Труды Карельского научного центра РАН
№0. 2015. С. 1–28

*УДК 81.32*

# ОБЗОР МЕТОДОВ И АЛГОРИТМОВ РАЗРЕШЕНИЯ ЛЕКСИЧЕСКОЙ МНОГОЗНАЧНОСТИ: ВВЕДЕНИЕ

**Т. В. Каушинис[2], А. Н. Кириллов[1], Н. И. Коржицкий[2], А. А. Крижановский[1], А. В. Пилинович[2], И. А. Сихонина[2], А. М. Спиркова[2], В. Г. Старкова[1], Т. В. Степкина[2], С. С. Ткач[2], Ю. В. Чиркова[1], А. Л. Чухарев[1], Д. С. Шорец[2], Д. Ю. Янкевич[2], Е. А. Ярышкина[2]**

[1] *Институт прикладных математических исследований Карельского научного центра РАН*
[2] *Петрозаводский Государственный Университет*

Разрешение лексической многозначности — это задача выбора между разными значениями слов и словосочетаний в словаре в зависимости от контекста. В статье представлен краткий обзор методов и алгоритмов разрешения лексической многозначности. Эти методы используют различный математический и алгоритмический аппарат для решения WSD-задачи: нейронные сети, адаптивные алгоритмы улучшения точности обучения (AdaBoost), построение лексических цепочек, методы на основе применения теоремы Байеса и методы кластеризации контекстных векторов и семантически близких слов. Работу завершает исследование, в котором сравниваются время обучения, время работы и результаты работы разных алгоритмов решения WSD-задачи. Статья распространяется на правах свободной лицензии “CC Attribution”.

Ключевые слова: разрешение лексической многозначности, нейронная сеть, бустинг, лексическая цепочка, наивный байесовский классификатор, байесовская сеть, сочетаемостные ограничения, различение значений слов.

**T. V. Kaushinis[2], A. N. Kirillov[1], N. I. Korzhitsky[2], A. A. Krizhanovsky[1], A. V. Pilinovich[2], I. A. Sikhonina[2], A. M. Spirkova[2], V. G. Starkova[1], T. V. Stepkina[2], S. S. Tkach[2], J. V. Chirkova[1], A. L. Chuharev[1], D. S. Shorets[2], D. Y. Yankevich[2], E. A. Yaryshkina[2].** **A REVIEW OF WORD-SENSE DISAMBIGUATION METHODS AND ALGORITHMS: INTRODUCTION**

The word-sense disambiguation task is a classification task, where the goal is to predict the meaning of words and phrases with the help of surrounding text. The purpose of this short review is to acquaint the reader with the general directions of word sense disambiguation methods and algorithms. These approaches include the following groups of methods: neural network, machine learning meta-algorithms (AdaBoost), lexical chain computation, methods based on Bayes' theorem, context clustering and words clustering algorithms. The experimental comparison of different algorithms concludes this review. This paper is licensed under the CC Attribution license.

Key words: word-sense disambiguation, neural network, boosting, lexical chain, naive Bayes classifier, Bayesian network, selectional preferences, word sense discrimination.

## Введение

В статье представлен обзор методов и алгоритмов разрешения лексической многозначности (word-sense disambiguation или WSD). Верный выбор в словаре одного из значений многозначного слова или фразы в зависимости от контекста является успешным результатом решения WSD-задачи.

Приведем несколько примеров употребления слов «коса» и «косой», найденных с помощью Национального корпуса русского языка (http://ruscorpora.ru) по запросу «коса»:

1. Поп сам в первой *косе* идет, но прихожане не торопятся, смотрят на солнышко и часа через полтора уже намекают, что обедать пора. [*М. Е. Салтыков-Щедрин. Мелочи жизни (1886-1887)*]
2. Но работа даже и после этого идет все вялее и вялее; некоторые и *косы* побросали. [*М. Е. Салтыков-Щедрин. Мелочи жизни (1886-1887)*]
3. В особенности жестоко было крепостное право относительно дворовых людей: даже волосы крепостных девок эксплуатировали, продавая их *косы* парикмахерам. [*М. Е. Салтыков-Щедрин. Мелочи жизни (1886-1887)*]
4. Это одинокая скала, соединяющаяся с материком намывной *косой* из песка и гальки. [*В. К. Арсеньев, «По Уссурийскому краю», 1917 г.*]
5. Первая черепашка подскочила к гвардейцу и воткнула ему в спину сверкающий *косой* меч. [*Виктор Пелевин. S.N.U.F.F, 2011*]

Первые четыре примера дают три разных значения существительного «коса»: ряд косарей, сельскохозяйственное орудие, заплетенные волосы, протяженная речная отмель. Последний пример содержит прилагательное «косой», совпадающее с одной из форм существительного «коса». Все эти значения и часть речи читатель легко определяет по контексту.

Именно многозначность слов, их неоднозначность и зависимость значений слов от контекста являются причиной возникновения такой задачи и одновременно обуславливают сложность ее решения. Уверенное решение WSD-задачи необходимо во многих приложениях, связанных с автоматической обработкой текста (например, информационный поиск, машинный перевод) и, на наш взгляд, является предтечей искусственного интеллекта.

Среди основных методов разрешения лексической многозначности выделяют: методы, использующие внешние источники информации и методы, базирующиеся на машинном обучении, работающие на размеченных корпусах текстов. Также применяются комбинации этих методов [4] (с. 191–192).

По другой классификации методы разрешения лексической многозначности различают по типу используемых *внешних источников информации* [46] (с. 10:6–10:8):

- структурированные источники данных (машиночитаемые словари, тезаурусы, онтологии). Тезаурусы содержат информацию об отношениях между словами, такими как: синонимия, антонимия и другие. Классическим примером тезауруса и машиночитаемого словаря для английского языка является WordNet, в котором слова организованы в виде *синсетов* (от англ. ***synonym set***, группа синонимов), отношения указаны между синсетами.

- неструктурированные источники данных в виде корпусов текстов делятся на (а) неразмеченные корпуса (raw corpora) и (б) синтаксически и/или семантически размеченные корпуса.

На сегодняшний день на русском языке нет, по-видимому, достаточно объёмных и серьёзных обзоров по разрешению многозначности. Наиболее полное описание истории развития методов (20 страниц) есть в диссертации Д. Ю. Турдакова [7]. Такое положение дел послужило одной из причин написания этой статьи, которая будет заделом для полновесного обзора по данной теме.

Далее будут представлены примеры методов и алгоритмов разрешения лексической многозначности, разбитые на группы:

- нейронные сети — многообещающие методы с богатой историей;
- бустинг как метод улучшения точности алгоритма обучения;
- лексические цепочки — построение последовательности семантически связанных слов;
- метод ансамбля байесовских классификаторов и сочетаемостные ограничения на основе байесовских сетей;

- контекстная кластеризация — кластеризация контекстных векторов, где разные кластеры соответствуют разным значениям слова;
- кластеризация слов — это кластеризация семантически близких слов, при этом кластер соответствует некоторому значению.

Данная статья является «введением» в проблематику WSD, поскольку эта тема является чрезвычайно обширной и существуют сотни интересных работ по каждому из затронутых направлений.

## WSD на основе нейронных сетей, построенных по данным машиночитаемых словарей

*А. Н. Кириллов*

Использование нейронных сетей (NN) для WSD было предложено в 80-е гг. в работах [16, 62]. В типичной NN на вход подается слово, значение которого требуется установить, то есть целевое (*target*) слово, а также — контекст (фраза) его содержащий. Узлы выхода соответствуют различным значениям слова. В процессе обучения, когда значение тренировочного целевого слова известно, веса связующих узлы соединений (связей) настраиваются таким образом, чтобы по окончании обучения выходной узел, соответствующий истинному значению целевого слова, имел наибольшую активность. Веса соединений могут быть положительными или отрицательными, и настраиваются посредством рекуррентных алгоритмов (алгоритм обратного распространения ошибки, рекуррентный метод наименьших квадратов и так далее). Сеть может содержать скрытые (hidden) слои, состоящие из узлов, соединенных как прямыми, так и обратными связями. Для представления входной информации обычно используется одна из двух схем: распределенная (distributed) или локалистская (localist ) ([9], [17], [28] ).

