УДК 519.6:410

АССОЦИАТИВНЫЕ ПОЛЯ ПРЕДМЕТНЫХ ОБЛАСТЕЙ

О. Г. Чанышев, А. М. Пуртов

Омский филиал Института математики им. С. Л. Соболева СО РАН

Поступила в редакцию 02.04.2013 г.

Аннотация. Вводится понятие «контекстного ассоциативного поля слова». На его основе строятся ассоциативные сети текстов и предметных областей (ПО), представленных множествами текстов. С их помощью определяются попарные близости ПО и близости заданного термина к ПО.

Ключевые слова: ассоциативное поле, сеть ассоциативных полей, обратная ассоциативная сеть.

Annotation. The concept of " the contextual associative field of a word» is entered. Associative networks of texts and subject domains (SD) (presented by sets of texts) are construction on this basis. Paired affinity SD and affinity any term to SD are determined with their help.

Keywords: associative field, network of associative fields, backhanded associative network.

ВВЕДЕНИЕ

Наиболее общим подходом, отправной точкой к построению алгоритмов автоматического анализа естественно-языковых текстов является, по нашему мнению, учение о вербальных ассоциациях. Поскольку «...слова, служащие для обозначения ... предметов, ассоциированы между собою в языке и, конечно, потому, что соответствующие предметы ассоциированы в человеческом сознании» [1]. Таким образом, в вербальных ассоциациях отражены экстралингвистические явления, что ограничивает область применения чисто лингвистических моделей: «авторы лингвистических моделей зачастую явно или неявно апеллируют к языковой интуиции человека, носителя описываемого языка, опуская ряд «очевидных» деталей, чрезвычайно существенных при автоматической обработке текста» [2]. В работе Г. А. Мартиновича [3] отмечается, что «...многие проблемы словесных ассоциаций, процессов ассоциирования как в естественном речевом общении людей, так и в эксперименте остаются все еще недостаточно изученными: до сих пор нет единого определения самих вербальных ассоциаций, их однородной (научной) классификации, не установлены их лингвистический и психолингвистический статус, отношение к ряду традиционных лингвистических категорий и т. п.»

Одним из центров внимания исследователей вербальных ассоциаций находятся ассоциативные поля слов. Они получаются в ходе ассоциативного эксперимента и используются «для экспериментального исследования субъективных семантических полей слов, формируемых и функционирующих в сознании человека, а также характера семантических связей слов внутри семантического поля» [4].

В статье [5] изложены основные положения психо- и нейролингвистики «Ассоциативной модели реального текста», дающие основания рассматривать слово как логический адрес нейронного ансамбля, предложение как элементарный способ образования или вызова ассоциаций, а текст в целом как «программу» закрепления ассоциаций и установления связей между ними. Множество слов, встречающихся с некоторым фиксированным словом хотя бы в одном предложении текста или группы текстов, явно принадлежащих одной предметной области, естественно назвать контекстным ассоциативным полем этого слова.

1. АССОЦИАТИВНЫЕ ПОЛЯ И АССОЦИАТИВНЫЕ СЕТИ СЛОВ

Рассматриваются слова, с повторяемостью в тексте не менее 2 раз и не принадлежащих словарю стоп-слов.

Положим Q_i, Q_j – области существования слов l_i и l_j $(i \neq j)$. В зависимости от решаемых задач они могут быть представлены либо номерами включающих предложений, либо множествами слов предложений. ρ – размер множества.

Слово l_j принадлежит к ассоциативному полю слова l_i если

$$K_{i,j} = \frac{\rho(Q_i \cap Q_j)}{\rho(Q_j)}.$$

Если $0 < K_{i,j} < 0.5$, то j-ое слово является элементом полного ассоциативного поля i-го, если же $K_{i,j} > 0.5$, то j-ое слово является элементом основного ассоциативного поля i-го. Множество общих элементов пары полных полей рассматривается в качестве ребер, соединяющих два этих поля. Определив все пересечения между ассоциативными полями, получаем сеть ассоциативных полей.

