7. Валгина Н.С. Активные процессы в современном русском языке. — М.: Логос, 2003. — 304 с.
8. McNamee B. Paul. Language identification: a solved problem suitable for undergraduate instruction // Journal of Computing Sciences in Colleges. – 2005. –Vol. 20(3). – P. 94–101.
9. Cavnar W. B., Trenkle J. M. N-gram-based text categorization // Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval. - Las Vegas, US, 1994. – P. 161–175.
10. Vatanen Tommi, Väyrynen Jaakko J., Virpioja Sami. Language identification of short text segments with n-gram models // Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), 2010. – P. 3423-3430.
11. Гусев С.В., Чеповский А.М. Модель для идентификации естественного языка текста // Бизнес-информатика. - 2011. - № 3(17). - С. 31-35.

Материал поступил в редакцию 16.10.12.

Сведения об авторах

ЧЕПОВСКИЙ Андрей Михайлович - кандидат технических наук, доцент кафедры Анализа данных и искусственного интеллекта Национального исследовательского университета – Высшая школа экономики (НИУ ВШЭ), Москва.

E-mail: achep@adde.math.msu.su; achepovskiy@hse.ru

ГУСЕВ Сергей Валерьевич - старший преподаватель кафедры Анализа данных и искусственного интеллекта НИУ ВШЭ, Москва

E-mail: unk379@mail.ru

РОССИЙСКАЯ АКАДЕМИЯ НАУК

Федеральное государственное бюджетное учреждение науки
ВСЕРОССИЙСКИЙ ИНСТИТУТ НАУЧНОЙ И ТЕХНИЧЕСКОЙ ИНФОРМАЦИИ
РОССИЙСКОЙ АКАДЕМИИ НАУК

предлагает научным работникам, аспирантам и другим специалистам в области естественных, точных и технических наук, желающим быстро и эффективно опубликовать результаты своей научной и научно-производственной деятельности, использовать способ публикации своих работ через *систему депонирования.*

«Депонирование (передача на хранение) – особый метод публикации научных работ (отдельных статей, обзоров, монографий, сборников научных трудов, материалов научных мероприятий – конференций, симпозиумов, съездов, семинаров) узкоспециального профиля, разрешенных в установленном порядке к открытому опубликованию, которые нецелесообразно издавать полиграфическим способом печати, а также работ широкого профиля, срочная информация о которых необходима для утверждения их приоритета. Депонирование предусматривает прием, учет, регистрацию, хранение научных работ и обязательное размещение информации о них в специальных информационных изданиях».

Подготовка и передача на депонирование научных работ происходит в соответствии с «Инструкцией о порядке депонирования научных работ по естественным, техническим, социальным и гуманитарным наукам» (М., 2003).

Результатом депонирования является публикация информации о депонированных научных работах в информационных изданиях ВИНИТИ РАН – Реферативном журнале и аннотированном библиографическом указателе «Депонированные научные работы».

В соответствии с “Положением о порядке присуждения ученых степеней”, утвержденным Постановлением Правительства Российской Федерации от 30.01.2002 № 74 (в ред. Постановлений Правительства РФ от 20.04.2006 № 227, от 02.06.2008 № 424, от 20.06.2011 № 475), научные работы, депонированные в организациях государственной системы научно-технической информации, признаны публикациями, учитывающимися при защите кандидатских и докторских диссертаций.

Подать научную работу на депонирование можно обратившись в Отдел депонирования ВИНИТИ РАН по адресу:

125190, Москва, ул. Усиевича, 20.
ВИНИТИ РАН, Отдел депонирования научных работ.
Тел.: 8 (499) 155-43-28, Факс: 8 (499) 943-00-60.
e-mail: dep@viniti.ru

С инструкцией о порядке депонирования можно ознакомиться на сайте ВИНИТИ РАН: **http://www.viniti.ru**

УВАЖАЕМЫЕ ЧИТАТЕЛИ!

ЦЕНТР НАУЧНО-ИНФОРМАЦИОННОГО ОБСЛУЖИВАНИЯ ВИНИТИ РАН

ПРЕДОСТАВЛЯЕТ КОПИИ ПЕРВОИСТОЧНИКОВ

ВИНИТИ РАН осуществляет обслуживание копиями первоисточников, хранящихся в фонде научно-технической литературы ВИНИТИ, в фондах других библиотек, а также в доступных ВИНИТИ электронных ресурсах.

