Программа выявления в тексте двучленных статистически значимых осмысленных коллокаций (на материале русского языка)

В.В. Залесская

Санкт-Петербургский государственный университет

vera-zalesskaya@yandex.ru

Аннотация

В статье рассмотрены меры ассоциации и проблемы выявления коллокаций. Представлена программа выявления коллокаций, которая обеспечивает выявление двучленных статистически значимых осмысленных коллокаций в текстах на русском языке и предлагает решение некоторых проблем выявления коллокаций.

Введение

С появлением компьютеров возникла возможность автоматической обработки лингвистических данных. Разнообразные программные средства помогают исследователям решать многие проблемы лингвистики, позволяют проводить недоступные ранее исследования. Появление корпусной лингвистики дало возможность изучать сочетаемость слов в больших массивах текстовых данных. Однако использование стандартных корпусных менеджеров не всегда дает удовлетворительные результаты.

В данной работе мы обратимся к проблеме повышения эффективности автоматического поиска коллокаций. Коллокацией называется «лексико-фразеологически обусловленная сочетаемость слов в речи как реализация их полисемии» [1, с. 193]. Степень обусловленности такой сочетаемости можно определить автоматически при помощи статистических методов — так называемых мер ассоциации. Мы постараемся увеличить их эффективность при выявлении коллокаций за счёт решения проблем, препятствующих корректной работе современных корпусных менеджеров, и создать собственное программное обеспечение, способное выявлять двусловные статистически значимые осмысленные коллокации в текстах на русском языке.

1. Меры ассоциации и проблемы выявления коллокаций

В качестве статистических методов выявления коллокаций мы рассмотрим в данной работе меры ассоциации — «математический аппарат для установления синтагматической связи между словами в тексте» [4, с. 11]. Меры ассоциации определяют силу ассоциации между коллокатами на основе частот их собственной и совместной встречаемости в корпусе.

При описании мер ассоциации в этой главе мы будем опираться на материалы Интернет-сайта [7].

1.1. Классификация мер ассоциации

На Интернет-сайте [7] в разделе, посвящённом мерам ассоциации, даются описание, характеристика и оценка эффективности различных типов этих мер. С.Эверт выделяет семь типов мер ассоциации:

- меры правдоподобия: binomial-likelihood, multinomial-likelihood, Poisson-Stirling, Poisson-likelihood, hypergeometric-likelihood;
- точные критерии для проверки гипотезы: binomial, Poisson, Fisher;
- асимптотические критерии для проверки гипотезы: z-score, t-score, chi-squared, log-likelihood;
- точечные оценки силы ассоциации: MI, odds-ratio, relative-risk, MS, Liddell, Dice, Jaccard;
- оценки силы ассоциации с запасом: $MI_{conf,\alpha}$;
- меры из теории информации: pointwise MI, average MI, local MI;
- эвристические меры: frequency, MI^2, MI^3, random.

Далее мы подробнее рассмотрим те меры ассоциации, которые будут использованы нами при создании программы выявления коллокаций.

1.2. Широко используемые меры ассоциации

1.2.1.Мера Log-likelihood

Как мы уже указали ранее, мера Log-likelihood (критерий отношения правдоподобия) относится к типу асимптотических критериев для проверки гипотезы. Формула для её вычисления существует в нескольких вариантах, однако мы воспользуемся модификацией оригинальной формулы Т.Даннинга [5], предложенной в [7] (см. рис.1).

$$\text{log-likelihood}_{\text{Dunning}} = -2\log\frac{L(O_{11},C_1,r)\cdot L(O_{12},C_2,r)}{L(O_{11},C_1,r_1)\cdot L(O_{12},C_2,r_2)}$$
$$L(k,n,r) = r^k(1-r)^{n-k}$$
$$r = \frac{R_1}{N},\ r_1 = \frac{O_{11}}{C_1},\ r_2 = \frac{O_{12}}{C_2}$$

Рис. 1. Модификация формулы для вычисления меры log-likelihood, используемая в нашей программе.

При создании программы мы будем пользоваться стандартным способом представления и зададим значения наблюдаемых частот следующим образом:

O_{11} = f(n,c);

O_{12} = f(n) – f(n,c);

O_{21} = f(c) – f(n,c);

O_{22} = N – f(n,c) – (f(n) – f(n,c)) – (f(c) – f(n,c)) = N + f(n,c) – f(n) – f(c),

где *n* — ключевое слово; *c* — коллокат; *f(n,c)* — частота встречаемости ключевого слова *n* в паре с коллокатом *c*; *f(n), f(c)* — абсолютные (независимые) частоты ключевого слова *n* и коллоката *c* в корпусе (тексте); *N* — общее число словоформ в корпусе (тексте).