В работе [61] описан метод автоматического построения ***очень больших нейронных сетей*** (VLNN) с помощью текстов, извлекаемых из машиночитаемых словарей (MRD), и рассмотрено использование этих сетей в задачах разрешения лексической неоднозначности. Поясним основную идею VLNN. Широко известен метод Леска [34] использования информации из MRD для задачи WSD. Суть этого метода состоит в вычислении так называемой *степени пересечения*, то есть количества общих слов в словарных определениях слов из контекста («окна») условного размера, содержащего целевое слово. Основной недостаток метода Леска — зависимость от словарной статьи, то есть от слов, входящих в нее. Стратегия преодоления этого недостатка — использование словарных статей, определяющих слова, входящие в другие словарные статьи, начиная со словарных статей, соответствующих словам из контекста. Таким образом, образуются достаточно длинные пути из слов, входящих в словарные статьи. Эта идея лежит в основе топологии VgN. В работе [61] для построения VLNN использован словарь Collins English Dictionary.

**Топология сети.** Целевое слово представлено узлом, соединенным активирующими связями со смысловыми узлами, представляющими все возможные значения слова, имеющиеся в словарных статьях. Каждый смысловой узел, в свою очередь, соединен активирующими связями с узлами, представляющими слова в словарной статье, соответствующей толкованию данного значения. Процесс соединения повторяется многократно, создавая сверхбольшую сеть взаимосвязанных узлов. В идеале сеть может содержать весь словарь. Авторы, по практическим соображениям, ограничиваются несколькими тысячами узлов и 10–20 тысячами соединений. Слова представлены своими леммами (каноническими формами). Узлы, представляющие различные значения данного слова, соединены запрещающими (inhibitory) связями.

**Алгоритм.** При запуске сети первыми активируются узлы входного слова (согласно принятой кодировке). Затем каждый входной узел посылает активирующий сигнал своим смысловым узлам, с которыми он соединен. В результате сигналы распространяются по всей сети в течение определенного числа циклов. В каждом цикле узлы слова и его значений получают обратные сигналы от узлов, соединенных с ними. Узлы конкурирующих значений посылают взаимно подавляющие сигналы. Взаимодействие сигналов обратной связи и подавления, в соответствии со стратегией «победитель получает все», позволяет увеличить активацию узлов-слов и соответствующих им правильных узлов-значений, одновременно уменьшая активацию узлов, соответствующих неправильным значениям. После нескольких десятков циклов сеть стабилизируется в состоянии, в котором активированы только узлы-значения с наиболее активированными связями с узлами-словами. При обучении сети используется метод обратного распространения (*back propagation*).

## Бустинг

*Т. В. Степкина, Ю. В. Чиркова*

Бустинг — это общий и доказуемо-эффективный метод получения очень точного правила предсказания путем комбинирования грубых и умеренно неточных эмпирических правил [22]. Метод бустинга разработан на основе модели обучения "PAC" (probably approximately correct learning).

Метод бустинга имеет множество реализаций. Работы, посвященные бустингу, обычно описывают какой-либо из его алгоритмов. Так, например, в работах [5, 12] рассматривается алгоритм arc-x4. В [6, 24] приводится алгоритм AdaBoost.M1. Мы рассмотрим бустинг на примере алгоритма AdaBoost, который является базовым для многих модификаций, а также имеет прочный теоретический фундамент и является результатом строгого вывода [5].

Алгоритм AdaBoost был предложен в 1995 году Фройндом и Шапиро [23]. В нём исправлены многие недостатки предыдущих алгоритмов бустинга.

AdaBoost является адаптивным алгоритмом [22], поскольку он может адаптироваться к уровням ошибок отдельных слабых гипотез. В названии "Ada" является сокращением от "adaptive" (адаптивный).

На вход алгоритма поступает обучающая выборка $(x_i; y_i); ..; (x_m; y_m)$, где каждый элемент $x_i$ принадлежит некоторому домену или признаковому пространству $X$, и каждая метка $y_i$ принадлежит некоторому набору меток $Y$. Для каждого обучающего примера $i$ вес распределения для целых $t$ обозначается $D_t$ $(i)$, где $t$ — это шаг алгоритма. За начальное распределение весов принимается $D_1(i) = 1/m$. Пусть метки принимают значения из множества $Y = \{-1, 1\}$.

Далее на каждом шаге $t$, где $t = 1 \dots T$, выполняется обучение с использованием текущего распределения $D_t$, после чего строится слабая гипотеза $h_t : X \to \{-1; 1\}$ с ошибкой первого рода $\varepsilon_t = \Sigma_{(i\,:\,h_t(x_i) \neq y_i)} D_t(i)$, по которой выбирается уровень значимости $\alpha_t = \frac{1}{2} ln(\frac{1-\varepsilon_t}{\varepsilon_t})$ и строится новое распределение для следующего шага

$$D_{t+1}(i) = \frac{D_{t^i}}{Z_t} \times \begin{cases} e^{-\alpha t}, \text{ если } h_t(x_i) = y_i \\ e^{\alpha t}, \text{ если } h_t(x_i) \neq y_i \end{cases} = \\ = \frac{D_t(i) exp(-\alpha_t y_i h_t(x_i))}{Z_t}$$

Конечная гипотеза $H(x)$ — это среднее из большинства решений $T$ слабых гипотез, где $\alpha_t$ — вес, присвоенный гипотезе $h_t$.

$$H(x) = sign(\sum_{t=1}^{T} \alpha_t h_t(x))$$

Идея алгоритма заключается в определении набора весов для обучающей выборки. Первоначально все веса примеров устанавливаются одинаково, но на каждом круге цикла веса неправильно классифицированных по гипотезе $h_t$ примеров увеличиваются. Таким образом получаются веса, которые относятся к сложным примерам.

Основное теоретическое свойство AdaBoost — это способность алгоритма уменьшать ошибку обучения [22]. Фройнд и Шапиро показали, что, так как каждая слабая гипотеза немного лучше случайного выбора, ошибка обучения уменьшается с экспоненциальной скоростью.

В статье [22] показано, как ограничена ошибка обобщения конечной гипотезы в терминах ошибки обучения, размера выборки m, VC размерности (размерности Вапника — Червоненкиса [57]) пространства слабых гипотез и количества циклов T. Также получена граница, не зависящая от T. Это показывает, что бустинг AdaBoost не подвержен эффекту переобучения.

Так как ошибка обучения и ошибка обобщения ограничены, как показано в статье [22], этот алгоритм действительно является бустинговым алгоритмом в том смысле, что он может эффективно преобразовать слабый алгоритм обучения в сильный, который может породить гипотезу со сколь угодно малой частотой ошибок, имея достаточное количество данных.

После того, как авторы рассмотрели бинарный случай, где целью является различие лишь между двумя возможными классами, они переходят к рассмотрению мультиклассного, более приближенного к реальности. Есть несколько способов приведения AdaBoost к мультиклассному случаю. Самое простое обобщение называется AdaBoost.M1 [24], которое является приемлемым, если слабый обучаемый может достичь достаточно высокой точности на распределениях, созданных AdaBoost. Тем не менее, этот метод завершается неудачно, если слабый ученик не может достичь хотя бы 50% точности при работе на этих распределениях. Для такого случая было разработано несколько методов:

1. Методы, которые работают за счет преобразования мультиклассной задачи в боль-

шую бинарную задачу или в набор бинарных задач. Эти методы требуют дополнительных усилий в разработке слабого алгоритма обучения.

2. Технология, которая включает в себя метод Диттерича и Бакири — метод выходных кодов, исправляющих ошибки [57].

AdaBoost обладает определенными преимуществами. Его быстро и просто запрограммировать. Он не имеет никаких параметров для настройки, за исключением количества циклов. Он не требует никаких предварительных знаний о слабом обучаемом и поэтому может быть скомбинирован с любым методом для нахождения слабых гипотез.

Недостатки метода заключаются в следующем. Фактическая производительность бустинга на конкретной задаче явно зависит от данных и слабого обучаемого. Теоретически, бустинг может выполниться плохо, если данных недостаточно, слабые гипотезы слишком сложные, или наоборот слишком слабые. Также бустинг особенно восприимчив к шуму.

