УДК 81’322.2’367

Э.С. Клышинский, Н.А. Кочеткова, В.К. Логачева

Метод кластеризации слов с использованием информации об их синтаксической связности[1]

Излагаются результаты экспериментов в области кластеризации слов русского языка. Разработан новый метод кластеризации, позволяющий разделить слова по семантическим классам в соответствии с их синтаксическими связями с другими словами. Для проведения экспериментов был взят большой (около 7,2 млрд слов) неразмеченный корпус текстов, из которого были извлечены синтаксически связанные группы слов различного вида. Для извлечения подобных связей использовался набор конечных автоматов, обрабатывающих контактные группы слов, не омонимичных по части речи. Для экспериментов были использованы связи вида «существительное + прилагательное» и «глагол (+ предлог) + существительное». Для разделения слов по кластерам использовался модифицированный метод группового среднего. Качество разделения слов по кластерам было проверено с использованием «Русского семантического словаря» под общей редакцией Н. Ю. Шведовой. В итоге были получено значение NMI, равное 0,457, и F_1-мера, равная 0,607.

Ключевые слова: *кластеризация слов, синтаксическая связанность, семантическая близость*

ВВЕДЕНИЕ

В ряде случаев при анализе текстов, исследуя наблюдаемые факты и применяя к ним методы извлечения фактов, можно выделить скрытые законы, которым подчиняется язык. Одной из областей, к которой можно применить подобный подход, является кластеризация слов по смыслу с использованием информации об их сочетаемости. Общеизвестно, что синтаксическая роль слова в предложении зависит от семантических характеристик как самого этого слова, так и окружающих его слов, а различия в синтаксической структуре сходных предложений отражает разницу в их смыслах. Подобная связь заставляет думать, что мы можем извлечь семантическую информацию, используя информацию о синтаксической структуре текстов (обратное уже активно используется системами синтаксического анализа). В соответствии с этим предположением З. Харрис в 1954 г. сформулировал гипотезу о том, что слова, употребляющиеся в сходных контекстах, должны иметь сходный смысл [1]. К сожалению, данное утверждение не является истинным для всех слов языка. Существует большое количество примеров, соответствующих данной гипотезе, однако имеется и определенное число контрпримеров. Анализ результатов экспериментов, проводимых в этой области (см., например, [2–5]), позволяет утверждать, что в соответствии с данной гипотезой можно получить положительные результаты, качество и объем которых наталкиваются на некоторые ограничения (вероятно, вызванные особенностями естественных языков и сходством применяемых методов).

В настоящей статье мы изложим метод кластеризации глаголов, существительных и прилагательных, основанный на выделении из текста на русском языке связей вида «существительное + прилагательное» и «глагол (+ предлог) + существительное» с помощью набора конечных автоматов. Для выделения подобных групп был использован большой неаннотированный корпус текстов различной стилистики. Нашей задачей являлось создание метода автоматической кластеризации сходных слов русского языка в семантически однородные группы. При этом использовалась информация о синтаксической сочетаемости этих слов. Кластер считался семантически однородным в случае, когда все его слова относились к одному семантическому классу (в идеальном случае являлись синонимами).

СУЩЕСТВУЮЩИЕ ПОДХОДЫ К РЕШЕНИЮ ЗАДАЧИ

Успешное решение задачи классификации слов по семантическим группам позволяет повысить качество решений в различных областях автоматической обработки текстов. К таким областям относятся, например, автоматизация составления и пополнение онтологий и тезаурусов [6, 7], проведение синтаксического анализа [8, 9], выделение сочетаемостных свойств слов языка, используемых, например, при обучении иностранному языку, снятие омонимии [10] и целый ряд других задач. Одной из первых в данной области была работа Ф. Перейры и др. [11]. При построении онтологий и тезаурусов предварительная группировка слов позволяет проводить опе-

[1] Работа выполнена при финансовой поддержке грантов РГНФ № 12-04-00060 и РФФИ № 11-01-00793.

рации не над отдельными словами, а над их группами, что существенно повышает скорость работы. Наши эксперименты показали, что предварительная кластеризация слов позволяет повысить скорость работы специалиста, работающего над созданием тезауруса предметной области, в 2–3 раза. Кроме того, разделение слов на кластеры может использоваться как дополнительная сторонняя информация, при помощи которой проверяется связанность и корректность уже внесенной информации. При выполнении синтаксического анализа с помощью информации о сочетаемости слова или о его модели управления проверяется возможность установления связи между словами, а также уточняется роль слова в предложении. При снятии омонимии проверяется возможность разных вариантов слова формировать связи с другими словами. Использование данного метода как дополнения к уже существующим позволяет повысить качество снятия омонимии на 1–3% (что является хорошим результатом, так как современные системы, решающие эту задачу, преодолели порог в 95% и стремятся к 98%) [10]. Информация о принадлежности слова к некоторой семантической группе или о модели его управления используется как дополнительная оценка для каждого из вариантов, полученного от модуля лексического анализа слова. Также группирование слов проводится для выделения их модели управления [3, 4, 12], которая помогает в дальнейшем разделить семантические классы слов, устранить полисемию и выделить их возможные синтаксические связи. В работе [5] приведен весьма объемный и качественный обзор различных задач, которые могут быть решены при помощи анализа информации о сочетаемости слов.