1.2.2.Мера MI

Мера MI (коэффициент взаимной информации) относится к точечным оценкам силы ассоциации. В основе MI лежит понятие взаимной информации (mutual information), заимствованное из теории информации. Коэффициент взаимной информации сравнивает зависимые контекстно-связанные частоты с независимыми (при случайном появлении слов в контексте) [4, с.12]:

MI (n, c) = $\log_2\frac{f(n,c)\times N}{f(n)\times f(c)}$;

где *n* — ключевое слово; *c* — коллокат; *f(n,c)* — частота встречаемости ключевого слова *n* в паре с коллокатом *c*; *f(n), f(c)* — абсолютные (независимые) частоты ключевого слова *n* и коллоката *c* в корпусе (тексте); *N* — общее число словоформ в корпусе (тексте).

Мы видим, что наиболее высокие значения MI получают сочетания, для которых f(n,c) стремится к $f(n)\times f(c)$. Это характерно в частности для сочетаний низкочастотных элементов, которые могут оказаться случайными, а также для разного рода опечаток.

Мера MI существует в разных вариантах — обычном и нескольких нормализованных. Мы будем использовать при создании программы как обычный вариант, так и один из нормализованных – эвристическую меру MI^3. Её значение вычисляется по формуле [7]:

MI^3 = $\log_2\frac{f(n,c)^3\times N}{f(n)\times f(c)}$;

где *n* — ключевое слово; *c* — коллокат; *f(n,c)* — частота встречаемости ключевого слова *n* в паре с коллокатом *c*; *f(n), f(c)* — абсолютные (независимые) частоты ключевого слова *n* и коллоката *c* в корпусе (тексте); *N* — общее число словоформ в корпусе (тексте).

Как видно, эвристическая мера MI^3 увеличивает вес частоты совместной встречаемости в числителе, что не даёт MI завышать значения для низкочастотных сочетаний. Таким образом, MI^3 должна показать лучшие результаты при выявлении коллокаций на практике, чем обычная мера MI.

1.2.3.Мера T-score

T-score - мера ассоциации, которая, как и log-likelihood, относится к асимптотическим критериям для проверки гипотезы. Она вычисляется по формуле [8]:

t-score (n, c) = $\frac{f(n,c)-\left(\frac{f(n)\times f(c)}{N}\right)}{\sqrt{f(n,c)}}$;

где n — ключевое слово; c — коллокат; f(n,c) — частота встречаемости ключевого слова n в паре с коллокатом c; f(n), f(c) — абсолютные (независимые) частоты ключевого слова n и коллоката c в корпусе (тексте); N — общее число словоформ в корпусе (тексте).

Эта формула показывает, насколько распределения ключевого слова и коллоката в корпусе (тексте) зависят друг от друга. Однако возможна переоценка некоторых случайных результатов, в частности, сочетаний высокочастотного элемента с низкочастотным. По этой причине t-score обычно используется в комбинации с другими мерами, чаще всего с MI.

1.2.4.Мера Dice

Мера Dice относится к точечным оценкам силы ассоциации. Она вычисляется по формуле [8]:

Dice = $\frac{2f(n,c)}{f(n)+f(c)}$;

где n — ключевое слово; c — коллокат; f(n,c) — частота встречаемости ключевого слова n в паре с коллокатом c; f(n), f(c) — абсолютные (независимые) частоты ключевого слова n и коллоката c в корпусе (тексте).

В отличие от рассмотренных выше мер, Dice не учитывает размер корпуса (текста), будучи основанной только на частоте совместной встречаемости и независимых частотах. Опять же, как и в случае с MI, мы видим здесь возможность переоценки низкочастотных сочетаний.

Мера Dice имеет также нормализованную форму – logDice:

logDice = $14 + \log_2\frac{2f(n,c)}{f(n)+f(c)}$;

где n — ключевое слово; c — коллокат; f(n,c) — частота встречаемости ключевого слова n в паре с коллокатом c; f(n), f(c) — абсолютные (независимые) частоты ключевого слова n и коллоката c в корпусе (тексте).

1.2.5.Мера Minimum sensitivity

Мера minimum sensitivity – ещё одна точечная оценка силы ассоциации. Она рассчитывается по формуле[8]:

minimum sensitivity = $\min\left(\frac{f(n,c)}{f(n)}, \frac{f(n,c)}{f(n)}\right)$;

где *n* — ключевое слово; *c* — коллокат; *f(n,c)* — частота встречаемости ключевого слова n в паре с коллокатом c; *f(n), f(c)* — абсолютные (независимые) частоты ключевого слова n и коллоката c в корпусе (тексте).

Как и мера Dice, minimum sensitivity не учитывает размер корпуса (текста). Опять же, возникает возможность переоценки низкочастотных сочетаний и элементов.