AdaBoost был протестирован эмпирическим путем многими исследователями. Например, Фройнд и Шапиро проверили AdaBoost на множестве эталонных наборов данных UCI [41] с использованием C4.5 [52] как слабого алгоритма обучения, а также алгоритм, который находит самое лучшее дерево решений с одним тестом. После проведения эксперимента был сделан вывод, что бустинг даже слабых деревьев решений с одним тестом, как правило, дает хорошие результаты, в то время как бустинг C4.5, как правило, дает алгоритм дерева принятия решений значительно улучшенной производительности.

Почти во всех этих экспериментах и для всех показателей эффективности бустинг работает так же хорошо или значительно лучше, чем в других методах испытаний. Бустинг также применяется к фильтрации текстов, проблемам ранжирования и проблемам классификации, возникающих при обработке естественного языка. В работе [14] бустинг наряду с другими семью WSD-методами, используется для решения тестовой задачи с китайской лексикой. Проведенные эксперименты показали, что бустинг по точности уступает только методу максимальной энтропии и классификатору, комбинирующему бустинг, наивный байесовский классификатор, метод максимальной энтропии и PCA-модель. Для задачи с английской лексикой опыт применения бустинга описан в работах [20, 21]. В них авторы рассматривают алгоритм LazyBoosting – модификация AdaBoost.MH [55]. По результатам сравнительных экспериментов бустинг оказывается по точности лучше таких методов, как наивный байесовский классификатор, метод основанный на примерах (Exemplar Based) и MFS (naive Most-Frequent-Sense classifier).

## Использование лексических цепочек для реферирования текстов

*А. В. Пилинович*

В статье [10] с целью реферирования текста строится модель в виде лексических цепочек. Реферирование включает четыре этапа: оригинальный текст делится на блоки (сегменты), строятся лексические цепочки, определяются сильные цепочки, извлекаются важные предложения.

Реферирование — это процесс сжатия исходного текста в более компактный при сохранении информативности текста. Реферирование выполняется для решения разных задач — от обзорного анализа текстов какой-либо научной области до быстрого выделения главных тем текста. Создание качественной информативной аннотации произвольного текста является сложной задачей, требующей полного понимания текста. Легче создавать приблизительные, указательные аннотации (indicative summaries), позволяющие принять решение — стоит ли читать текст. В работе [10] описан метод создания указательных аннотаций по произвольным текстам.

Интуитивное понятие *cohesion* (связность, склеивание, слияние), введенное в [25], указывает на объединение разных частей (фрагментов) текста в одно целое, в то, что имеет значение, смысл. Одним из видов связности является *лексическая связность* (lexical cohesion) [30]. Лексическая связность формируется с помощью семантически связанных слов. Халлидей и Хасан [25] выделили два способа формирования лексической связности: (1) с помощью категории повторений и (2) категории словосочетаний.

1. *Лексическая связность повторений* (reiteration category) достигается повтором слов, использованием синонимов и гипонимов.

2. *Лексическая связность словосочетаний* (collocation category) определена для слов, которые часто употребляются вместе, то есть встречаются в одних и тех же контекстах.

Слова и фразы, между которыми существует лексическая связность, формируют

лексическую цепочку (lexical chains) [30]. Метод лексических цепочек, предложенный Барзилей и Эльхадад [10], основан на анализе совместной встречаемости слов и лексических связей между словами.

**Алгоритм построения цепочек.** Достоинство лексических цепочек в том, что их легко распознать и построить. Первая вычислительная модель для лексических цепочек была представлена в работе Морриса и Хирста [45]. Цепочки создавались путем взятия нового слова из текста и поиска родственной (связанной) цепочки для слова в соответствии с критериями родства. Недостатком подхода в [45] было то, что в одну цепочку могло входить слово с разными значениями (для многозначных слов). Таким образом, выбор подходящей цепочки для слова эквивалентен решению WSD-задачи.

Метод построения лексических цепочек включает шаги:

1. Выбирается набор слов-кандидатов (существительные и составные существительные). Это кандидаты на включение в цепочки.

2. Строится список всех значений для каждого слова-кандидата (по данным словаря).

3. Для каждого значения каждого слова-кандидата находится (вычисляется) отношение (расстояние) до каждого слова во всех уже построенных цепочках (слово в цепочке имеет строго определённое значение, задаваемые другими словами в той же цепочке). Между двумя словами есть отношение (будет указана связь в цепочке), если мало расстояние между этими словами в тексте (text distance) или между значениями этих слов существует путь в тезаурусе WordNet. Выделяют три вида отношений ([18], стр. 36):

   (а) *Extra-strong* отношение существует для слов, повторяющихся в тексте. Повтор может быть на любом расстоянии от первого употребления слова.

   (б) *Strong* отношение определено между словами, связанными отношением в WordNet. Два таких слова должны находиться в окне не более семи предложений.

   (в) *Medium-strong* отношение указывается для слов, синсеты которых находятся на расстоянии больше одного в WordNet (но есть ещё и дополнительные ограничения на путь между синсетами). Слова в тексте должны находиться в пределах трех предложений.

4. Слово-кандидат добавляется в цепочки, со словами которых найдена связь. Смысловая неоднозначность устраняется, то есть в цепочку добавляется не просто слово, а его конкретное значение (благодаря выбору значения в словаре на шаге 2).

Для выбора приоритетной цепочки (для вставки слова-кандидата) отношения упорядочены так: extra-strong, strong, medium-strong. В работе Хирста и Ст-Онж [29] предложен жадный алгоритм выбора цепочек. При этом слово-кандидат попадает ровно в одну цепочку и после этого выбор уже не может быть изменён, даже если последующий текст покажет ошибочность первоначального решения. В работе Барзилей и Эльхадад [10] предложена более сложная схема выбора «подходящего значения», требующая рассмотрения всех возможных цепочек. Таким образом, будут сформированы цепочки с учетом всех возможных значений слов с последующим выбором наилучшей цепочки. Эта более сложная схема и рассматривается далее.

Для иллюстрации метода приводится пример на отрывке текста, представленном ниже, и посмотрим, какое значение будет выбрано для слова *machine*. Во-первых, для слова *Mr.* создается узел [лексема “Mr.”, значение {mister, Mr.}]. Следующим по тексту существительным, представленным в тезаурусе WordNet, будет слово *person*, у него есть два значения: *“human being”* ($person_1$) и *“grammatical category of pronouns and verb forms”* ($person_2$). Наличие двух значений у слова *person* разбивает пространство цепочек на два множества интерпретаций: в первой интерпретации используется значение $person_1$, во второй — $person_2$, (Рис. 1).

***Mr.*** *Kenny is the* ***person*** *that invented an anesthetic* ***machine*** *which uses* ***micro-computers*** *to control the rate at which an anesthetic is pumped into the blood. Such* ***machines*** *are nothing new. But his* ***device*** *uses two* ***micro-computers*** *to achieve much closer monitoring of the* ***pump*** *feeding the anesthetic into the patient.*

*Компонентой* в [10] называют список взаимоисключающих интерпретаций. Именно посредством компонент выбор одного из значений слов ведёт к выбору соответствующей ин-

терпретации, а, следовательно, к невозможности других интерпретаций из этой компоненты. Интерпретации 1 и 2 на Рис. 1 являются компонентой.

Следующее слово *anesthetic* не связано со словами из первой компоненты, поэтому для него создается компонента с одним значений (то есть новая компонента содержит ровно одну интерпретацию).

Следующее слово *machine* имеет 5 значений: от $machine_1$ до $machine_5$. В первом значении $machine_1$ [лексема "machine", значение {an efficient person}] слово связано со значениями слов *person* и *Mr.*, поэтому слово *machine* вставляется в первую компоненту. После этой вставки изображение первой компоненты становится таким, как показано на Рис. 2. Если продолжить этот процесс и вставить слова *micro-computer, device* и *pump,* то количество альтернативных вариантов значительно увеличивается. Самые сильные интерпретации представлены на Рис. 3. При условии, что текст связный, лучшей интерпретацией считается та, которая имеет больше всего связей. В данном случае в конце шага 3 выбрана другая интерпретация $machine_4$ [лексема "machine", значение {*any mechanical or electrical device that performs or assists in the performance*}], что верно отражает значение слова *machine* в этом контексте.