Основные ассоциативные поля оказались полезными для генерирования частично связных «квазирефератов» текстов [6]. Однако, с точки зрения построения естественно-языковых интерфейсов к базам знаний в качестве первичной («входной») сети удобнее использовать обратные ассоциативные поля.

$$K_{i,j} = \frac{\rho(Q_i \cap Q_j)}{\rho(Q_i)}.$$

Т.е., если с «точки зрения» вершины ассоциативного поля в нем собраны все «кто близок ко мне», то при вершине обратного поля собраны все «к кому я близка».

В обратном ассоциативном поле меру близости можно рассматривать и как меру «проводимости возбуждения» при моделировании мышления.

Ассоциативные поля предметных областей получаются простым объединением ассоциативных полей текстов, представляющих предметную область. Каждому слову соответствует запись следующей структуры:

<вершина_слово><список_ассоциированных_слов>

<список_ассоциированных_слов>::=

{<слово>(<номер_файла>,<близость>)}

Рассматриваются следующие предметные области:

Системы управления базами данных (BD) – (13 текстов)

Сетевые операционные системы (NetOpSyst) – (10 текстов)

Искусственный интеллект (AI) – (18 текстов)

Психология (Psih) – (18 текстов)

Философия (FLS) – (47 текстов)

2. БЛИЗОСТИ ПО

Определять близости ПО через функцию от множества пересечения их слов достаточно бессмысленно, поскольку пересечения будут состоять преимущественно из общеупотребительных слов. Однако можно ожидать иного результата, если использовать множества вершин основных ассоциативных полей.

Положим T_i, T_j – множества вершин основных ассоциативных полей в i-ой и j-ой предметных областях.

Определим близость пары ПО следующим образом:

$$B_{i,j} = \frac{\rho(T_i \cap T_j)}{\rho(T_i) + \rho(T_j)}.$$

Для данного состава ПО результат упорядочения убыванию значения $B_{i,j}$ выглядит в соответствии с интуитивным представлением о лексико-семантических близостях ПО:

BD∩ NetOpSyst 0.144
AI∩ BD 0.113
Psih∩ FLS 0.104
AI∩ Psih 0.092
AI∩ NetOpSyst 0.090
AI∩ FLS 0.0845
BD∩ Psih 0.088
BD∩ FLS 0.068
Psih∩ NetOpSyst 0.066
FLS∩ NetOpSyst 0.054

3. ОПРЕДЕЛЕНИЕ ПО, НАИБОЛЕЕ ПОЛНО ПРЕДСТАВЛЯЮЩЕЙ СЛОВО

В названной задаче главной проблемой является выбор критерия определения ПО. По нашему мнению, в критерии должны быть учтены связи заданного слова с другими словами текстов, а не просто частота слова. Для подтверждения этого положения была разработана программа, определяющая близость ПО к слову по двум критериям:

а) по сумме частот заданного слова по всем текстам каждой ПО (критерий Ω),

б) по сумме частот всех возбужденных слов (вершин) по всей цепочке ассоциаций от заданного слова для всех текстов каждой ПО (критерий Ω^+).

Цепочка ассоциаций слова и сумма частот возбужденных вершин.

Пусть $A(W) = (w_1, w_2, ..., w_i, ..., w_n)$ – множество слов, к которые прямо ассоциированы с W (встречаются в одном предложении) во всех текстах ПО.

Слова частично упорядочены по убыванию близости и близость W к w_i не должна быть меньше, чем W к w_1 (в таком случае полагается, что слово-вершина возбуждена) Далее в цепочку ассоциаций включаются прямые ассоциации всех w_i. Для каждого слова из цепочки ассоциаций вычисляется частота. Частоты суммируются.

Сумма частот заданного слова: Для заданного слова суммируются его частоты в каждом тексте ПО.