1.2.6.Мера salience

Мера salience (MI.log-f[6]) не рассматривается в рамках классификации [7], однако мы считаем возможным рассматривать её как нормализованный вариант меры MI. Формула для её расчёта:

salience = $\mathrm{MI} \times \ln(f(n,c) + 1)$;

где n — ключевое слово; c — коллокат; f(n,c) — частота встречаемости ключевого слова n в паре с коллокатом c; f(n), f(c) — абсолютные (независимые) частоты ключевого слова n и коллоката c в корпусе (тексте).

Эта мера увеличивает вес частоты совместной встречаемости ключевого слова и коллоката по сравнению с MI. Таким образом, эффективность salience также должна быть выше, чем у MI.

2.Проблемы выявления коллокаций

Как мы указали ранее, существует множество разнообразных мер ассоциации. Не все из них используются при решении практических задач, но даже те, которые уже утвердились в традиционной практике, не лишены серьёзных недостатков. В нашем прошлом исследовании [3, с.26—28] мы выявили и наглядно проиллюстрировали некоторые проблемы, возникающие при решении практических задач с использованием распространённых мер ассоциации. Здесь мы упомянем главные из них, решить которые мы собираемся в нашей программе.

Во-первых, главной проблемой таких мер ассоциации, как log-likelihood и t-score, является выделение сочетаний слов со знаками препинания (или комбинациями знаков препинания) в качестве коллокаций. Такие сочетания в текстах встречаются очень часто и в результате имеют большие значения этих мер, чем осмысленные коллокации. Нужно также отметить, что мера MI имеет заниженные значения для таких сочетаний и таким образом не рассматривает их как полноценные коллокации.

Второй важной проблемой мер этого рода является выделение в качестве коллокаций сочетаний знаменательного слова со служебным. Такие сочетания, как правило, неосмысленны и не характеризуются зависимостью распределений коллокатов друг от друга. Мера MI и здесь отличается более правильным подходом к таким сочетаниям и занижением их значений.

Во-третьих, возникает проблема выделения в качестве коллокаций случайных неосмысленных сочетаний. К ним относятся опечатки разного рода и просто случайные употребления сочетаний слов в некоторых контекстах (например, в заголовках статей). Это становится наиболее очевидным при использовании меры MI – она завышает вес таких неосмысленных сочетаний. Однако такие меры, как log-likelihood и t-score, придают этим сочетаниям низкие значения, и они не попадают в список частотных коллокаций.

Таким образом, ни одна из самых распространённых мер ассоциаций не решает задачу выявления коллокаций безошибочно. Поэтому необходимо либо использовать эти меры в комбинации друг с другом, либо создать новый алгоритм выявления коллокаций, который занижал бы значения мер ассоциации для неосмысленных сочетаний, или вовсе исключал бы их на промежуточной стадии выделения.

3.Программа выявления коллокаций

3.1.Используемое программное обеспечение

Разработанная нами программа написана на языке программирования Python, что позволяет ей работать практически на всех платформах. Для корректной работы программы необходим Python версии 2.7 и установленная библиотека для морфологического анализа pymorphy со словарями для русского языка, а также установленная библиотека PyQt4 для поддержки программ с графическим интерфейсом.

3.2.Требования к входным данным

На вход программе подаётся текстовый файл на русском языке с расширением .txt, сохранённый в кодировке UTF-8. Для того, чтобы привести текст в необходимый формат, можно выполнить, например, такую последовательность действий:

- открыть текстовый файл в Microsoft Word;
- сохранить его как обычный текст;
- при сохранении указать кодировку Юникод (UTF-8).

3.3.Оценка эффективности программы выявления коллокаций

Для оценки эффективности программы выявления коллокаций мы использовали специальный текст — реферат по философии, так как для специальных текстов характерно наличие большого числа устойчивых и терминологических словосочетаний. Объём текста — 1098 слов (6785 знаков). Мы специально выбрали текст небольшого объёма, чтобы количество коллокаций было не очень большим (не более 50) и можно было подробно проанализировать результаты выполнения программы. Полный список коллокаций, найденных программой в специальном тексте, для двух типов поиска приведён в Приложении 2.

Как мы можем увидеть, программа находит в специальном тексте разные типы сочетаний:

- терминологические сочетания (*теория познания, простые идеи, внутренний опыт*);
- общеязыковые сочетания (*представлять собой, органы чувств, основные черты*);
- имена собственные (*Джон Локк*);
- вводные конструкции (*то есть*);
- сочетания, характеризующие тему текста (*черты эмпиризма, получение знаний, наш ум*);
- свободные сочетания (*ум получает, существует три*);
- сочетания с нераспознанными служебными словами (*при помощи*);
- сочетания с нераспознанными окказиональными элементами (*познание Дж*).