Оценка интерпретации определяется как сумма оценок ее цепочек. Оценка цепочки определяется количеством и весом отношений между участниками цепочки. В эксперименте авторы зафиксировали следующий вес: повторения и синонимы — 10, антонимы — 7, гиперонимы и гипонимы — 4. Описанный алгоритм вычисляет все возможные интерпретации, не допуская противоречий между ними. Когда число возможных интерпретаций превышает определенный порог, слабые интерпретации удаляются, это необходимо для предотвращения экспоненциального роста использования памяти.

**Объединение цепочек из разных сегментов.** Текст предварительно разбивается на сегменты (несколько предложений или абзац). Пример выше (*Mr. Kenny...*) соответствует одному сегменту. Цепочки строятся для каждого сегмента на основе найденных отношений между словами (extra-strong, strong, medium-strong). На следующем этапе объединяются цепочки из разных сегментов, но для объединения нужно, чтобы выполнялось еще более жесткое условие: две цепочки объединяются, если они содержат одно и то же слово в одном и том же значении. Поскольку есть прямая связь между цепочками и смысловыми блоками текста, постольку с помощью лексических цепочек можно решать и обратную задачу — разбиение текста на сегменты [10].

**Вычисление оценок цепочек.** Для того чтобы использовать лексические цепочки для построения аннотации, в первую очередь следует выявить сильнейшие цепочки среди всех тех, которые создаются описанным выше алгоритмом. Барзилей и Эльхадад в [10] предложили эмпирическую методику для оценки силы цепочки. Они разработали среду, чтобы вычислить и графически визуализировать лексические цепочки, чтобы оценить экспериментально, насколько хорошо идентифицируются (определяются) основные темы текстов. Авторы собрали данные из 30 текстов, выбранных случайным образом из популярных журналов (например, "The Economist", "Scientific American"). Для каждого текста вручную выполнили ранжирование цепочек по степени соответствия основным темам текста.

Из множества параметров, которые можно измерить (длина цепочки; объем текста, покрываемого цепочкой; плотность; диаметр слов цепочки в графе тезауруса; число повторений), Барзилей и Эльхадад [10] опытным путем нашли следующие показатели значимости цепочек для построение реферата:

- *Длина (Length):* число употреблений в тексте элементов цепочки.
- *Индекс однородности (HomogeneityIndex):* 1 — количество различных употреблений в тексте элементов цепочки, деленное на длину (*Length*).

Таким образом, значимость цепочек оценивается так:

$$\text{Score(Chain)} = \text{Length} \times \text{HomogeneityIndex}$$

При ранжировании цепочек в соответствии с этой оценкой было найдено, что для построения реферата нужны цепочки, удовлетворяющие «критерию прочности (силы)»:

$$\text{Score(Chain)} > \text{Average(Scores)} + 2 \times \text{StandardDeviation(Scores)}$$

где *Average* — это средняя оценка по всем цепочкам, *StandardDeviation* — среднеквадратическое отклонение.

**Извлечение важных предложений.** После того как сильные цепочки отобраны, выполняется поиск соответствующих им предложений и извлечение этих предложений целиком из исходного текста.

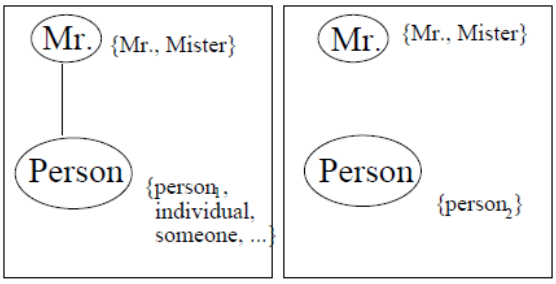
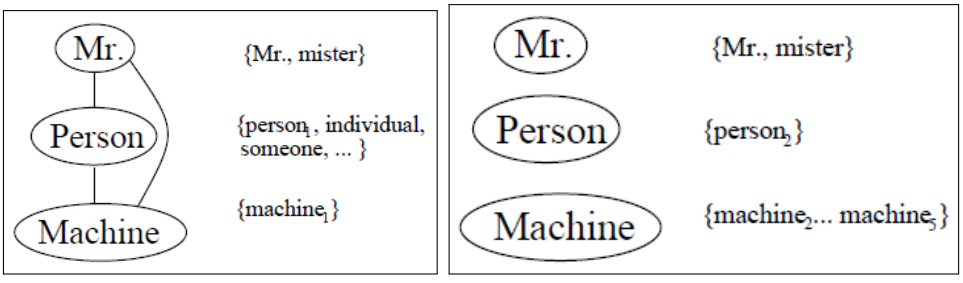
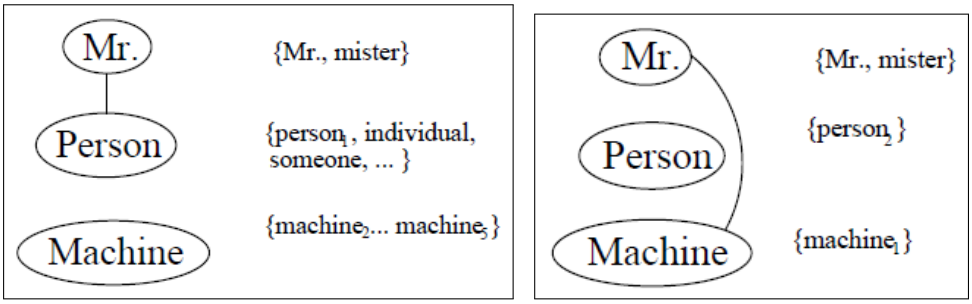



*Рис. 1.* Шаг 1, интерпретация 1 (слева) и 2 (справа)

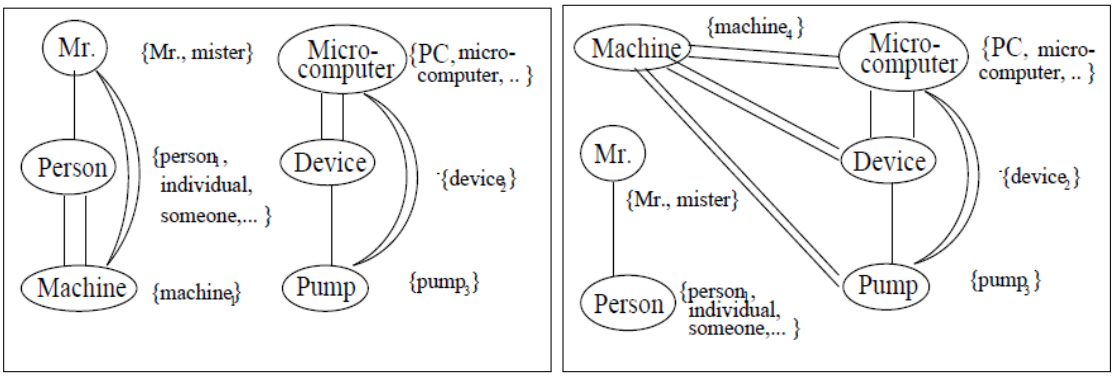



(a) Интерпретация 1 (b) Интерпретация 2




(c) Интерпретация 3 (d) Интерпретация 4

*Рис. 2.* Четыре интерпретации на втором шаге




(a) Интерпретация 1 (b) Интерпретация 2

*Рис. 3.* Две самые сильные интерпретации, полученные на третьем шаге

Для каждой сильной цепочки на основе разработанных эвристик выбирается ровно одно предложение для включения в текст реферата:

Эвристика 1: Для каждой цепочки для включения в реферат выбрать то предложение, которое содержит первое появление члена цепочки в тексте.

Эвристика 2: Для каждой цепочки для включения в реферат выбрать предложение, которое содержит первое появление *показательного* элемента цепочки в тексте. *Показательные* слова (representative words), служащие представителями цепочки, — это такие слова цепочки, которые встречаются в цепочке не реже, чем в среднем по всем словам цепочки.