Для проведения кластеризации необходимо определить меру близости между словами. В качестве данных, на основе которых рассчитывается мера близости, чаще всего берется информация о связях определенного слова с другими словами. М. Барони и А. Ленчи в работе [5] выделяют несколько видов связности между словами или, скорее, понятиями, обозначаемыми словами:

- атрибутивная связность, когда два слова, находящиеся, например, в отношении синонимии или гиперонимии, имеют одинаковые признаки или набор действий;
- связность по отношениям, когда два слова не принадлежат одному классу, а связываются между собой отношением, например, «часть–целое» (в этом случае они должны относительно часто встречаться в тексте вместе).

Следует заметить, что в случае атрибутивной связности два слова, являющиеся синонимами, но употребляющиеся в текстах разного стиля, могут иметь различный набор действий или объектов и субъектов. Так, например, слова «доктор» и «эскулап» обозначают сходные предметы, но второе слово, в отличие от первого, употребляется обычно в возвышенном или, напротив, уничижительном смысле. И наоборот, существуют слова с разным значением, но со сходным употреблением. Например, как «резервирование», так и «санация» могут быть «дополнительными», «заблаговременными», «стопроцентными», «полными» и «частичными». А для слов «газета» и «полиция» в русском языке нами было найдено более 300 прилагательных, употребляемых как с одним, так и с другим словом (в основном относящихся к географическим названиям или обозначающих принадлежность к тем или иным органам власти, эмоциональным оценкам). Вследствие этого мы не сможем полностью избавиться от ошибок при автоматической группировке слов. Однако, как было замечено выше, сходство выявляется корректно для большого списка слов.

Вне зависимости от типа используемой информации различают несколько методов ее выделения. Самыми простыми из них являются методы, основанные на модели «набор слов» (bag of words model). Методы этой группы предполагают, что если два слова появляются в тексте недалеко друг от друга достаточно часто, то они связаны друг с другом по смыслу [13]. Также в качестве входных данных здесь могут использоваться слова, входящие в одно предложение [14]. Последний метод прост в реализации, позволяет получить связи различного вида между словами, находящимися достаточно далеко друг от друга. С другой стороны, метод не позволяет получить результаты приемлемого качества. В связи с этим чаще используются результаты синтаксического анализа [5, 15]. В этом случае информация о связях характеризуется более высоким качеством, но времени на разбор текста требуется гораздо больше. Кроме того, при определении связей в тексте синтаксический анализ совершает 5–10% ошибок (что достаточно много), покрывая при этом практически 100% текста. Помимо информации о наличии связей между словами часть исследователей используют информацию о роли слова в предложении [5], выдаваемую семантическим анализом. Однако, как отмечалось выше, в ряде случаев роль слова не может быть корректно выявлена без привлечения семантической информации, получение которой и является нашей целью. Таким образом, мы приходим к замкнутому кругу.

Промежуточное положение занимают методы, основанные на лексических шаблонах, когда с помощью, например, регулярных выражений выделяется часть текста, синтаксическая связность которой может быть определена очевидным образом без привлечения синтаксического анализа. Данный метод дает более высокое качество определения синтаксических связей (около 95–98%), но при этом обеспечивает более низкое покрытие (около 20–40% текста). Кроме того, в большинстве языков встает проблема частеречной омонимии, которая не позволяет выделять связанные конструкции напрямую. Применение здесь автоматического снятия омонимии снижает качество результатов. Однако, как будет показано ниже, в русском языке имеется возможность преодолеть эту проблему.

Синтаксические связи выделяются для слов, относящихся к различным частям речи: глаголов [16], существительных [17] и прилагательных [18]. Большинство исследований в данной области посвящены английскому языку, но имеются работы, выполненные для русского [19], немецкого [20], каталонского [21] и других языков.

МЕТОД КЛАСТЕРИЗАЦИИ СЛОВ

Особенности входных данных

В качестве входных данных для метода кластеризации мы использовали информацию, извлеченную из большого неразмеченного корпуса (коллекции) текстов. В наших предыдущих работах [22] была собрана база данных, содержащая в себе информацию о синтаксической сочетаемости слов. Для того чтобы собрать указанную базу данных, мы применили поверхностный синтаксический анализатор, извлекающий именные, предложные и глагольные группы, используя для этого набор регулярных выражений. Из текста извлекались лишь контактные группы слов, для которых можно было достоверно получить синтаксические связи, не прибегая при этом к полному синтаксическому анализу. Приведем несколько примеров используемых регулярных выражений.