Наибольшее количество найденных в тексте сочетаний относится к первой группе — это различные термины философии и других наук. Также программа распознаёт много общеязыковых сочетаний, сочетаний, характеризующих тему текста, и свободных сочетаний.

Два последних типа сочетаний встречаются крайне редко. Проблемы с нераспознанными служебными словами и окказиональными элементами обусловлены не недоработками в алгоритме программы, а некорректностью работы библиотеки для морфологического анализа pymorphy в ряде случаев. Эти недостатки мы рассмотрим подробнее в следующем подразделе.

3.4.Случаи некорректной работы библиотеки pymorphy

Как можно увидеть из таблиц в Приложении 2, некорректная работа библиотеки pymorphy в ряде случаев приводит к следующим недостаткам:

- неправильное определение части речи слова может привести к включению сочетаний со служебными словами в выдачу (в сочетании «при помощи» часть речи предлога ‘при’ определена как имя существительное, поэтому это сочетание было включено программой в итоговую таблицу);
- неправильное определение леммы (при правильном определении части речи) может мешать правильному подсчёту частот встречаемости (в сочетании «получения знаний» лемма второго слова была определена как «ЗНАНЬЕ»);
- приписывание леммы и части речи окказиональным элементам (в сочетании «познание Дж» элементу «Дж» была приписана лемма «ДЖ» и часть речи — имя существительное);
- нераспознавание леммы (в сочетании «врождённых идей» лемма первого слова не была распознана, и оно было оставлено в исходном виде);

Так как случаи выдачи некорректных результатов крайне редки и обусловлены недостатками в работе библиотеки для морфологического анализа, а не алгоритма программы, то, в целом, можно считать, что поставленная нами задача решена и программа решает заявленные нами проблемы.

3.5.Решение проблем выявления коллокаций в разработанной программе

Разработанная нами программа решает несколько проблем выявления коллокаций, исследованных нами в прошлом году.

Во-первых, решена проблема поиска сочетаний слова и знака препинания в качестве коллокаций. Знаки препинания в нашей программе не рассматриваются как словоупотребления в тексте и исключаются из рассмотрения на одной из ранних стадий обработки текста.

Во-вторых, решена проблема поиска сочетаний знаменательного и служебного слова в качестве коллокаций. При помощи библиотеки pymorphy программа определяет часть речи для каждого словоупотребления в тексте и на стадии формирования списка потенциальных коллокаций исключает такие сочетания из рассмотрения.

В-третьих, решена проблема поиска низкочастотных сочетаний в качестве коллокаций. Наша программа исключает из рассмотрения все сочетания с частотой совместной встречаемости, меньшей или равной единице. Так как наша программа разработана для работы с небольшими по объёму текстами, этого вполне достаточно для достижения необходимых результатов.

В-четвёртых, словоупотребления, разделённые знаком препинания, в нашей программе не считаются коллокациями. Возможно, это некорректно в ряде случаев, но всё же, в основном, это очень эффективная мера отсеивания нерелевантных сочетаний.

Таким образом, наша программа решает некоторые важные проблемы выделения коллокаций, которые затрудняют корректную работу многих современных корпусных менеджеров.

3.6.Сравнение эффективности используемых в программе мер ассоциации

Чтобы оценить эффективность используемых мер, мы отобрали для каждой меры и типа поиска 10 выданных коллокатов с наибольшими значениями мер. Список коллокаций, использованный для оценки эффективности мер в нашей программе, мы поместили в Приложении 1.

Из таблиц Приложения 1 можно увидеть, что практически все меры ассоциации в нашей программе работают достаточно эффективно. Ни в одной из таблиц не оказалось сочетаний с окказиональными элементами. Сочетание «при помощи», выделенное многими мерами, оказалось в таблицах только из-за упомянутой нами выше

ошибки библиотеки pymorphy в приписывании частей речи.

Если считать выделенные мерами ассоциации свободные сочетания слов отрицательным показателем эффективности, то с этой точки зрения наименее эффективными оказались меры MI (по 2 свободных сочетания из 10 словоформ и лемм) и minimum sensitivity (2 свободных сочетания для словоформ и 1 для лемм). Чуть более эффективны меры Dice и Log-Dice (по одному свободному сочетанию для словоформ и лемм), причём нормализованная мера показала такую же эффективность, как и ненормализованная версия. Остальные меры не выделили ни одного свободного сочетания, все выявленные ими коллокации относятся к пяти первым категориям из пункта 2.5., то есть, они действительно выделяют словосочетания, воспринимаемые нами как несвободные или неслучайные. Следует отметить, что нормализованные версии меры MI — MI^3 и salience — оказались эффективнее, чем ненормализованная мера.