Эвристика 3: Для каждой цепи найти блок текста, где есть высокая концентрация цепочки (то есть много употреблений элементов из этой цепочки). Извлечь предложение с первого появления цепочки в этом блоке. Концентрация вычисляется как число появлений членов цепи в сегменте, разделенное на количество существительных в сегменте. Цепочка имеет высокую концентрацию, если ее концентрация является максимальной из всех цепочек. Кластер представляет собой группу последовательных сегментов, таких, что каждый сегмент содержит какие-либо элементы цепочки.

Эксперименты в [10] показали значительное преимущество алгоритма на основе лексических цепочек (точность 47-61% и полнота 64-67%) по сравнению с программой Microsoft Summarizer, доступной в Word'97 (точность 32-33% и полнота 37-39%). Эти результаты указывают на большой потенциал лексических цепочек в задаче реферирования.

## Байесовский классификатор и сочетаемостные ограничения

В первой части главы строится ансамбль наивных байесовских классификаторов. Наивный байесовский классификатор — это простой вероятностный классификатор на основе применения теоремы Байеса. Для различения значений учитывается совместная встречаемость слов в окне заданного размера в текстах корпуса.

Во второй части главы для каждого глагола строится байесовская сеть. Байесовская модель обучается сочетаемостным ограничениям глаголов, то есть обучается тому — с какими существительными глаголы могут употребляться. Сочетаемостные ограничения позволяют ограничить число значений целевого слова по данным контекста [46] (с. 10:32). Связи глагол-существительное извлекаются из корпуса текстов, а классы существительных задаются тезаурусом WordNet. Маловероятные значения слов при построении сочетаемостных ограничений отбрасываются (стратегия "explaining away").

### Разрешение лексической многозначности методом ансамбля байесовских классификаторов

*А. Л. Чухарев, Т. В. Каушинис*

В работе Педерсена [49] рассматривается подход к разрешению лексической многозначности слов (WSD), подразумевающий создание ансамбля наивных байесовских классификаторов, каждый из которых основан на оценке вероятности вхождения определенных слов в контекст целевого слова, значение которого определяется.

При разрешении лексической многозначности, представленном в виде задачи обучения с учителем, применяют статистические методы и методы машинного обучения к размеченному корпусу. В таких методах словам корпуса, для которых указано значение, соответствует набор языковых свойств. Педерсен [49] относит к языковым свойствам два вида особенностей: так называемые простые лексические признаки (shallow lexical features) и более сложные лингвистически обусловленные признаки (lingvistically motivated features). К первым относятся совместная встречаемость слов (co-occurence) и словосочетания (collocations), в то время как вторые включают в себя такие свойства, как часть речи и отношение действие-объект. Обычно алгоритмы обучения строят модели классификаторов значений по этим языковым свойствам.

Автор статьи [49] предлагает подход, основанный на объединении ряда простых классификаторов в ансамбль, который разрешает многозначность с помощью голосования простым большинством голосов. Педерсен утверждает [49], что, во-первых, более сложные алгоритмы обычно не улучшают точность разрешения. Во-вторых, совместная встречаемость слов и словосочетаний имеют большее влияние

на точность разрешения, чем оперирование более сложной лингвистической информацией.

В рассматриваемой статье [49] в ансамбль объединяются наивные байесовские классификаторы. При таком подходе предполагается, что все переменные, участвующие в представлении проблемы, — условно независимы при фиксированном значении переменной классификации. В проблеме разрешения лексической многозначности существует понятие контекста, в котором встречается многозначное слово. Этот контекст представляется в виде функции переменных $(F_1, F_2, \ldots, F_n)$, а значение многозначного слова представлено в виде классификационной переменной (S). Все переменные бинарные. Переменная, соответствующая слову из контекста, принимает значение ИСТИНА, если это слово находится на расстоянии определенного количества слов слева или справа от целевого слова. Совместная вероятность наблюдения определенной комбинации переменных контекста с конкретным значением слова выражается следующим образом:

$$p(F_1, F_2, \ldots, F_n S) = p(S)\Pi_{i=1}^{n} p(F_i \vee S)$$

где $p(S)$ и $p(F_i|S)$ — параметры данной модели. Для оценки параметров достаточно знать частоты событий, описываемых взаимозависимыми переменными $(F_i, S)$. Эти значения соответствуют числу предложений, где слово, представляемое $F_i$, встречается в некотором контексте многозначного слова, упомянутого в значении $S$. Если возникают нулевые значения параметров, то они сглаживаются путем присвоения им по умолчанию очень маленького значения. После оценки всех параметров модель считается обученной и может быть использована в качестве классификатора.

Контекст в [49] представлен в виде bag-of-words (модель «мешка слов»). В этой модели выполняется следующая предобработка текста: удаляются знаки препинания, все слова переводятся в нижний регистр, все слова приводятся к их начальной форме (лемматизация). В [49] контексты делятся на два окна: левое и правое. В первое попадают слова, встречающиеся слева от неоднозначного слова, и, соответственно, во второе — встречающиеся справа.

Окна контекстов могут принимать 9 различных размеров: 0, 1, 2, 3, 4, 5, 10, 25 и 50 слов. Первым шагом в ансамблевом подходе является обучение отдельных наивных байесовских классификаторов для каждого из 81 возможных сочетаний левого и правого размеров окон. В статье [49] наивный байесовский классификатор $(l, r)$ включает в себя $l$ слов слева от неоднозначного слова и $r$ слов справа. Исключением является классификатор (0,0), который не включает в себя слов ни слева, ни справа. В случае нулевого контекста классификатору присваивается **априорная вероятность** многозначного слова (равная вероятности встретить наиболее употребимое значение).

Следующий шаг в [49] при построении ансамбля — это выбор классификаторов, которые станут членами ансамбля. 81 классификатор группируется в три общие категории, по размеру окна контекста. Используются три таких диапазона: узкий (окна шириной в 0, 1 и 2 слова), средний (3, 4, 5 слов), широкий (10, 25, 50 слов). Всего есть 9 возможных комбинаций, поскольку левое и правое окна отделены друг от друга. Например, наивный байесовский классификатор (1,3) относится к диапазону категории (узкий, средний), поскольку он основан на окне из одного слова слева и окне из трех слов справа. Наиболее точный классификатор в каждой из 9 категорий диапазонов выбирается для включения в ансамбль. Затем каждый из 9 членов классификаторов голосует за наиболее вероятное значение слова с учетом контекста. После этого ансамбль разрешает многозначность путем присвоения целевому слову значения, получившего наибольшее число голосов.

**Экспериментальные данные.** Для экспериментов были выбраны английские слова *line* и *interest*. Источником статистических данных по этим словам послужили работы [33], [13]. В статье приводится информация о частоте использования шести значений для каждого из этих слов (Табл. 1, Табл. 2).

*Таблица 1.* Число употреблений слова *line* для шести наиболее часто встречаемых значений (из тезауруса WordNet) по данным корпусов *ACL/DCI Wall Street Journal и American Printing House for the Blind*

| **Значение** | **Частота** |
|---|---|
| product | 2218 |
| written or spoken text | 405 |
| telephone connection | 429 |
| formation of people or things; queue | 349 |
| an artificial division; boundary | 376 |
| a thin, flexible object; cord | 371 |
| **Всего** | 4148 |

*Таблица 2.* Число употреблений слова *interest* для шести наиболее часто встречаемых значений (из словаря Longman Dictionary of Contemporary English). Этот набор данных был получен в 1994 году Брюсом и Виебе [13] путем указания значений для всех вхождений слова *interest* в корпус ACL/DCI Wall Street Journal

| **Значение** | **Частота** |
|---|---|
| money paid for the use of money | 1252 |
| a share in a company or business | 500 |
| readiness to give attention | 361 |
| advantage, advancement or favor | 178 |
| activity that one gives attention to | 66 |
| causing attention to be given to | 11 |
| **Всего** | 2368 |

**Результаты экспериментов.** Итогом проделанной работы стали обучение и проверка 81 наивного байесовского классификатора на многозначных словах *line* и *interest*. Точность разрешения лексической многозначности составила 89% для слова *interest* и 88% для слова *line*. В [49] было получено, что ансамбль классификаторов с голосованием простым большинством дает более высокую точность, чем взвешенное голосование. Например, для слова *interest* при голосовании простым большинством точность составила 89%, а взвешенное голосование дало только 83%.