Фонд научно-технической литературы ВИНИТИ включает более 2 млн изданий по точным, естественным и техническим наукам, в т.ч.:

– отечественные и иностранные периодические и продолжающиеся издания – с 1987 г.;

– отечественные книги – с 1987 г.;

– иностранные книги – с 1991 г.;

– рукописи, депонированные в ВИНИТИ, – с 1962 г.

Заказы на бумажные или электронные копии первоисточников принимает Центр научно-информационного обслуживания (ЦНИО) ВИНИТИ. ЦНИО ВИНИТИ обслуживает коллективных (организации и учреждения) и индивидуальных пользователей.

Формы обслуживания:

– абонементная (на основе договоров и предоплаты);

– разовые заказы (с предоплатой заказа по счету);

– индивидуальная форма обслуживания в читальном зале ЦНИО ВИНИТИ.

На сайте ВИНИТИ (http://www.viniti.ru) представлен полный Электронный каталог научно-технической литературы (http://catalog.viniti.ru), зарегистрированной в ВИНИТИ с 1994 г. Доступ для просмотра и поиска по Каталогу свободный. Постоянные абоненты ЦНИО ВИНИТИ, имеющие логин и пароль для работы с Каталогом, могут делать заказ копий непосредственно через Каталог.

Услуги по изготовлению копий первоисточников из фондов других библиотек предоставляются только постоянным абонентам. Место хранения первоисточников указывается в Электронном каталоге.

За подробной информацией обращаться по адресу:

125190, Россия, Москва, ул. Усиевича, 20, ВИНИТИ РАН. ЦНИО

Телефоны: 8 (499)155-42-43, 155-42-09, 152-54-59

Факс: 8 (499) 943-00-60

E-mail: cnio@viniti.ru; **URL:** http://www.viniti.ru

БАЗА ДАННЫХ ВИНИТИ РАН

ВИНИТИ предлагает к использованию через WWW-сервер (**http://www.viniti.ru**) крупнейшую Федеральную базу отечественных и зарубежных публикаций по естественным, точным и техническим наукам. БД ВИНИТИ РАН генерируется с 1981 г., обновляется ежемесячно, пополнение составляет около 1 млн документов в год. БД ВИНИТИ представлена ретроспективными тематическими фрагментами и единой политематической БД (ретроспектива с 2001 г.), объединяющей все тематические фрагменты БД ВИНИТИ.

БД ВИНИТИ РАН в сети INTERNET

Сервер ВИНИТИ – http://www.viniti.ru – обеспечивает on-line доступ к Базе данных ВИНИТИ РАН круглосуточно без выходных.

На основе БД ВИНИТИ РАН предоставляются следующие услуги:

- Диалоговый поиск научно-технической информации **в режиме on-line**;
- **Демо-версия**, позволяющая ознакомиться с основными функциями поисковой системы, составом данных, формами представления документов и получить навыки работы с системой;
- **Поисковые эксперты** ВИНИТИ выполнят тематический поиск по разовым или постоянным запросам, а также окажут **консультационные услуги**.

БД ВИНИТИ РАН на CD-ROM

Любые наборы тематических фрагментов БД ВИНИТИ или их разделов могут быть предоставлены **на CD-ROM в поисковой системе (ИПС) "Сокол"**, обеспечивающей все поисковые функции, доступные в режиме on-line:

- Поиск можно вести в годовом или ретроспективном массиве (за несколько лет сразу) в одном или нескольких тематических фрагментах .
- Поиск по словам и любым словосочетаниям из заглавия, реферата, ключевых слов.
- Использование года, языка, рубрик, шифров тематических разделов БД для уточнения поиска.
- Поиск по словарю, выполняющему функции многоаспектного указателя, в том числе авторского, предметного, источников, индексов МПК, номеров патентных документов и депонированных рукописей и т.д.
- Возможность запоминания запросов для последующего использования и/или редактирования их.
- Чтение документов не только как в РЖ (последовательный просмотр документов одного номера за другим), но и чтение документов нужных тематических фрагментов (разделов) по оглавлению за весь период заказанной ретроспективы.

ИПС "Сокол" является прикладной программой Microsoft Windows.

Любые наборы тематических фрагментов БД ВИНИТИ или их разделов могут быть подготовлены **в коммуникативных форматах** ISO-2709, МЕКОФ, txt на любых видах электронных носителей.

Продукты предоставляются на договорной основе.

Информационная служба БД ВИНИТИ: 125190, Москва, ул. Усиевича 20, ВИНИТИ
Телефон: (499) 155-45-01, 155-45-02, **Факс:** (499) 152-62-31 **e-mail:** csbd@viniti.ru