<s> (NP|PP) (adv)? VP – единственная предложная или именная группа, стоящая в начале предложения перед единственным глаголом в предложении (перед глаголом может идти причастие),

prep (adj)* noun – одно или более прилагательных, перед прилагательными находится предлог, после прилагательных – существительное, согласующееся с каждым из прилагательных.

Приведенные выражения обеспечивают высокую вероятность корректного определения синтаксических связей в предложениях на русском языке. В нашем проекте мы использовали лишь несколько видов полученных связей: <глагол, существительное>, <глагол, предлог + существительное> и <существительное, прилагательное>. В приведенных кортежах второй элемент подчиняется первому, но сами слова в исходном тексте могут следовать в произвольном порядке. Так, например, из предложения «В школу пришли новые учителя» будет извлечено три группы следующего вида <приходить + в + школа>, <приходить + учитель>, <учитель + новый>. Созданная база данных содержит полученные коллокации (при этом сами слова приведены к нормальной форме) и их частоты встречаемости.

На практике омонимичность слов не позволяет извлекать подобные сочетания с высокой точностью. Определение границ групп здесь становится затруднительным, так как отсутствует точная информация о части речи, которую представляет слово. Так, например, зачастую сложно определить, является ли рассматриваемое слово прилагательным, существительным или прилагательным в роли существительного. Как следствие, становится непонятно, следует ли рассматривать данное слово как прилагательное и продолжать выделять группу или необходимо принять, что оно существительное, и закончить разбор. Наши предыдущие исследования показали, что природа омонимии в русском языке позволяет использовать в данном случае группы неомонимичных слов. Дело в том, что в русском языке около 75–85% слов не омонимичны по части речи. По этой причине при разборе текста существует высокая вероятность обнаружить неомонимичную группу слов, отвечающую одному из синтаксических шаблонов, использовавшихся для сбора информации в базу данных зависимостей. В табл. 1 приведено сравнение омонимии в русском и английском языках (подробнее см. [23]). За счет высокого процента слов, не омонимичных по части речи, мы имеем возможность извлекать синтаксически связанные конструкции из большой части корпуса. Слова не обязаны быть неомонимичными по параметрам. Вне зависимости от значений лексических параметров мы можем констатировать факт наличия синтаксической связи и поместить сочетания в базу.

Применение методов снятия омонимии, основанных на использовании n-грамм [24, 25] или деревьев принятия решений [26], в данном случае не оправдано. Это связано с тем, что, несмотря на высокий уровень точности работы (до 97%), системы снятия омонимии вносят достаточное количество ошибок, чтобы повлиять на качество итоговых результатов. Наши эксперименты показали, что предложенный метод, основанный на частичном синтаксическом анализе неомонимичных групп слов, позволяет собирать информацию с высоким качеством: сочетания глагола с существительным содержат около 3% ошибок, существительного с прилагательными – около 1% ошибок. Самый большой процент ошибок наблюдался в сочетаниях <существительное + существительное в родительном падеже> – около 8% ошибок.

Заметим, что нам удалось понизить процент ошибок за счет применения модели управления глагола, а также проверки согласования между прилагательным и существительным. Несмотря на высокий уровень качества, покрытие в предложенном методе является низким (20–40% текста в зависимости от его стиля и предметной области). В связи с этим для получения представительных результатов нам пришлось использовать большой корпус текстов (несколько млрд слов).

Под ошибкой мы понимаем здесь ошибку работы предложенного метода или явные ошибки в тексте. Небольшая часть (около 1–2%) корректных с точки зрения метода сочетаний может вызывать сомнения с точки зрения стилистики или физической реализуемости. При этом извлеченные сочетания могут быть обусловлены окружающим текстом, являться выразительными приемами. Таким образом, полученная база в большей мере отражает узус, а не грамматические особенности языка.

Таблица 1

Сравнение видов омонимии в русском и английском языках

Вид омонимии	Лента.ру 2005–2009	Компьюлента 2001–2009	Reuters 2009
Неомонимичные	48,28%	46,55%	38,87%
Неизвестные	4,38%	4,28%	7,65%
Омонимичные по части речи	5,26%	6,33%	50,35%
Омонимичные по лексическим параметрам	27,68%	28,22%	2,79%
Омонимичная начальная форма	4,67%	4,01%	0,32%
Омонимичные по нескольким параметрам	9,92%	10,61%	0,96%

Мера сходства слов и вектор признаков кластеризации

Полученная база данных синтаксической сочетаемости слов может быть формально описана как множество кортежей $\mathbf{B}=\{<w_m, w_s, f>\}$, где w_m – главное слово, w_s – зависимое слово или часть предложной группы (предлог + существительное), а f – частота встречаемости данного сочетания в рассматриваемом тексте. В этом случае можно построить словарь главных слов $\mathbf{D}^+ = <w>$ (прямой словарь). Он будет содержать в себе множество всех главных слов w_m из базы данных **B**. Кроме того, можно построить словарь зависимых слов $\mathbf{D}^- = <w>$ (обратный словарь), который будет содержать в себе множество всех зависимых слов w_s из базы данных **B**. Словари $\mathbf{D}^+$ и $\mathbf{D}^-$ будут использоваться ниже как пространство признаков в ходе кластеризации (т. е. для расчета меры сходства будут использоваться частоты встречаемости слов из этих словарей).