## Построение сочетаемостных ограничений на основе байесовских сетей для разрешения многозначности

*И. А. Сихонина*

В статье [15] представлена байесовская модель, применяемая для разрешения лексической многозначности глаголов. Авторы рассматривают такое понятие, как сочетаемостные ограничения (selectional preferences). *Сочетаемостные ограничения* (далее SP) — это закономерности использования глагола относительно семантического класса его параметров (субъект, объект (прямое дополнение) и косвенное дополнение).

Модели автоматического построения SP важны сами по себе и имеют приложения в обработке естественного языка. Сочетаемостные ограничения глагола могут применяться для получения возможных значений неизвестного параметра при известных глаголах; например, из предложения *«Осенние хххх жужжали и бились на стекле»* легко определить, что "хххх" — мухи. При построении предложения SP позволяют отранжировать варианты и выбрать лучший среди них. Исследование SP могло бы помочь в понимании структуры ментального лексикона.

Системы обучения SP без учителя обычно комбинируют статистические подходы и подходы, основанные на знаниях. Компонент базы знаний (здесь WordNet [42]) — это обычно база данных, в которой слова сгруппированы в классы.

Статистический компонент состоит из пар предикат-аргумент, извлеченных из неразмеченного корпуса. В тривиальном алгоритме можно было бы получить список слов (прямых дополнений глагола), и для тех слов, которые есть в WordNet, вывести их семантические классы. В работе [15] семантическим классом называется *синсет* тезауруса WordNet, то есть класс соответствует одному из значений слова. Таким образом, в тривиальном алгоритме на основе данных WordNet можно выбрать классы (значения слов), с которыми употребляются (встречаются в корпусе) глаголы.

Например, если в исходном корпусе текстов глагол *ползать* употребляется со словом *ящерица*, принадлежащим классу РЕПТИЛИИ, то в модели построения SP будет записано, что «глагол *ползать* употребляется со словами из класса РЕПТИЛИИ». Если слово *крокодил*, во-первых, также встречается в тексте с глаголом *ползать*, во-вторых, слово *крокодил* принадлежит сразу двум классам: РЕПТИЛИЯ и ВЕРТОЛЁТ, то из этого следует, что модель SP будет расширена информацией о том, что «глагол *ползать* употребляется со словами из классов и РЕПТИЛИЯ, и ВЕРТОЛЁТ».

В ранее разработанных моделях (Резник (1997) [53], Абни и Лайт (1999) [8]) было обнаружено, что главная трудность в таком тривиальном алгоритме — это наличие неоднозначных слов в обучающих данных. В тех же работах ([53], [8]) были предложены более сложные модели, в которых предполагается, что все значения многозначных слов появляются с одинаковой частотой.

**Байесовские сети** или байесовские сети доверия (БСД) состоят из множества переменных (вершин) и множества ориентированных ребер, соединяющих эти переменные. Такой сети соответствует ориентированный ациклический граф. Каждая переменная может принимать одно из конечного числа взаимоисключающих состояний. Пусть все переменные будут бинарного типа, то есть принимают одно из двух значений: истина или ложь. Лю-

бой переменной $A$ с родителями $B_1, ..., B_n$ соответствует таблица условных вероятностей (conditional probability table, далее CPT).

Например, построим SP для глагола *ползать* и сеть на Рис. 4 будет базой знаний.

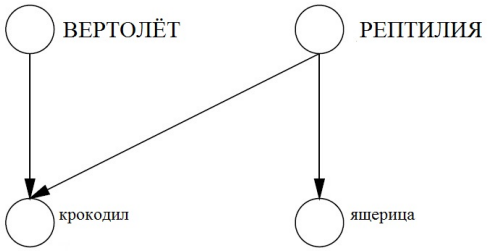



*Рис. 4.* Байесовская сеть для многозначного существительного *крокодил*

Глагол *ползать* употребляется со словами *крокодил* и *ящерица*. Переменные ВЕРТОЛЁТ и РЕПТИЛИЯ соответствуют более общим абстрактным значениям, переменные *крокодил* и *ящерица* являются более узкими, конкретными значениями. Переменная РЕПТИЛИЯ может принимать одно из двух значений, соответствующих словам *крокодил* и *ящерица*, именно эту задачу определения значения и нужно решить.

*Таблица 3.* Условные вероятности переменных *крокодил* и *ящерица* в зависимости от значений переменных ВЕРТОЛЁТ и РЕПТИЛИЯ, где (В, Р, к, я — это аббревиатуры слов ВЕРТОЛЁТ, РЕПТИЛИЯ, *крокодил* и *ящерица*)

| | $P(X = x \mid Y_1 = y_1, Y_2 = y_2)$ | | | |
|---|---|---|---|---|
| | В,Р | В,¬Р | ¬В,Р | ¬В, ¬Р |
| к = *true* | 0,99 | 0,99 | 0,99 | 0,01 |
| к = *false* | 0,01 | 0,01 | 0,01 | 0,99 |
| я = *true* | 0,99 | 0,99 | 0,01 | 0,01 |
| я = *false* | 0,01 | 0,01 | 0,99 | 0,99 |

При построении Табл. 3 условных вероятностей (CPT), учтем следующие предположения:

- вероятность, что выбираем какой-либо из концептов (ВЕРТОЛЁТ и РЕПТИЛИЯ) очень мала, то есть $P(B{=}true) = P(P{=}true) = 0{,}01$, следовательно, велика вероятность, что концепты не выбраны: $P(B{=}false) = P(P{=}false) = 0{,}99$;
- если какой-либо из концептов истинен (В, Р), то «выпадает» слово *крокодил*;
- если концепт РЕПТИЛИЯ истинен, то растут шансы встретить слово *ящерица*;

Из Табл. 3 вероятности появления слов следует вывод, что использование разу двух значений слова *крокодил (рептилия и вертолёт МИ-24)* маловероятно. Вероятность использования значения РЕПТИЛИЯ намного больше чем значения ВЕРТОЛЁТ. Таким образом, гипотеза «вертолёт» «отброшена» ("explaining away").

**Байесовские сети для построения SP.** Иерархия существительных в WordNet представлена в виде ориентированного ациклического графа. Синсет узла принимает значение «истина», если глагол «выбирает» существительное из набора синонимов. Априорные вероятности задаются на основе двух предположений: во-первых, маловероятно, что глагол будет употребляться только со словами какого-то конкретного синсета, и во-вторых, если глагол действительно употребляется только со словами из данного синсета (например, синсет ЕДА), тогда должно быть правомерным употребление этого глагола с гипонимами этого синсета (например, ФРУКТ).

Те же предположения (что для синсетов) верны и для употреблений слов с глаголами:

1. слово, вероятно, является аргументом глагола в том случае, если глагол употребляется с каким-либо из значений этого слова;
2. отсутствие связки глагол-синсет говорит о малой вероятности того, что слова этого синсета употребляются с глаголом.

Словам «вероятно» и «маловероятно» должны быть приписаны такие числа, сумма которых равна единице.

Находкой работы [15] является разъяснение стратегии "explaining away", то есть отбрасывание маловероятных значений слов при построении сочетаемостных ограничений. Такая стратегия является неотъемлемым свойством байесовских сетей и байесовского вывода, полезным свойством при разрешении лексической многозначности.

Каждому вхождению анализируемого слова в корпус соответствует контекстный вектор. Выполняется кластеризация векторов, где разные кластеры соответствуют разным значениям слова [46] (с. 10:26–10:28). Алгоритмы кластеризации полагаются на дистрибутивную гипотезу (Distributional Hypothesis) [26], в соответствии с которой слова, употребляемые в схожих контекстах, считаются близкими по смыслу.

В первой части выполняется разрешение лексической многозначности и поиск новых значений на основе контекстных векторов, построенных по биомедицинским текстам.

Во второй части представлена задача *различения значений слов*. Эта задача отличается от задачи разрешения лексической многозначности тем, что при различении значений слов нет никаких предопределенных значений слова, присоединенных к кластерам; здесь слова, употребляющиеся в схожих контекстах, группируются в кластеры (значения).