Ещё одно интересное наблюдение состоит в том, что набор терминологических сочетаний, выделенных различными мерами, несколько отличается, в то время как те немногие сочетания, которые были выделены всеми мерами без исключения, характерны для общей системы языка. Поэтому наше решение использовать в программе 9 мер ассоциации было полностью оправданным.

Таким образом, при правильной комбинации алгоритма выявления коллокаций и эффективных мер ассоциации можно достичь достаточно точных результатов при выявлении коллокаций.

Заключение и направления совершенствования программы

В данной работе мы привели подробное описание использующихся на практике мер ассоциации — статистических методов выявления сочетаемости слов в тексте, рассмотрели основные проблемы, возникающие при выявлении коллокаций, и попытались решить их при создании собственной программы поиска коллокаций в текстах на русском языке. Проблема выделения сочетаний со знаками препинания в нашей программе решена полностью, проблемы выделения сочетаний знаменательного слова со служебным и случайных сочетаний решены в той мере, в какой позволяет корректная работа библиотеки морфологического анализа. Разработанная нами программа выделяет двусловные коллокации нескольких типов: терминологические и общеязыковые сочетания, имена собственные, вводные конструкции, сочетания, характеризующие тему текста, а также некоторые свободные сочетания. Используемые в программе меры ассоциации в целом показали высокую эффективность при поиске коллокаций и не переоценивали свободные сочетания. Результаты разных мер совпадали при выделении общеупотребительных слов и расходились при выделении терминов. Это свидетельствует о том, что для эффективного выявления в тексте коллокаций разных типов необходимо использовать комбинацию нескольких мер. Однако главным является то, что решение проблем выявления коллокаций позволяет существенно повысить эффективность поиска осмысленных сочетаний и мер ассоциации.

Литература

[1] Ахманова О.С. Словарь лингвистических терминов. Москва: Советская энциклопедия, 1966.

[2] Залесская В.В. Внешний и внутренний опыт, идеи и качества в эмпиризме Дж. Локка: реферат. Санкт-Петербург, 2013.

[3] Залесская В.В. Сравнительный анализ статистических методов выявления словосочетаний на материале корпусов текстов различных функциональных стилей: курсовая работа. Санкт-Петербург, 2013.

[4] Хохлова М.В. Исследование лексико-семантической сочетаемости в русском языке с помощью статистических методов (на базе корпусов текстов). Санкт-Петербург, 2010.

[5] Dunning T. Accurate Methods for the Statistics of Surprise and Coincidence. URL: http://www.coli.uni-saarland.de/~schulte/Teaching/ESSLLI-06/Referenzen/Distributions/dunning-1993.pdf (дата обращения: 13.10.2014).

[6] Kilgarriff A., Tugwell D. Sketching words. URL: http://www.kilgarriff.co.uk/Publications/2002-KilgTugwell-AtkinsFest.pdf (дата обращения: 13.10.2014).

[7] Evert S. Association Measures//Computational Approaches to Collocations. — URL: http://collocations.de/AM/index.html (дата обращения: 13.10.2014).

[8] Statistics used in the Sketch Engine//SkE/Getting started — Sketch Engine. URL: https://www.sketchengine.co.uk/documentation/wiki/SkE/Help/PageSpecificHelp/collocconc (дата обращения: 13.10.2014).

Приложения

Приложение 1

Результаты выполнения программы выявления коллокаций на материале специального текста по философии

Таблица 1. 10 коллокаций с наибольшими значениями меры frequency

№	Поиск по словоформам	Поиск по леммам
1	ТЕОРИИ ПОЗНАНИЯ	ПРОСТОЙ ИДЕЯ
2	ДЖОН ЛОКК	ТЕОРИЯ ПОЗНАНИЕ
3	ПРОСТЫХ ИДЕЙ	НАШ УМ
4	ВНУТРЕННИЙ ОПЫТ	ДЖОН ЛОКК
5	ЭМПИРИЧЕСКОЙ ТЕОРИИ	СЛОЖНЫЙ ИДЕЯ
6	ЧЕРТЫ ЭМПИРИЗМА	ПЕРВИЧНЫЙ КАЧЕСТВО
7	ТО ЕСТЬ	ВНУТРЕННИЙ ОПЫТ
8	ПРОСТЫЕ ИДЕИ	ЭМПИРИЧЕСКИЙ ТЕОРИЯ
9	ПРИ ПОМОЩИ	ЧЕРТА ЭМПИРИЗМ
10	НАШЕГО УМА	ТОТ ЕСТЬ

