КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА И ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА

УДК 001.103:002

АЛГОРИТМЫ ОПРЕДЕЛЕНИЯ СЕМАНТИЧЕСКОЙ БЛИЗОСТИ КЛЮЧЕВЫХ СЛОВ ПО ИХ ОКРУЖЕНИЮ В ТЕКСТЕ

И. Е. Воронина, А. А. Кретов, И. В. Попова

Воронежский государственный университет

Поступила в редакцию 01.03.2010 г.

Аннотация. Рассматриваются алгоритмы определения семантической близости ключевых слов: алгоритм Гинзбурга и его программная реализация и алгоритм с учетом частей речи и проблемы его реализации. Анализируются результаты вычислительного эксперимента.
Ключевые слова: компьютерная лингвистика семантическое поле слов, выделение ключевых слов, вычислительный эксперимент, алгоритм Гинзбурга, алгоритм с учетом частей речи.

Abstract. The article focuses upon two algorithms of semantic proximity assessment of the keywords: the Ginzburg's algorithm and its program implementation and the part of speech algorithm and the problems of its softwear implementation. The results of the computational experiment are discussed.
Key words: computational linguistics, semantic field of the word, .key words selection, computational experiment, Ginzburg's algorithm

ВВЕДЕНИЕ

Задача выделения ключевых слов в тексте на сегодняшний день является одной из самых актуальных и важных в таких областях как классификация текстов, информационный поиск и автоматическое реферирование. Разработано множество методов, различных по своим характеристикам и параметрам (например, [1]). Выделенные с их помощью ключевые слова могут служить основой для определения предметной области текста, его тематики и стилистической отнесенности.

Но при переходе к задаче описания содержательной структуры текста по ключевым словам необходима информация о том, как эти слова соотносятся друг с другом. Поэтому важно иметь инструмент для выявления отношений между ключевыми словами, для определения семантической близости этих слов. Подобный анализ может быть полезен как для проведения разнообразных лингвистических исследований, так и в задачах онтологического моделирования предметных областей, интеллектуальной обработки данных и др.

Рассмотрим алгоритмы, с помощью которых можно количественно оценить силу связи между словоформами в рамках исследуемого текста.

АЛГОРИТМ ГИНЗБУРГА

Алгоритм Гинзбурга [2] предназначен для поиска контекста данного слова в рамках рассматриваемого текста. Данный алгоритм состоит в следующем:

1. Находим в тексте T для каждой словоформы a ее относительную частоту – ОЧТ(a) (частное от деления наблюдаемой, абсолютной частоты на количество слов в тексте T).

2. Находим в T те части этого текста одного и того же размера, которые содержат заданное слово C – предложения с этим словом. Для совокупности всех этих предложений T* построим частотный словарь V(T*), содержащий абсолютную и относительную (частное от деления наблюдаемой, абсолютной частоты на количество слов в T*) частоты.. Относительную частоту словоформы a в V(T*) обозначим ОТЧ*(a).

3. Сравниваем полученные относительные частоты в T и T*: Вводится Индекс значимости словоформы a в контексте слова C (ИнЗ(a)), вычисляемый по формуле:

$$\text{ИнЗ}(a) = \frac{\text{ОЧТ}^{*}(a)}{\text{ОЧТ}(a)}. \quad (1)$$

4. Если полученный ИнЗ(а) больше единицы, то словоформу а относим к контексту слова С.

Опираясь на выявленные контекстные множества, можно получить информацию о взаимосвязях между словоформами внутри рассматриваемого текста. В частности, можно вычислить силу связи между двумя словоформами.