## Кластеризация фрагментов биомедицинских текстов

*Е. А. Ярышкина*

В статье [54] изучаются уже существующие методы кластеризации без учителя и их эффективность для решения лексической многозначности при обработке текстов по биомедицине. Решение проблем лексической многозначности в данной области включает в себя не только традиционные задачи присвоения ранее определенных смысловых значений для терминов, но также и обнаружения новых значений для них, еще не включенных в данную онтологию.

Для разрешения лексической многозначности Савова и др. [54] предложили разделять контексты (фрагменты текста), содержащие определенное целевое слово, на кластеры, где разные кластеры будут соответствовать различным значения целевого слова. Каждый кластер состоит из близких по значению контекстов. Предполагается, что используемое целевое слово в аналогичном контексте будет иметь одно и то же или очень близкое значение (дистрибутивная гипотеза).

Процесс кластеризации продолжается до тех пор, пока не будет найдено предварительно заданное число кластеров. Выбор шести кластеров в работе [54] основан на том факте, что это больше, чем максимальное число возможных значений любого английского слова, наблюдаемое среди данных (большинство слов имеют два-три значения). Нормализация текста не выполняется.

Данные в этом исследовании состоят из ряда контекстов, которые включают данное целевое слово, где у каждого целевого слова вручную отмечено — какое значение из словаря было использовано в этом контексте. Контекст — это единственный источник информации о целевом слове. Цель исследования — преобразовать контекст в контекстные вектора первого и второго порядка [2]. Контекстные вектора содержат следующие «лексические свойства»: биграммы, совместную встречаемость и совместную встречаемость целевого слова. Биграммами являются как двухсловные словосочетания, так и любые два слова, расположенные рядом в некотором тексте. Для лингвистических исследований могут быть полезны только упорядоченные наборы биграмм [1].

Экспериментальные данные — это набор NLM WSD [60] (NLM — национальная библиотека медицины США), в котором значения слов взяты из UMLS (единая система медицинской терминологии). UMLS имеет три базы знаний:

- *Метатезаурус* включает все термины из контролируемых словарей (SNOMED-CT, ICD и другие) и понятия, которые представляют собой кластера из терминов, описывающих один и тот же смысл.
- *Семантическая сеть* распределяет понятия на 134 категории и показывает отношения между ними. SPECIALIST-лексикон содержит семантическую информацию для терминов Метатезауруса.
- *Medline* — главная библиографическая база данных NLM, которая включает приблизительно 13 миллионов ссылок на журнальные статьи в области науки о жизни с уклоном в биомедицинскую область.

Авторы успешно проверили по три конфигурации существующих методов (PB — Pedersen and Bruce [50], SC — Schütze [58]) и оценили эффективность использования SVD (сингулярное разложение матриц). Методы PB основаны на контекстных векторах первого порядка — признаки одновременного присутствия целевого слова или биграммы. Рассчитывается среднее расстояние между кла-

стерами или применяется метод бисекций. PB методы подходят для работы с довольно большими наборами данных. Методы SC основаны на представлениях второго порядка — матрицы признаков одновременного присутствия или биграммы, где каждая строка и столбец — вектор признаков первого порядка данного слова. Так же рассчитывается среднее расстояние между кластерами или применяется метод бисекций. SC методы подходят для обработки небольших наборов данных.

Метод SC2 (признаки одновременного присутствия второго порядка, среднее расстояние между элементами кластера в пространстве подобия) с применением и без SVD показал лучшие результаты: всего 56 сравниваемых экземпляров, в 47 случаях метод SC2 показал наилучшие результаты, в 7 случаях результаты незначительно отличаются от других проверяемых методов.

Все эксперименты, указанные в исследовании, выполнялись с помощью пакета SenseClusters [59]. В ходе исследования было проведено два эксперимента для разных наборов данных. Маленький тренировочный набор — это набор NLM WSD, который включает 5000 экземпляров для 50 часто встречаемых неоднозначных терминов из Метатезауруса UMLS. Каждый неоднозначный термин имеет по 100 экземпляров с указанным вручную значением. У 21 термина максимальное число экземпляров находится в пределах от 45 до 79 экземпляров. У 29 терминов число экземпляров от 80 до 100 для конкретного значения. Стоит отметить, что каждый термин имеет категорию «ни одно из вышеупомянутых», которая охватывает все оставшиеся значения, не соответствующие доступным в UMLS. Большой тренировочный набор является реконструкцией «1999 Medline», который был разработан Weeber [63]. Были определены все формы из набора NLM WSD и сопоставлены с тезисами «1999 Medline». Для создания тренировочного набора экземпляров использовались только те тезисы из «1999 Medline», которым было найдено соответствие в наборе NLM WSD.

Использование целиком текста аннотации статьи в качестве контекста приводит к лучшим результатам, чем использование отдельных предложений. С одной стороны, большой объем контекста, представленный аннотацией, дает богатую коллекцию признаков, с другой стороны, в коллекции WSD представлено небольшое число контекстов.

## Различение значений слов на основе векторов свойств, расширенных словарными толкованиями

*А. М. Спиркова*

Амрута Пурандаре и Тед Педерсен в 2004 году разработали «Алгоритм различения значений на основе контекстных векторов» (*Context vector sense discrimination*) [51]. В этом алгоритме (1) берется набор примеров употреблений исследуемого слова, (2) выполняется кластеризация этих примеров так, чтобы близкие по значению или связанные каким-либо образом слова объединились в одну группу [51].

*Word sense discrimination* — это задача группировки нескольких употреблений данного слова в кластеры, где каждому кластеру соответствует определенное значение целевого слова. Подходы к решению этой проблемы основываются на дистрибутивной гипотезе. Следует различать понятия *различение значений слов* и *разрешение лексической многозначности*. При *различении значений слов* нет никаких предопределенных значений слова, присоединенных к кластерам; здесь, скорее, слова, употребляющиеся в схожих контекстах, группируются в кластеры (значения).

При решении задачи *различения значений* используются контекстные вектора: если целевое слово встречается в тестовых данных, то контекст этого слова представляется в виде вектора контекста. *Вектор контекста* — это средний вектор по векторам свойств каждого из слов контекста. *Вектор свойств* содержит информацию о совместной встречаемости данного слова с другими словами, этот вектор строится по данным корпуса текстов на этапе обучения.

Метод различения значений Пурандаре и Педерсена [51] предназначен для работы при недостаточном объеме текстовых данных, при этом вектор свойств расширяется данными, извлеченными из толкований словарей. Этот метод группирует в кластеры близкие по значению употребления целевого слова.

**Построение матрицы встречаемости слов.** Первоначально строится матрица совместной встречаемости слов по данным обучающего корпуса (были использованы тексты Wall Street Journal и Британского национального корпуса).

Вектор свойств (строка матрицы) содержит информацию о совместной встречаемости данного слова с другими. Было решено в [51], что слова «встречаются», если они находятся

в тексте на расстоянии не более пяти словопозиций (то есть между ними находится не более трех слов).

**Обработка матрицы.** После создания матрицы выполняется разделение тестовых данных, то есть группировка примеров употреблений (фраз) с целевым словом. Каждому слову в примере употребления в тестовых данных соответствует вектор свойств из матрицы встречаемости. Средний вектор свойств по всем словам соответствует вектору контекста. Таким образом, набор тестовых данных, включающих употребление исследуемого слова, преобразуется в набор контекстных векторов, каждый из которых соответствует одному из употреблений целевого слова.

Различение значений происходит путем кластеризации контекстных векторов с помощью разделяющего (partitional) или иерархического «сверху вниз» (agglomerative) алгоритма кластеризации [31], [32], [64]. Получающиеся кластеры составлены из употреблений близких по значению фраз, и каждый кластер соответствует отдельному значению целевого слова.

**Векторы свойств, расширенные текстами толкований из словаря.** Векторы свойств, полученные по небольшому корпусу текстов, имеют очень малую размерность (несколько сотен), что не позволяет полностью описать закономерности совместной встречаемости слов. Для решения этой проблемы векторы свойств слов расширяются содержательными словами (content words), извлеченными из словарных толкований разных значений данного слова. В Табл. 4 представлены примеры толкований и содержательные слова для восьми значений слова «история» из Русского Викисловаря.