Таблица 2. 10 коллокаций с наибольшими значениями меры log-likelihood

№	Поиск по словоформам	Поиск по леммам
1	ТЕОРИИ ПОЗНАНИЯ	ТЕОРИЯ ПОЗНАНИЕ
2	НАШЕГО УМА	ПРЕДСТАВЛЯТЬ СЕБЯ
3	ДЖОН ЛОКК	ЧЕРТА ЭМПИРИЗМ
4	ЧЕРТЫ ЭМПИРИЗМА	ПРЯ ПОМОЩЬ
5	ТО ЕСТЬ	ОРГАН ЧУВСТВО
6	ВНУТРЕННИЙ ОПЫТ	ПЕРВИЧНЫЙ КАЧЕСТВО
7	ПОЛИТИЧЕСКОГО УСТРОЙСТВА	ЭМПИРИЧЕСКИЙ ТЕОРИЯ
8	НАУЧНОЙ ДЕЯТЕЛЬНОСТИ	НАШ УМ
9	ПРИ ПОМОЩИ	УСТРОЙСТВО ГОСУДАРСТВО
10	ЭМПИРИЧЕСКОЙ ТЕОРИИ	ВТОРИЧНЫЙ КАЧЕСТВО

Таблица 3. 10 коллокаций с наибольшими значениями меры t-score

№	Поиск по словоформам	Поиск по леммам
1	ТЕОРИИ ПОЗНАНИЯ	ПРОСТОЙ ИДЕЯ
2	ДЖОН ЛОКК	ТЕОРИЯ ПОЗНАНИЕ
3	ВНУТРЕННИЙ ОПЫТ	НАШ УМ
4	ПРОСТЫХ ИДЕЙ	ДЖОН ЛОКК
5	НАШЕГО УМА	ПЕРВИЧНЫЙ КАЧЕСТВО
6	ЧЕРТЫ ЭМПИРИЗМА	СЛОЖНЫЙ ИДЕЯ
7	ТО ЕСТЬ	ВНУТРЕННИЙ ОПЫТ
8	ПРИ ПОМОЩИ	ПРЕДСТАВЛЯТЬ СЕБЯ
9	ЭМПИРИЧЕСКОЙ ТЕОРИИ	ЧЕРТА ЭМПИРИЗМ
10	ПРОСТЫЕ ИДЕИ	ОРГАН ЧУВСТВО

Таблица 4. 10 коллокаций с наибольшими значениями меры MI

№	Поиск по словоформам	Поиск по леммам
1	НАУЧНОЙ ДЕЯТЕЛЬНОСТИ	УСТРОЙСТВО ГОСУДАРСТВО
2	ПОЛИТИЧЕСКОГО УСТРОЙСТВА	ПРЕДСТАВЛЯТЬ СЕБЯ
3	ОСНОВНЫЕ ЧЕРТЫ	ОСНОВНЫЙ ЧЕРТА
4	УСТРОЙСТВА ГОСУДАРСТВА	ИМЕТЬ ОСНОВАНИЕ
5	НАШЕГО УМА	ПОЛИТИЧЕСКИЙ УСТРОЙСТВО
6	ЧЕРТЫ ЭМПИРИЗМА	ОРГАН ЧУВСТВО
7	ТО ЕСТЬ	ЧЕРТА ЭМПИРИЗМ
8	МОГУТ БЫТЬ	ПРЯ ПОМОЩЬ
9	СУЩЕСТВУЕТ ТРИ	СВОЙ ПРООБРАЗ
10	ПРЕДСТАВЛЯЮТ СОБОЙ	СУЩЕСТВОВАТЬ ТРИ

Таблица 5. 10 коллокаций с наибольшими значениями меры MI^3

№	Поиск по словоформам	Поиск по леммам
1	ТЕОРИИ ПОЗНАНИЯ	ТЕОРИЯ ПОЗНАНИЕ
2	НАШЕГО УМА	ПРЕДСТАВЛЯТЬ СЕБЯ
3	ЧЕРТЫ ЭМПИРИЗМА	ОРГАН ЧУВСТВО
4	ТО ЕСТЬ	УСТРОЙСТВО ГОСУДАРСТВО
5	НАУЧНОЙ ДЕЯТЕЛЬНОСТИ	ЧЕРТА ЭМПИРИЗМ
6	ПОЛИТИЧЕСКОГО УСТРОЙСТВА	ПРЯ ПОМОЩЬ
7	ДЖОН ЛОКК	НАШ УМ
8	ВНУТРЕННИЙ ОПЫТ	ПЕРВИЧНЫЙ КАЧЕСТВО
9	ОСНОВНЫЕ ЧЕРТЫ	ПРОСТОЙ ИДЕЯ
10	УСТРОЙСТВА ГОСУДАРСТВА	ОСНОВНЫЙ ЧЕРТА