Вектор признаков $\mathbf{v}^+_a$ слова a содержит в себе частоты встречаемости слова a со словами из обратного словаря. Данный вектор строится следующим образом: $\mathbf{v}^+_a = <v_1, v_2, \ldots, v_n>$, $n=|\mathbf{D}^-|$: $v_i = \{f_i$, если в **B** имеется запись вида $<a, d^-_i, f_i>$; 0 – в противном случае$\}$, здесь d^-_i это i-й элемент $\mathbf{D}^-$. Вектор признаков $\mathbf{v}^-_a$ для слова a определяется аналогичным образом, но с использованием частот для слов из обратного словаря $\mathbf{D}^+$. Мера сходства слов a и b определяется в таком случае с использованием косинусной меры сходства между векторами признаков $\mathbf{v}^+_a$ и $\mathbf{v}^+_b$ или между векторами признаков $\mathbf{v}^-_a$ и $\mathbf{v}^-_b$.

Эксперименты показали, что использование собственно частоты встречаемости приводит к большому количеству ошибок. Так, например, слова «Орлеан» и «Зеландия» будут весьма похожи, так как оба чаще всего встречаются в сочетании со словом «Новый» и гораздо реже с другими словами. Для того чтобы устранить подобную ситуацию предлагается использовать логарифм частоты встречаемости вместо самой частоты. Это уменьшает разницу между редко и часто используемыми словами, в результате чего они вносят примерно равный вклад в определение меры сходства.

Фильтрация данных и метод кластеризации

В целях снижения числа ошибок рекомендуется отфильтровать часть объектов, вносящих шум во входные данные. В связи с этим в своих экспериментах мы использовали только те сочетания, у которых частота встречаемости превосходила заданный порог. Кроме того, мы использовали сочетания только с теми главными словами, которые встретились более чем в заданном количестве различных комбинаций, т. е. число зависимых слов для которых превышало определенный порог. В терминах вектора признаков можно сказать, что число ненулевых значений в векторе превышает заданный порог. Математически формирование векторов признаков описывается следующим образом. Пусть $\mathbf{S}^+ = \{\mathbf{v}^+_a\}$, $a \in \mathbf{D}^+$. Тогда $\mathbf{v}^+_a = <v_1, v_2, \ldots, v_n>$, $n=|\mathbf{D}^-|$: $v_i = \{f_i$, если в **B** имеется запись вида $<a, d^-_i, f_i>$ и $f_i \geq \alpha$, 0 – в противном случае$\}$, здесь d^-_i это i-й элемент $\mathbf{D}^-$. При этом $\mathbf{v}^+_a \in \mathbf{S}^+$ только если $NZCount(\mathbf{v}^+_a) > \beta$, где функция $NZCount(\mathbf{v}^+_a)$ возвращает количество ненулевых элементов $\mathbf{v}^+_a$. Аналогичным образом строится множество векторов $\mathbf{v}^-_a$: $\mathbf{S}^- = \{\mathbf{v}^-_a\}$, $a \in \mathbf{D}^-$. Тогда $\mathbf{v}^-_a = <v_1, v_2, \ldots, v_n>$, $n=|\mathbf{D}^+|$: $v_i = \{f_i$, если в **B** имеется запись вида $<d^+_i, a, f_i>$ и $f_i \geq \alpha$, 0 – в противном случае$\}$, здесь d^+_i это i-й элемент $\mathbf{D}^+$. При этом $\mathbf{v}^-_a \in \mathbf{S}^-$ только если

$$NZCount(\mathbf{v}^-_a) > \beta.$$

На первом шаге метода кластеризации рассчитывается матрица расстояний, основанная на расчете косинусной меры сходства. В зависимости от используемого словаря для расчета используются различные векторы признаков: $cos(\mathbf{v}^+_a, \mathbf{v}^+_b)$ или $cos(\mathbf{v}^-_a, \mathbf{v}^-_b)$, $a \neq b$. Из данной матрицы извлекается n самых близких (максимальных) значений или все значения, превосходящие заданный порог γ. Для определения расстояния между кластерами применяется модифицированный метод средней связи [2]. Исходный метод средней связи заключается в том, что расстояние между кластерами рассчитывается как среднее расстояние между всеми элементами двух кластеров. В нашем случае рассчитывалось среднее расстояние между каждым элементом каждого кластера и центроидом второго кластера.