3.1. ИДЕЯ ЭКСПЕРИМЕНТА

Выбираются грамматические формы некоторого слова, заведомо имеющего широкое распространение в качестве термина в рассматриваемом множестве ПО. Расчитывем критерии Ω и Ω^+ для каждого элемента грамматической парадигмы каждой из ПО. Упорядочиваем результаты по убыванию значений критериев. В полученных распределениях для каждого элемента парадигмы каждой ПО присваиваем балл равный обратному порядковому номеру в распределении (если число ПО = 5, то первому месту присваивается балл 5, последнему – 1). Суммируем баллы. Полагаем, что ПО с максимальным баллом наиболее полно представляет заданное слово.

Если результаты для двух критериев существенно различны, то в предметную область, набравшую наименьшее количество баллов по одному из критериев, добавляем тексты, заведомо посвященные раскрытию «смысла» термина в рамках данной ПО и смотрим, как изменится результат.

3.2. ЭКСПЕРИМЕНТ

Для множества наших предметных областей одним из наиболее общих является термин «система». В качестве элементов грамматической парадигмы выбираем «система», «системе», «системы», «системой».

Результаты эксперимента с исходными текстами представлены в таблице 1.

Таблица 1

Предметные области с начальными текстами по психологии

По убыванию суммы частоты вершины (критерий Ω)	По убыванию суммы частот всех возбужденных вершин (критерий Ω^+)
--- Вершина <система> ---	--- Вершина <система> ---
NetOpSyst 0.372 5	FLS 23.821 5
AI 0.312 4	AI 23.088 4
BD 0.254 3	BD 6.686 3
FLS 0.198 2	NetOpSyst 0.060 2
Psih 0.064 1	Psih 0.058 1
--- Вершина <системе> ---	--- Вершина <системе> ---
FLS 0.239 5	FLS 23.821 5
NetOpSyst 0.199 4	AI 23.088 4
BD 0.121 3	BD 6.686 3
AI 0.078 2	NetOpSyst 0.331 2
Psih 0.061 1	Psih 0.050 1
--- Вершина <системы> ---	--- Вершина <системы> ---
AI 0.976 5	Psih 25.984 5
NetOpSyst 0.757 4	FLS 23.821 4
BD 0.703 3	BD 0.300 3
FLS 0.418 2	NetOpSyst 0.244 2
Psih 0.052 1	AI 0.015 1
--- Вершина <системой> ---	--- Вершина <системой> ---
NetOpSyst 0.121 5	FLS 23.821 5
BD 0.075 4	AI 23.088 4
AI 0.054 3	BD 6.686 3
FLS 0.044 2	NetOpSyst 0.322 2
Psih 0.002 1	Psih 0.060 1
Итого по сумме баллов	Итого по сумме баллов
NetOpSyst 5+4+4+5=18	FLS 5+5+4+5=19
AI 4+2+5+3=14	AI 4+4+1+4=13
BD 3+3+3+4=13	BD 3+3+3+3=12
FLS 2+5+2+2=11	NetOpSyst 2+2+2+2= 8
Psih 1+1+1+1= 4	Psih 1+1+5+1= 8

Как видим, последнее место в распределениях по обоим критериям занимает ПО «Психология». Первое место по критерию **Ω** занимает ПО NetOpSyst («Сетевые операционные системы», а по критерию Ω^+ – FLS («Философия»).

Добавление к подборке «Психология» 6-ти глав монографии Ганзена выводит эту ПО на второе место по критерию Ω^+, оставляя «философию» на первом месте (Таблица 2). На критерий **Ω** сделанные изменения фактически не повлияли, изменились только количественные показатели.

4. ВЫВОДЫ

Результаты экспериментов, представленных в настоящей работе и в статье «Автореферирование на основе ассоциативных полей доминант» [6], дают основание полагать, что концепция «контекстных ассоциативных полей слов» может претендовать на единую методологическую основу автоматического анализа естественно-языковых текстов.

СПИСОК ЛИТЕРАТУРЫ

1. *Покровский М.М.* Избранные работы по языкознанию. – М., Изд-во АН СССР, 1959, С. 382.