Для вычисления силы связи между словоформами a_1 и a_2 необходимо составить контекстные множества M_1 и M_2 по алгоритму Гинзбурга с положительным ИнЗ. Число словоформ в M_1 обозначим L_1, в M_2 – L_2. Множества могут иметь две части: общую для M_1 и M_2 и специфичную для M_1 или M_2. Алгоритм обработки множеств:

1) специфичные части суммируются по модулю (СуммаСпец);

2) для каждой словоформы в общей части вычисляется разность:

$$\text{ИнЗ}(M_1) - \text{ИнЗ}(M_2) \quad (2)$$

3) полученные разности суммируются по модулю (СуммаОбщ);

4) вычисляется общая сумма (СуммаТотал):

$$\text{СуммаТотал} = \text{СуммаСпец} + \text{СуммаОбщ} \quad (3)$$

5) вычисляется сумма всех ИнЗ всех словоформ, принадлежащих множеству M_1 (СумМ1);

6) вычисляется сумма всех ИнЗ всех словоформ, принадлежащих множеству M_2 (СумМ2);

7) вычисляется сила связи между словоформами a_1 и a_2:

$$\text{СилаСвязи} = 1 - \frac{\text{СуммаТотал}}{\text{СумМ1} + \text{СумМ2}}. \quad (4)$$

Полученный результат находится в диапазоне от 0 до 1. При этом 0 соответствует полному отсутствию общих слов в контексте, а 1 – полному совпадению множеств слов контекста и их частот.

Величину силы связи между словоформами также можно интерпретировать и как величину, обратную расстоянию между словоформами в семантическом пространстве исследуемого текста.

ПРОГРАММНАЯ РЕАЛИЗАЦИЯ

На основе изложенного подхода был разработан набор инструментальных и алгоритмических средств, позволяющих анализировать текст и вычислять силу связи между входящими в него словоформами. Он был добавлен к разработанной ранее программе выделения тематически маркированной лексики [3], реализующей подход А. А. Кретова [4] и предназначенной для выявления в тексте ключевых слов (рис. 1, 2).

Программа поддерживает следующие функции:

1) построение контекстного множества по алгоритму Гинзбурга для данной словоформы;

2) вычисление силы связи между двумя словоформами (рис. 3);

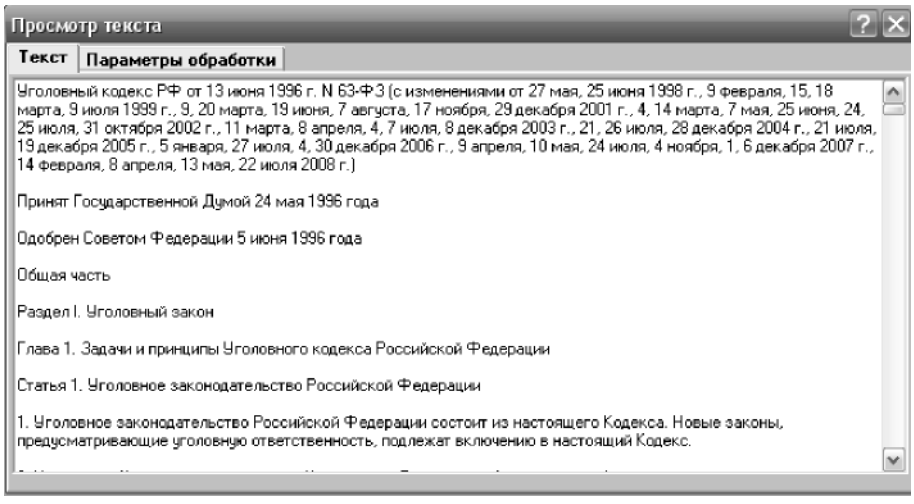

Рис. 1. Окно просмотра текста

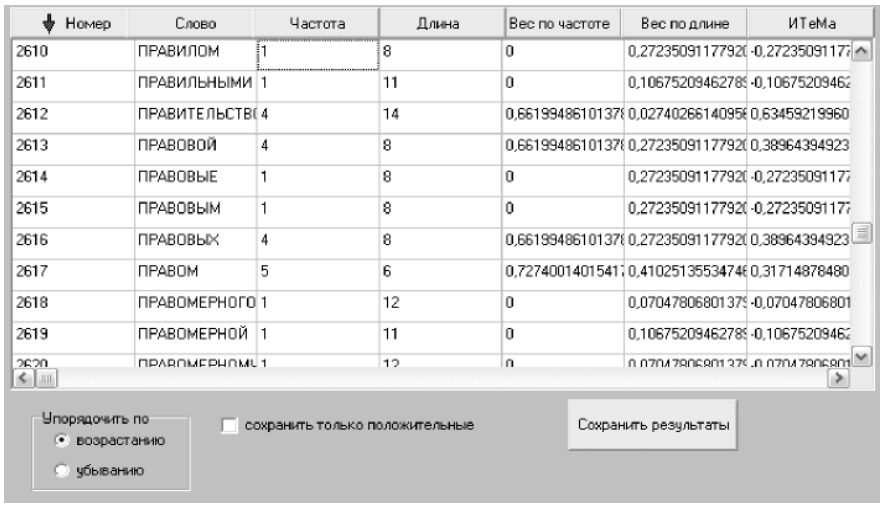

Рис. 2. Фрагмент окна просмотра результатов с таблицей Индекса тематической маркированности