На начальном шаге кластеризации каждое слово считается отдельным кластером. Далее на каждом шаге кластеризации выбираются два самых близких кластера, после чего проводится их объединение в один с пересчетом расстояний до других кластеров. Алгоритм кластеризации останавливается, когда расстояние между объединяемыми кластерами становится меньше заданного порога (напомним, что меньшее значение косинусной меры соответствует большему расстоянию между кластерами). Данное пороговое значение неявно задает количество получаемых кластеров. В случае слишком малого значения порога все слова будут объединены в кластеры, среди которых будут встречаться достаточно большие (до 10–15 слов). Большие кластеры при этом могут объединять в себе несколько более мелких, не связанных между собой по смыслу. Слишком большое значение порога приведет к тому, что будут кластеризованы не все слова, однако качество кластеров будет выше. Тогда возможно применение иерархических методов кластеризации [27]. В этом случае слова образуют бинарное дерево (так как на каждом шаге объединяется ровно два кластера), и перед нами встанет новая проблема – определение границ кластеров.

ОПИСАНИЕ ВХОДНЫХ ДАННЫХ И МЕТОДА ОЦЕНКИ РЕЗУЛЬТАТОВ

Для своих экспериментов мы использовали неразмеченный корпус текстов на русском языке объемом 7,2 млрд слов, состоящий из беллетристики (около 6,6 млрд слов), новостных (около 550 млн слов) и научных текстов (статьи, тексты диссертаций, отчеты – всего около 50 млн слов). Как это было описано выше, синтаксически связанные комбинации слов собирались при помощи набора регулярных выражений, действующих по заданным шаблонам. Морфологическая разметка проводилась с помощью

анализатора «Кросслейтор» [28]. База данных синтаксически связанных комбинаций слов для русского языка составила более 23 млн различных глагольных групп (<глагол + существительное> или <глагол + предлог + существительное>) и около 5,5 млн различных именных групп (существительное + прилагательное). В базу данных попали группы как с прямым, так и с обратным порядком слов, однако все они были нормализованы.

Как и в работе [29], для оценки результатов плоской кластеризации мы использовали формулу нормализованной взаимной информации NMI(A, B) и F_1-меру [2].

$$NMI(A,B) = \frac{I(A,B)}{\left[H(A)+H(B)\right]/2};$$

$$I(A,B) = \sum_k \sum_j \frac{\left|v_k \cap c_j\right|}{N} \log \frac{N\left|v_k \cap c_j\right|}{\left|v_k\right|\left|c_j\right|},$$

где $|v_k \cap c_j|$ – количество общих элементов в кластере v_k и классе золотого стандарта c_i, $H(A)$ – энтропия кластеров, $H(B)$ – энтропия классов золотого стандарта. F_1-мера представляет собой удвоенное среднегармоническое значение от покрытия (recall) и точности (precision). Энтропия определяется по формуле, предложенной Шенноном: $H(x) = -\sum_{i=1}^{n} p(i) * \log_2 p(i)$, где $p(i)$ – вероятность вхождения элемента в кластер: $p(i) = n / N$, n – число элементов в кластере, N – общее число элементов.

Под золотым стандартом обычно понимается набор классов, разбиение для которых задано заранее, например, вручную [2]. В нашей работе в качестве золотого стандарта мы использовали словарь [30]. Нами было взято 97 глаголов движения, приобретения и найма. Кроме того, для того чтобы внести некоторый шум в исходные данные, было добавлено еще 49 глаголов из других групп. Все глаголы были разделены на 25 классов: 14 из них были взяты из [30], 9 были получены вручную по таким естественным параметрам, как переходность/непереходность или совершённость/несовершённость действия.

РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ ДЛЯ РУССКОГО ЯЗЫКА

В ходе экспериментов с корпусом текстов на русском языке мы использовали следующие пороговые значения: частота встречаемости сочетания в корпусе не ниже 5; количество слов, подчиняющихся данному – не менее 7; алгоритм останавливается при расстоянии между кластерами менее 0,3. Для выбранных 146 глаголов значение NMI составило 0,457, а F_1-мера составила 0,607.

В ходе экспериментов мы не могли оценить качество всех полученных данных с помощью золотого стандарта, так как расхождения между лексикой, полученной в экспериментах, и лексикой золотого стандарта слишком велики. Так, например, из рассматриваемых текстов было извлечено достаточно много специфичной лексики, отсутствующей в выбранном в качестве золотого стандарта словаре. В связи с тем, что для русского языка отсутствует достаточно большая (50–70 тыс. слов) открытая онтология (такая, как, например, WordNet [31]), мы приняли решение провести проверку полученных результатов при помощи асессора.