2. *Сулейманов Д. Ш.* Аналитический обзор отечественных и зарубежных работ обработки естественного языка в аспекте прагматически-ориентированного подхода. \ Электронный журнал Казанского госуниверситета „Информационные технологии“, Казань – 1999. http://www.kcn.ru/tat_en/science/ittc/vol000/st.doc, 19.06.2012.

3. *Мартинович Г. А.* Вербальные ассоциации в ассоциативном эксперименте. СПб., 1997. // Электронный ресурс, http://rudocs.exdat.com/docs/index-130837.html.

Таблица 2

Предметные области с добавленными текстами по психологии (к Psih добавлены 6 глав монографии Ганзена «Системные описания в психологии»)

По убыванию суммы частот вершины	По убыванию суммы частот всех возбужденных вершин
--- Вершина <система> ---	--- Вершина <система> ---
NetOpSyst 0.372 5	Psih_G 27.266 5
AI 0.312 4	FLS 23.821 4
BD 0.254 3	AI 23.088 3
FLS 0.198 2	BD 6.686 2
Psih_G 0.161 1	NetOpSyst 0.061 1
--- Вершина <системе> ---	--- Вершина <системе> ---
FLS 0.239 5	FLS 23.821 5
NetOpSyst 0.199 4	AI 23.088 4
BD 0.122 3	BD 6.686 3
Psih_G 0.120 2	NetOpSyst 0.331 2
AI 0.078 1	Psih_G 0.050 1
--- Вершина <системы> ---	--- Вершина <системы> ---
AI 0.976 5	Psih_G 27.266 5
NetOpSyst 0.757 4	FLS 23.821 4
BD 0.703 3	BD 0.301 3
FLS 0.418 2	NetOpSyst 0.244 2
Psih_G 0.379 1	AI 0.015 1
--- Вершина <системой> ---	--- Вершина <системой> ---
NetOpSyst 0.121 5	Psih_G 27.266 5
BD 0.075 4	FLS 23.821 4
AI 0.054 3	AI 23.088 3
FLS 0.044 2	BD 6.686 2
Psih_G 0.022 1	NetOpSyst 0.322 1
Итого по сумме баллов	Итого по сумме баллов
NetOpSyst 5+4+4+5=18	FLS 4+5+4+4=17
AI 4+1+5+3=13	Psih_G 5+1+5+5=16
BD 3+3+3+4=13	AI 3+4+1+3=11
FLS 2+5+2+2=11	BD 2+3+3+2=10
Psih_G 1+2+1+1= 5	NetOpSyst 1+2+2+1= 6

4. *Глухов В.П.* Основы психолингвистики: учеб. пособие для студентов педвузов. М.: АСТ: Астрель, 2005. // Электронный ресурс, http://www.pedlib.ru/Books/4/0356/4_0356-299.shtml.

5. *Чанышев О.Г.* Ассоциативная модель естественноязыкового текста. // Вестник Омского государственного университета, вып. 4, 1997 г.,Омск: ОмГУ, – 1997. – С. 17–20.

6. *Чанышев О.Г.* Автореферирование на основе ассоциативных полей доминант // Вестник Омского университета, № 4, 2011, С. 50–54.

Чанышев О. Г. – к. т. н., старший научный сотрудник, Омский филиал Института математики им С. Л. Соболева СО РАН. E-mail: fedorov22@yandex.ru

Пуртов А. М. – к. т. н., старший научный сотрудник, Омский филиал Института математики им С. Л. Соболева СО РАН.

Chanyshev O. G. – candidate of technical sciences, Senior Researcher of the Omsk Branch of the Institute of Mathematics of the Siberian Branch of the Russian Academy of Sciences. E-mail: fedorov22@yandex.ru

Purtov A. M. – candidate of technical sciences, Senior Researcher of the Omsk Branch of the Institute of Mathematics of the Siberian Branch of the Russian Academy of Sciences