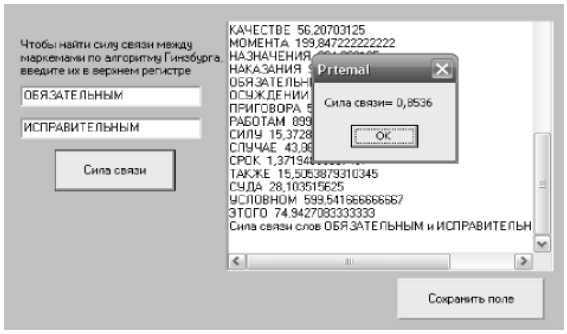

Рис. 3. Фрагмент окна просмотра результатов для вычисления силы связи

3) построение для заданного числа ключевых слов с максимальными индексами значимости (рис. 4) матрицы связей между словоформами (рис. 5).

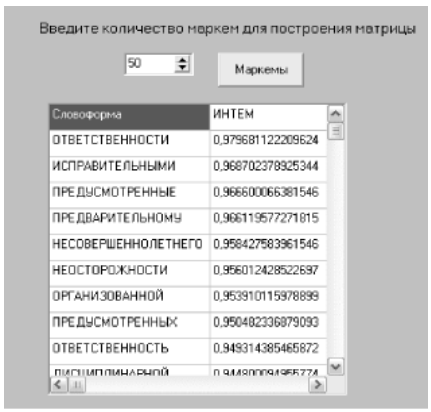

Рис. 4. Фрагмент окна просмотра для определения количества ключевых слов

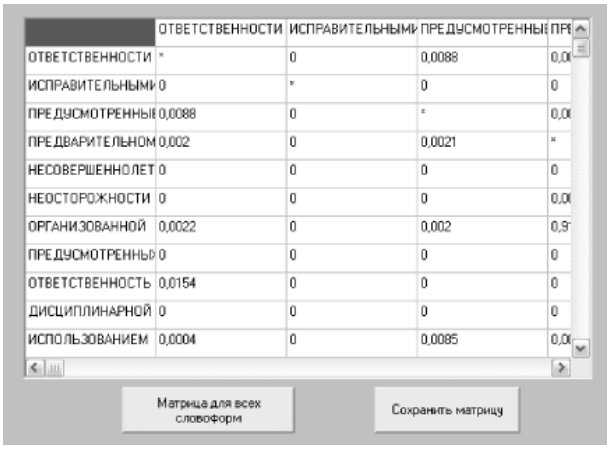

Рис. 5. Фрагмент матрицы для Общей части УК РФ

Пользователь, по своему выбору, может не включать в статистические данные слова, составляющие специфику естественного языка в целом, а также слова, не являющиеся значимыми в рамках конкретного исследования. Все получаемые результаты можно сохранить для последующего рассмотрения и обработки.

Матрица, построенная по заданному числу ключевых слов, содержит данные о силе связи между ними в рамках исследуемого текста. Она может служить основой для визуализации семантического пространства текста или его представления в виде графа.

В качестве вычислительного эксперимента с помощью программы был обработан текст Общей части Уголовного Кодекса РФ и построена матрица связей слов, рассчитанная для 200 ключевых слов с максимальным индексом тематической маркированности. В табл. 1 пред-

ставлены данные матрицы – ключевые слова, имеющие величину силы связи больше 0,1 (меньшие значения не рассматриваются как случайные).

Таблица 1.