Вместо того чтобы выбирать все пары слов с мерой сходства, превышающей заданное значение, мы отобрали 7000 пар с максимальным сходством. Прямой словарь для глаголов составил в этом случае 4200 различных слов, распределенных по 1550 кластерам. В данном эксперименте все слова были распределены по кластерам, т. е. покрытие составило 1. Оцененная асессором точность составила 0,79. Таким образом, F_1-мера равна 0,88.

Мы также провели эксперименты для существительных и прилагательных. Для существительных имелась возможность применить как прямой (существительное + прилагательное), так и обратный (глагол + предлог, существительное) словарь. Полученные кластеры также были проверены асессором. Для обратного словаря покрытие также равнялось 1, точность составила 0,85, F_1-мера была равна 0,92. Для прямого словаря мы получили покрытие, равное 0,85, точность – 0,88 и F_1-меру – 0,85. Эксперименты с прилагательными показали покрытие 0,85, точность 0,96 и F_1-меру 0,9. Результаты приведены в табл. 2.

Результаты, полученные другими авторами, приведены в табл. 3 (курсив означает, что значение F_1-меры было рассчитано нами на основе приведенных в работах данных). Как видно из приведенных данных, полученные нами результаты вполне согласуются с данными аналогичных работ (здесь мы не учитываем разницу между языками).

Таблица 2

Результаты экспериментов с прямым и обратным словарями

Вид связи	Покрытие	Точность	F_1-мера
<Глагол + предлог, существительное>	1	0,79	0,88
<Существительное + прилагательное>	0,85	0,88	0,85
< Существительное + предлог, глагол >	1	0,85	0,92
<Прилагательное + существительное >	0,85	0,96	0,9

Таблица 3

Сравнение результатов экспериментов

Работа	NMI	F_1-мера
[29]	0,378 – 0,573	0,367 – 0,4
[16]	–	*0,78*
[32]	–	0,46
[33]	–	*0,36*
Данная статья	0,457	0,607

Л. Сан и А. Корхонен в работе [29] получили значение меры NMI, находящееся между 0,378 и 0,573, а также значение F_1-меры, находящееся между 0,367 и 0,4. В работе [16] не приведено значение F_1-меры, однако, согласно приведенным данным, она может быть оценена как равная 0,78. Для арабского языка в работе [32] получена F_1-мера, равная 0,46. В работе [33], где F_1-мера также не рассчитывалась, она равна примерно 0,36. Для всех рассмотренных языков средний размер кластера достаточно мал (2–4 слова). В нашей работе мы получили значение меры NMI равное 0,457 и F_1-меру, равную 0,607. Заметим, что во всех экспериментах для сравнения использовался золотой стандарт. Относительно высокое значение F_1-меры, полученное в наших экспериментах, может быть объяснено значительно бóльшим размером использованного корпуса, а также особенностями русского языка. Также заметим, что методы кластеризации разнятся от работы к работе: начиная с простой косинусной меры и заканчивая выделением ведущих элементов при помощи метода LSA.

ЗАКЛЮЧЕНИЕ

Главным минусом изложенного подхода к кластеризации слов является малый размер получаемых кластеров. Это может быть объяснено ограничениями, накладываемыми на начальную гипотезу о корреляции семантического сходства слов с лексикой, связываемой с этими словами синтаксическими связями, а также особенностями естественных языков.

1. В языке присутствуют слова, которые имеют общий смысл, но обладают различной сочетаемостью с другими словами. Так, например, слова «врач» или «доктор» обладают очень широким и сходным словоупотреблением. Однако слово «эскулап» встречается гораздо реже и в последнее время носит иронический оттенок, если употребляется с определениями. Количество присоединяемых к нему прилагательных значительно меньше и степень сходства со словами «врач» и «доктор» в связи с этим невысока. Сходная ситуация наблюдается и при анализе присоединяемых глаголов, хотя здесь для слова «эскулап» преобладает возвышенная тональность. Как следствие, указанные три слова не могут быть объединены в один кластер предложенным методом. Справедливо и обратное. Так для слов «газета» и «полиция» в русском языке нами было найдено более 300 прилагательных, употребляемых как с одним, так и с другим словом (в основном относящихся к географическим названиям, принадлежности тем или иным органам власти, эмоциональным оценкам).

2. В случае анализа полисемичных слов множество слов, зависимых от данного, может быть разделено на несколько подмножеств. В связи с этим два полисемичных слова, сходных лишь по одному из значений, будут иметь лишь по одному совпадающему подмножеству. Как следствие, их мера сходства окажется весьма мала. С другой стороны, слова, имеющие несколько сходных значений, будут объединяться в первую очередь. Подобный подход позволяет, однако, решить две другие задачи: выделение сходных значений в полисемичных словах и выделение групп сходных признаков.

3. Значения лексических параметров играют очень важную роль в русском языке. Как уже отмечалось выше, все слова в нашей модели приводились к нормальной форме. Однако в модели глагольного управления синтаксическая роль в русском языке задается, помимо глагола, к которому присоединяется слово, или предлога, при помощи которого проводится присоединение, еще и посредством падежей. В связи с этим в нашем случае два глагола, имеющие различный набор валентностей, могут быть объединены в один кластер. Аналогичное можно утверждать и для других частей речи.