Таблица 6. 10 коллокаций с наибольшими значениями меры Dice

№	Поиск по словоформам	Поиск по леммам
1	НАУЧНОЙ ДЕЯТЕЛЬНОСТИ	ПРЕДСТАВЛЯТЬ СЕБЯ
2	ПОЛИТИЧЕСКОГО УСТРОЙСТВА	УСТРОЙСТВО ГОСУДАРСТВО
3	НАШЕГО УМА	ТЕОРИЯ ПОЗНАНИЕ
4	ЧЕРТЫ ЭМПИРИЗМА	ОРГАН ЧУВСТВО
5	ТО ЕСТЬ	ОСНОВНЫЙ ЧЕРТА
6	ОСНОВНЫЕ ЧЕРТЫ	ИМЕТЬ ОСНОВАНИЕ
7	УСТРОЙСТВА ГОСУДАРСТВА	ЧЕРТА ЭМПИРИЗМ
8	ТЕОРИИ ПОЗНАНИЯ	ПОЛИТИЧЕСКИЙ УСТРОЙСТВО
9	ВНУТРЕННИЙ ОПЫТ	ПРЯ ПОМОЩЬ
10	МОГУТ БЫТЬ	СВОЙ ПРООБРАЗ

Таблица 7. 10 коллокаций с наибольшими значениями меры log-Dice

№	Поиск по словоформам	Поиск по леммам
1	НАШЕГО УМА	ПРЕДСТАВЛЯТЬ СЕБЯ
2	НАУЧНОЙ ДЕЯТЕЛЬНОСТИ	УСТРОЙСТВО ГОСУДАРСТВО
3	ПОЛИТИЧЕСКОГО УСТРОЙСТВА	ТЕОРИЯ ПОЗНАНИЕ
4	ЧЕРТЫ ЭМПИРИЗМА	ОРГАН ЧУВСТВО
5	ТО ЕСТЬ	ОСНОВНЫЙ ЧЕРТА
6	ОСНОВНЫЕ ЧЕРТЫ	ИМЕТЬ ОСНОВАНИЕ
7	УСТРОЙСТВА ГОСУДАРСТВА	ЧЕРТА ЭМПИРИЗМ
8	ТЕОРИИ ПОЗНАНИЯ	ПОЛИТИЧЕСКИЙ УСТРОЙСТВО

9	ВНУТРЕННИЙ ОПЫТ	ПРЯ ПОМОЩЬ
10	МОГУТ БЫТЬ	СВОЙ ПРООБРАЗ

Таблица 8. 10 коллокаций с наибольшими значениями меры minumum sensitivity

№	Поиск по словоформам	Поиск по леммам
1	НАУЧНОЙ ДЕЯТЕЛЬНОСТИ	ПРЕДСТАВЛЯТЬ СЕБЯ
2	ПОЛИТИЧЕСКОГО УСТРОЙСТВА	УСТРОЙСТВО ГОСУДАРСТВО
3	НАШЕГО УМА	ОСНОВНЫЙ ЧЕРТА
4	ЧЕРТЫ ЭМПИРИЗМА	ИМЕТЬ ОСНОВАНИЕ
5	ТО ЕСТЬ	ОРГАН ЧУВСТВО
6	ОСНОВНЫЕ ЧЕРТЫ	ТЕОРИЯ ПОЗНАНИЕ
7	ТЕОРИИ ПОЗНАНИЯ	СВОЙ ПРООБРАЗ
8	УСТРОЙСТВА ГОСУДАРСТВА	ЧЕРТА ЭМПИРИЗМ
9	МОГУТ БЫТЬ	ПОЛИТИЧЕСКИЙ УСТРОЙСТВО
10	СУЩЕСТВУЕТ ТРИ	НАШ УМ

Таблица 9. 10 коллокаций с наибольшими значениями меры salience

№	Поиск по словоформам	Поиск по леммам
1	ТЕОРИИ ПОЗНАНИЯ	ТЕОРИЯ ПОЗНАНИЕ
2	НАШЕГО УМА	ПРЕДСТАВЛЯТЬ СЕБЯ
3	ДЖОН ЛОКК	НАШ УМ
4	ЧЕРТЫ ЭМПИРИЗМА	ОРГАН ЧУВСТВО
5	ТО ЕСТЬ	ЧЕРТА ЭМПИРИЗМ
6	ВНУТРЕННИЙ ОПЫТ	ПЕРВИЧНЫЙ КАЧЕСТВО
7	ПОЛИТИЧЕСКОГО УСТРОЙСТВА	ПРЯ ПОМОЩЬ
8	НАУЧНОЙ ДЕЯТЕЛЬНОСТИ	ПРОСТОЙ ИДЕЯ
9	ПРОСТЫХ ИДЕЙ	ЭМПИРИЧЕСКИЙ ТЕОРИЯ
10	ПРИ ПОМОЩИ	УСТРОЙСТВО ГОСУДАРСТВО