Слова, связанные со словом ИСПРАВИТЕЛЬНЫЕ

Слово	Сила связи
дисциплинарной	0,7265
деятельностью	0,5079
определенной	0,5069
определенные	0,5148
государственных	0,9351
федерального	0,9579
заниматься	0,5131
ограничение	0,9160

Таблица 2.

Слова, связанные со словом НАРКОТИЧЕСКИХ

Слово	Сила связи
вымогательство	0,9533
психотропных	0,9954
уничтожение	0,9628
повреждение	0,9954
четырнадцатилетнего	0,9358
несовершеннолетний	0,9358
боеприпасов	0,9622
транспортных	0,9461

Можно увидеть, что полученные результаты (табл. 1, 2) действительно отражают особенности сочетаемости слов и взаимосвязь понятий Общей части Уголовного Кодекса РФ. Кроме того, выявляются сложные терминологические сочетания, состоящие из большого количества слов типа «ограничение заниматься определенной деятельностью».

АЛГОРИТМ С УЧЕТОМ ЧАСТЕЙ РЕЧИ

Существует другой способ вычисления силы семантической связи. Рассмотрим его общие принципы на примерах.

Возьмем строку:

Прозрачный лес один чернеет…

(А. С. Пушкин, «Зимнее утро»)

Прилагательное «*прозрачный*» стоит перед словом «*лес*» и связано только с ним одним. У прилагательных всего одна сочетаемость – с существительным. Следовательно, для учёта сочетаемости слова *«прозрачный»* достаточно взять все слова, которые встречаются справа от него.

Но нормальный порядок слов может меняться. Например:

И на весь тот лес обжитый,
И на весь передний край
У землянок домовитый
Раздавался песий лай…

(А. Т. Твардовский, «Василий Тёркин»)

А может быть и так:

На болотный лес корявый,
На кусты – снега легли.

(А. Т. Твардовский, «Василий Тёркин»)

Следовательно, необходимо учитывать не только правую, но и левую ближайшую сочетаемость прилагательного.

У существительных может быть больше связей – *Лес* и «*прозрачный*», и «*чернеет*». А справа от словоформы «*леса*» располагается не слово *«чернеет»*, а слово «*один*».

Ещё больше сочетаемостей может быть у глагола: *легли > снега*; *легли > на кусты; легли > на лес*. Или: *раздавался > лай; раздавался > у землянок; раздавался > на весь лес; раздавался > на весь передний край*.

В связи с этим предлагается для любого слова брать N слов слева и M слов справа. Исследователь должен иметь возможность изменять эти параметры. После этого все слова из всех позиций объединяются в одно множество и ранжируются в порядке убывания частоты.

При этом во множестве окажутся как слова, связанные с данным, несущие информацию, так и случайные соседи, создающие информационный шум, помехи, хаос. Для отделения информации от шума члены полученного множества фильтруются по частоте. Предполагается, что случайность имеет примерно одинаковую частоту, а частоты закономерных членов распределены неравномерно: от самых частотных – до самых редких. За фильтр частотности (за границу между сигналом и шумом) принимается средняя частота слов в данном множестве. Что выше / равно средней – сигнал, что ниже – шум. Кроме того, слова множества должны принадлежать иной части речи, нежели слово, для которого контекст строится.

Для вычисления силы связи между двумя словами, сравниваем «сигнальные» множества слов L_1 и L_2 (только те слова, частота которых выше/равна средней).

Итак, получаем следующий алгоритм:

1) для данного слова в тексте находятся все N соседей слева и M соседей справа;

2) извлечённые слова, имеющие другую часть речи, чем слово, для которого строится контекст, объединяются в одно множество;

3) создаётся частотный словарь множества;

4) члены множества ранжируются по убыванию частоты;

5) вычисляется средняя частота слов множества;

6) отсекается та часть слов множества, частота которых ниже средней;

7) полученное множество сравнивается с аналогичным множеством другого слова и вычисляется сила их связи.

Для вычисления силы связи между двумя словами сравниваются относительные частоты, с которыми они сочетаются. При этом берутся и суммируются все их частоты, сочетающиеся с первым словом и только те извлеченные слова со своими частотами второго слова, которые имеются (повторяются) в сочетаемости первого слова. Все частоты суммируются и делятся на 200%. Получаем процент перекрытия сочетаемости двух слов. Эту величину и принимаем за меру семантической близости второго и первого слова. Она, как и для алгоритма Гинзбурга, находится в диапазоне от 0 до 1, 0 означает полное отсутствие общих слов, 1 – полное совпадение множеств.