Следует заметить, что все проанализированные методы обладают сходными недостатками (с учетом специфики рассматриваемого языка). В большинстве рассмотренных работ средний размер кластера равен примерно 3, увеличение размера кластера приводит к потере точности. Тот факт, что слово присоединяет к себе различные слова в зависимости от того, в каком значении оно используется, успешно используется рядом авторов для автоматического выделения этих значений [34, 35].

Указанные выше фундаментальные особенности языков не позволяют решить задачу автоматического построения кластеров семантически связанных слов. Получаемые кластеры слишком малы. Размер кластеров, как и объем обрабатываемой лексики, по всей видимости, может быть немного увеличен за счет увеличения количества анализируемых текстов и расширения их тематики. При этом, однако, не следует рассчитывать на существенное изменение ситуации.

Возможным представляется применение данной группы методов для начальной разметки кластеров с ее последующим постредактированием. Проведенные нами эксперименты показывают, что при классификации сходных слов вручную обработка предварительно выделенных кластеров позволяет повысить скорость работы оператора в 2–3 раза.

В дальнейшем мы планируем использовать нечеткие методы кластеризации, которые позволяют поместить одно и то же слово в несколько кластеров (см., например, [36]). Это должно увеличить размер кластеров и помочь учитывать полисемию слов. Учет лексических параметров должен улучшить качество кластеризации слов русского языка. Однако нам представляется, что совокупность указанных улучшений не позволит вывести качество решения задачи на кардинально иной уровень. Для этого требуется разработка новых методов обработки имеющейся информации.

СПИСОК ЛИТЕРАТУРЫ

1. Harris Z. Distributional structure // Word. – 1954, 10 (23). – P. 146–162.
2. Маннинг К., Рагхаван П., Шютце Х. Введение в информационный поиск. – М.: Вильямс, 2008. – 528 с.
3. Korhonen A., Krymolowski Yu., Marx Z. Clustering Polysemic Subcategorization Frame Distributions Semantically // Proceedings of the 41st an-

nual meeting of the Association for Computational Linguistics, 2003.