Приложение 2

Полный список коллокаций, выделенный программой для выявления коллокаций

Таблица 10. Коллокации, выделенные при поиске по словоформам

№	Коллокат 1	Коллокат 2
1	ТЕОРИИ	ПОЗНАНИЯ
2	ДЖОН	ЛОКК
3	ПРОСТЫХ	ИДЕЙ
4	ВНУТРЕННИЙ	ОПЫТ
5	ЭМПИРИЧЕСКОЙ	ТЕОРИИ
6	ЧЕРТЫ	ЭМПИРИЗМА
7	ТО	ЕСТЬ
8	ПРОСТЫЕ	ИДЕИ
9	ПРИ	ПОМОЩИ
10	НАШЕГО	УМА
11	УСТРОЙСТВО	ГОСУДАРСТВА
12	УМ	ПОЛУЧАЕТ
13	СУЩЕСТВУЕТ	ТРИ
14	СЛОЖНЫХ	ИДЕЙ
15	СЛОЖНЫЕ	ИДЕИ
16	ПРЕДСТАВЛЯЮТ	СОБОЙ
17	ПОСРЕДСТВОМ	ОПЫТА
18	ПОЛУЧЕНИЯ	ЗНАНИЙ
19	ПОЛИТИЧЕСКОГО	УСТРОЙСТВА
20	ПОЗНАНИЯ	ДЖ
21	ОСНОВНЫЕ	ЧЕРТЫ
22	ОНИ	МОГУТ
23	НАШ	УМ
24	НАУЧНОЙ	ДЕЯТЕЛЬНОСТИ
25	МОГУТ	БЫТЬ
26	ЕГО	ВКЛАД
27	ВТОРИЧНЫЕ	КАЧЕСТВА
28	ВРОЖДЁННЫХ	ИДЕЙ
29	ВНЕШНЕГО	ОПЫТА

Таблица 11. Коллокации, выделенные при поиске по леммам

№	Коллокат 1	Коллокат 2
1	ПРОСТОЙ	ИДЕЯ
2	ДЖОН	ЛОКК
3	ТЕОРИЯ	ПОЗНАНИЕ
4	НАШ	УМ
5	ВНУТРЕННИЙ	ОПЫТ
6	СЛОЖНЫЙ	ИДЕЯ
7	ПЕРВИЧНЫЙ	КАЧЕСТВО
8	ЧЕРТА	ЭМПИРИЗМ
9	ЭМПИРИЧЕСКИЙ	ТЕОРИЯ
10	ОРГАН	ЧУВСТВО
11	ПРЕДСТАВЛЯТЬ	СЕБЯ
12	ТОТ	ЕСТЬ
13	ПРЯ	ПОМОЩЬ
14	ВТОРИЧНЫЙ	КАЧЕСТВО
15	ОНО	ВКЛАД
16	ОСНОВНЫЙ	ЧЕРТА
17	ПРОЦЕСС	ПОЛУЧЕНИЕ
18	ПОЛУЧЕНИЕ	ЗНАНЬЕ
19	ОНО	НАУЧНЫЙ
20	НАУЧНЫЙ	ДЕЯТЕЛЬНОСТЬ
21	ПОЛИТИЧЕСКИЙ	УСТРОЙСТВО
22	УСТРОЙСТВО	ГОСУДАРСТВО
23	ВРОЖДЁННЫХ	ИДЕЯ
24	ПОЗНАНИЕ	ДЖ
25	ЧУВСТВЕННЫЙ	ОПЫТ
26	ПОСРЕДСТВО	ОПЫТ
27	ВНЕШНИЙ	ОПЫТ
28	ВИД	ОПЫТ
29	КОТОРЫЙ	УМ
30	НАШ	ЗНАНИЕ
31	ОНИ	МОЧЬ
32	МОЧЬ	БЫТЬ
33	СУЩЕСТВОВАТЬ	ТРИ
34	УМ	ПОЛУЧАЯТЬ
35	ИМЕТЬ	ОСНОВАНИЕ
36	СВОЙ	ПРООБРАЗ
37	НАШ	ЧУВСТВО
38	ОБЪЕКТ	ОЩУЩЕНИЕ

A Program for Extracting Two-word Statistically Significant Meaningful Collocations from Russian Texts

V.V. Zalesskaya

The paper presents a discussion of association measures and collocation extraction problems. A program for collocation extraction suggested here provides for the extraction of two-word statistically significant meaningful collocations in Russian texts and offers a solution to some collocation extraction problems.