ЗАКЛЮЧЕНИЕ

Программная реализация алгоритма, учитывающего части речи, требует автоматизированного определения части речи словоформы. Использование для этой цели словарей приведет к значительному увеличению объема памяти. Поэтому нужен качественный морфологический анализатор, что само по себе является отдельной и весьма нетривиальной задачей. Поэтому применение второго алгоритма представляется целесообразным лишь в случае создания эффективного анализатора, чтобы определить часть речи словоформы. Однако, само по себе наличие такого алгоритма рано или поздно сыграет свою роль в проведении соответствующих лингвистических исследований.

Анализ данных, полученных в результате программной обработки Общей части Уголовного кодекса РФ, вычисление силы связи между словами по алгоритму Гинзбурга действительно отражает реальные взаимосвязи ключевых понятий данного текста. Таким образом, алгоритм может использоваться для выявления терминологических словосочетаний, которые принципиально не могут быть получены в рамках статистических подходов без учета контекста.

Дальнейшее развитие программы может идти в направлении расширения механизмов анализа получаемых данных для детализации взаимосвязей. Для получения полной картины при некоторых исследованиях необходимо предусмотреть возможность снятия ограничения по индексу тематической маркированности, поскольку из-за особенностей алгоритма может происходить отсеивание частотных слов небольшой длины, учет которых необходим, например, при выявлении терминологических сочетаний.

Также представляется необходимым разработать процедуры визуализации результатов в виде графа или иного изображения семантического пространства текста, что помогло бы не только наглядно продемонстрировать обнаруженные связи, но и дать пищу новым размышлениям.

СПИСОК ЛИТЕРАТУРЫ

1. *Крештель Е. В.* Алгоритмы выделения ключевых слов в текстах на естественных языках / Е. В. Крештель, А. А. Кретов, И. Е. Воронина // Межвузовский сборник научных трудов. Воронеж, 2001. Вып. 3. – С. 35.

2. *Гинзбург Е. Л.* Идиоглоссы: проблемы выявления и изучения контекста. / Е. Л. Гинзбург // Семантика языковых единиц: Доклады VI Международной конференции. Т. I., М., 1998. – С. 26–28.

3. *Воронина И. Е.* Функциональный подход к выделению ключевых слов: методика и реализация / И. Е. Воронина, А. А. Кретов, И. В. Попова, Л. В. Дудкина // Вестник Воронеж. гос. ун-та. Серия Системный анализ и информационные технологии. – 2009. – № 1. – С. 68–72.

4. *Кретов А. А.* Метод формального выделения тематически нейтральной лексики (на примере старославянских текстов) // Вестник Воронеж. гос. ун-та. Серия Системный анализ и информационные технологии. – 2007. – № 1. – С. 81–90.

Воронина Ирина Евгеньевна – к.т.н., доцент кафедры ПОиАИС, факультет прикладной математики информатики и механики, Воронежский государственный университет. Тел. 208-337.E-mail: voronina@amm.vsu.ru.

Кретов Алексей Александрович – д-р филол. наук, проф.,зав. каф. Теоретической и прикладной лингвистики, Воронежский государственный университет. Тел. (4732)204-149.E-mail: tipl@rgph.vsu.ru.

Попова Ирина Викторовна – студентка 5 курса факультета ПММ, кафедра программного обеспечения и администрирования информационных систем, специальность «Математическое обеспечение и администрирование информационных систем».

Voronina I. Ye. – Candidate of Technical Sciences, Associate Professor, the dept. of Software and Information System Administration, Voronezh State University. Tel. 208-337. E-mail: voronina@amm.vsu.ru.

Kretov A.A. – Doctor of Philology Sciences, Professor, the Head of the dept. of Theoretical and Applied Lingvistics, Voronezh State University.Tel. (4732) 204-149. E-mail: tipl@rgph.vsu.ru.

Irina V. Popova – 5th year student of Department of Applied Mathematics, Computer Science & Mechanics, Software & Information Systems Administering Program.