4. Korhonen A., Krymolowski Yu., Briscoe T. A Large Subcategorization Lexicon for Natural Language Processing Applications // Proceedings of the 5th international conference on Language Resources and Evaluation, 2006.
5. Baroni M. and Lenci A. Distributional Memory: A general framework for corpus-based semantics // Computational Linguistics. – 2010. – Vol. 36 (4). – P. 673–721.
6. Волкова Г.А. Создание «онтологии всего». Проблемы классификации и решения // Сб. трудов научно-практического семинара «Новые информационные технологии в автоматизированных системах». – 2013. – С. 293–300.
7. Kipper K., Korhonen A., Ryant N., and Palmer M. Extending VerbNet with Novel Classes // Proceedings of the 5th International Conference on Language Resources and Evaluation. Genoa. Italy.
8. Гельбух А. Разрешение синтаксической неоднозначности и извлечение словаря моделей управления из корпуса текстов // Материалы VIII Международной конференции KDS-99.
9. Мальковский М.Г., Арефьев Н.В. Сочетаемостные ограничения в системе автоматического синтаксического анализа // Программные продукты и системы. № 1. – Тверь, 2012. – С. 28–31.
10. Клышинский Э.С., Кочеткова Н.А., Литвинов М.И., Максимов В.Ю. Метод разрешения частеречной омонимии на основе применения корпуса синтаксической сочетаемости слов в русском языке // Научно-техническая информация. Сер. 2. – 2011. – № 1. – С. 31–35.
11. Pereira F., Tishby N., Lee N. Distributional Clustering of English Words // Proceedings of ACL-93.
12. Manning C. Automatic acquisition of a large subcategorization dictionary from corpora // Proceedings of the 31st annual meeting on Association for Computational Linguistics.
13. Li H. and Abe N. Word Clustering and Disambiguation Based on Co-occurrence Data // Proceedings of COLING-ACL'98. – P. 749–755.
14. Bassiou N., Kotropoulos C. Long distance bigram models applied to word clustering // Pattern Recognition. – 2011. – Vol. 44, Issue 1. – P. 145–158.
15. Preiss J., Briscoe T., and Korhonen A. A system for large-scale acquisition of verbal, nominal and adjectival subcategorization frames from corpora // Proceedings of ACL, 2007. – P. 912–919.
16. Schulte im Walde S. Clustering Verbs Semantically According to their Alternation Behaviour // Proceedings of the 18th International Conference on Computational Linguistics, 2000.
17. Hindle D. Noun classification from predicate-argument structures // Proceedings of ACL-90.
18. Boleda Torrent G. and Alonso i Alemany L. Clustering Adjectives for Class Acquisition // Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics, 2003.
19. Митрофанова О.А., Мухин А.С., Паничева П.В. Автоматическая классификация лексики в русскоязычных текстах на основе латентного семантического анализа // Компьютерная лингвистика и интеллектуальные технологии : Труды международной конференции «Диалог-2007». – М.: Изд-во РГГУ, 2007.
20. Bohnet B., Klatt S., and Wanner L. A Bootstrapping approach to automatic annotation of functional information to adjectives with an application to German // Proceedings of the Workshop on Linguistic Knowledge Acquisition and Representation at the 3rd LREC Conference, 2002.
21. Boleda G., Badia T., Schulte im Walde S. Morphology vs. Syntax in Adjective Class Acquisition // Proceedings of the ACL-SIGLEX Workshop on Deep Lexical Acquisition, 2005. – P. 77–86.
22. Клышинский Э.С., Кочеткова Н.А., Литвинов М.И., Максимов В.Ю. Автоматическое формирование базы сочетаемости слов на основе очень большого корпуса текстов // Компьютерная лингвистика и интеллектуальные технологии : По материалам ежегодной Международной конференции «Диалог» (Бекасово, 26–30 мая 2010 г.). Вып. 9 (16). – М.: Изд-во РГГУ, 2010. – С. 181–185.
23. Клышинский Э.С., Кочеткова Н.А. Метод автоматической генерации модели управления глаголов русского языка // Сб. трудов 12 национальной конференции по искусственному интеллекту КИИ-2012 (Белгород, 16–20 сентября 2012 г.). Т. 1. – Белгород: изд-во БГТУ, 2012. – С. 227–235.
24. Brill E. Unsupervised Learning of Disambiguation Rules for Part of Speech Tagging // Proceedings of the Third Workshop on Very Large Corpora. Cambridge, USA, 1995.
25. Зеленков Ю.Г., Сегалович И.В., Титов В.А. Вероятностная модель снятия морфологической омонимии на основе нормализующих подстановок и позиций соседних слов // Компьютерная лингвистика и интеллектуальные технологии. Труды международного семинара «Диалог-2005».
26. Schmid H. Probabilistic Part-of-Speech Tagging Using Decision Trees // Proceedings of International Conference on New Methods in Language Processing. Manchester, 1994.
27. Sokal R.R. and Rohlf F.J. The comparison of dendrograms by objective methods. – Taxon, 1962.
28. Система морфологического анализа “Crosslator”. [Электрон. ресурс]. – URL: http://clschool.miem.edu.ru/ Материалы-школы.html.

29. Sun L. and Korhonen A. Hierarchical Verb Clustering Using Graph Factorization // Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP). Edinburgh, UK.

30. Русский семантический словарь. Толковый словарь, систематизированный по классам слов и значений / Ин-т рус. яз. им. В. В. Виноградова; под общ. ред. Н. Ю. Шведовой. – М.: Азбуковник, 1998.

31. Fellbaum F. WordNet: An Electronic Lexical Database. – Cambridge, MA: MIT Press, 1998.

32. Snider N. and Diab M. Unsupervised Induction of Modern Standard Arabic Verb Classes Using Syntactic Frames and LSA // Proceedings of ACL, 2006. – P. 795–802.

33. Sass B. First Attempt to Automatically Generate Hungarian Semantic Verb Classes // Proceedings of the 4th Corpus Linguistics conference. Birmingham, 2007.

34. Yarowsky D. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods // Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics. Cambridge, MA, 1995. – P. 189–196.

35. Толдова С.Ю., Кустова Г.И., Ляшевская О.Н. Семантические фильтры для разрешения многозначности в Национальном корпусе русского языка // Компьютерная лингвистика и интеллектуальные технологии. Диалог-2008. – М., 2008. – С. 522–529.

36. Yang M.-S. A Survey of Fuzzy Clustering // A survey of fuzzy clustering, Mathematical and Computer Modelling. – 1993, 18(11). – P. 1–16.

Материал поступил в редакцию 10.07.13.

Сведения об авторах

КЛЫШИНСКИЙ Эдуард Станиславович – кандидат технических наук, доцент Московского института электроники и математики Национального исследовательского университета «Высшая школа экономики» (МИЭМ НИУ ВШЭ); старший научный сотрудник Института прикладной математики им. М. В. Келдыша РАН, Москва
e-mail: eklyshinsky@hse.ru

КОЧЕТКОВА Наталия Александровна – аспирант Московского института электроники и математики Национального исследовательского университета «Высшая школа экономики», Москва
e-mail: natalia_k_11@mail.ru

ЛОГАЧЕВА Варвара Константиновна – кандидат физико-математических наук, научный сотрудник Института прикладной математики им. М. В. Келдыша РАН, Москва
e-mail: logacheva_vk@mail.ru