*Таблица 4.* Словарные толкования (и содержательные слова) по данным статьи «история» из Русского Викисловаря. Серым цветом и курсивом выделены те слова, которые уже были в векторе слов, черным – новые слова из толкований, которыми будет расширен вектор свойств

| № | Текст значения | Содержательные слова |
|---|---|---|
| 1 | закономерное, последовательное развитие, изменение действительности | *развитие,* изменение |
| 2 | наука, изучающая факты, тенденции и закономерности развития человеческого общества | *наука,* факт, тенденция, закономерность |
| 3 | наука, изучающая ход развития, последовательные изменения какой-либо области природы или культуры | *наука, развитие,* изменение |
| 4 | последовательный ход развития, изменения чего-либо, совокупность фактов о развитии какого-либо явления | *развитие,* изменение, факт |
| 5 | отдаленное время с его событиями, происшествиями; прошлое | время, событие, происшествие |
| 6 | эпическое повествование, рассказ | повествование, *рассказ* |
| 7 | смешная или неожиданная ситуация, происшествие, случай | ситуация, случай, происшествие |
| 8 | скандал, неприятность | скандал, неприятность |

Предположим, например, что вектор свойств (столбец в матрице встречаемости) для слова *история* имеет непустые значения в строках, соответствующих словам: *книга, мир, наука, образование, развитие, рассказ.*

В Русском Викисловаре различные значения слова *история* (Табл. 4) включают содержательные слова: *время, закономерность, изменение, наука, неприятность, повествование, происшествие,* развитие, рассказ, *ситуация, скандал, случай, событие, тенденция, факт.* Таким образом, вектор свойств, соответствующий слову «история», будет расширен новыми (отсутствующими ранее) словами из словаря: *время, закономерность, изменение, неприятность, повествование, происшествие, ситуация, скандал, случай, событие, тенденция, факт.*

В итоге, вектор свойств будет включать слова: *время, закономерность, изменение,* книга, мир, наука, *неприятность,* образование, *повествование, происшествие,* развитие, рассказ, *ситуация, скандал, случай, тенденция, факт.*

Для оценки результатов тестовым примерам употребления присваивали вручную теги значений. Кластеру присваивалось то значение, примеров употребления которого в нем было больше всего.

Авторами было проведено 75 экспериментов с использованием 72 слов из корпуса SENSEVAL-2 и со словами *line, hard* и *serve.*

В тестовых данных SENSEVAL-2 примеры употреблений включали 2-3 предложения. Для каждого слова было дано от 50 до 200 примеров употреблений в тестовых и тренировочных данных. Для этих слов известно много (порядка 8-12) значений. Малое число примеров при большем числе значений привело к тому, что для некоторых значений оказалось мало примеров употреблений. 43 из 72 слов SENSEVAL-2 показали улучшение F-меры и полноты (recall) при расширении вектора свойств текстами толкований словаря. Однако для 29 слов F-мера стала хуже, что, возможно, говорит о неправильном применении метода, в том числе о нерепрезентативности выборки. Для окончательной оценки необходима большая экспериментальная база: не десятки слов, а десятки и сотни тысяч.

Данный метод может быть полезен при различении значений слов без учителя при небольшом количестве обучающих данных.

## Кластеризация слов

Кластеризация слов — это кластеризация семантически близких слов, при этом кластер соответствует одному из значений исследуемого слова [46] (с. 10:28–10:29).

В первой части описан метод построения пары взаимных ближайших соседей и автоматического создания тезауруса. Для этого из текста извлекаются тройки зависимостей (слово 1, слово 2, отношение), затем эти тройки используются для вычисления близости значений слов.

Алгоритм кластеризации посредством комитетов, представленный во второй части раздела, также можно отнести к задаче *различения значения слов.* В алгоритме последовательно вычисляется сходство между словами, строится набор компактных кластеров (комитетов), все слова распределяются по этим кластерам

### Автоматический поиск и кластеризация похожих слов

*Д. С. Шорец*

В работе [35] представлен метод автоматического создания тезауруса, основанная на анализе корпуса текста и вычислении сходства слов, близости их значений. Значение незнакомого слова часто можно определить по контексту [47]. Рассмотрим, например, следующий текст:

(1) *Бутылка Tezgüino стоит на столе. Всем нравится Tezgüino. Tezgüino может привести к опьянению. Мы делаем Tezgüino из зерна.*

Из этого контекста можно предположить, что *Tezgüino* — это алкогольный напиток, приготовленный из зерна.

Задача поиска похожих слов *(similar words)* является первым шагом в определении значения слова. Тогда при обработке корпуса, включающего предложение (1), результатом должно быть определение близости значения слова *Tezgüino* к словам *пиво, вино, водка.*

**Методология автоматического создания тезауруса.** Для вычисления сходства между словами в работе [35] использован парсер [36], извлекающий тройки из текста. Тройки зависимостей (от англ. *dependency triple*, далее просто *тройки*) состоят из двух слов и грамматического отношения между ними. Символ $||w, r, w'||$ означает частоту в корпусе тройки $(w, r, w')$, где $w, w'$ — это слова в нормальной форме, $r$ — синтаксическое отношение. Произвольное слово или отношение обозначается символом-джокером «*». Например, $||$*cook, obj,* *$||$ означает число троек со словом *cook* и отношением *obj.*

Например из предложения *«У меня есть коричневая собака»* будут извлечены следующие тройки:

$||$*коричневый, прил_ сущ, собака*$||$
$||$*есть, гл_ сущ, собака*$||$

Определим следующие моменты:

1. *Описание слова* $w$ — это частоты всех троек ($w$, *, *) в корпусе, то есть всех троек, включающих $w$. Описание слова $w$ является вектором.
2. *«Пересечение» двух слов* — это тройки, представленные в описании обоих слов; это пересечение векторов.

Сходство между двумя объектами вычисляется как количество информации в «пересечении» двух объектов (2), деленное на количество информации в описании двух объектов (1), далее обозначено как функция $sim(w_1, w_2)$ [37].

Предположив, что частоты троек не зависят друг от друга, получаем, что информация, представленная в описание слова $w$, равна сумме информации по каждой из уникальных троек в описании слова $w$.

Для измерения информации в утверждении $||w,\ r,\ w'||{=}c$ выполним следующее:

1. измерим количество информации в утверждении, что произвольная тройка, извлеченная из текста, будет наша тройка *(w, r, w')* при условии, что значение $||w,\ r,\ w'||$ — не известно;
2. измерим то же при условии, что значение $||w,\ r,\ w'||$ — известно;
3. разница этих двух количеств является ответом.

Вероятность встретить в тексте тройку *(w, r, w')* можно рассматривать как одновременное возникновение трех событий:
**A:** случайно выбранное слово - это $w$;
**B:** случайно выбранное отношение- это $r$;
**C:** случайно выбранное слово - это $w'$;

1. Когда значение $||w,\ r,\ w'||$ неизвестно, то предполагаем, что **A** и **C** являются условно независимыми при наличии события **B**. Вероятность наступления сразу трех этих событий составляет $P_{MLE}(B)P_{MLE}(A|B)P_{MLE}(C|B)$, где $P_{MLE}$ — это оценка максимального правдоподобия распределения вероятностей (*maximum likelihood estimation*)
$$P_{MLE}(B) = \frac{||*,r,*||}{||*,*,*||}$$
$$P_{MLE}(A|B) = \frac{||w,r,*||}{||*,r,*||}$$
$$P_{MLE}(C|B) = \frac{||*,r,w'||}{||*,r,*||}$$
2. Когда значение $||w,\ r,\ w'||$ известно, можно сразу получить $P_{MLE}(A,B,C)$:
$$P_{MLE}(A,B,C) = \frac{||w,r,w'||}{||*,*,*||}$$
3. Пусть ***I(w,r, w')*** обозначает количество информации, содержащейся в утверждении $||w,\ r,\ w'||{=}c$. Можно вычислить это значение так:
$$\begin{aligned} I(w,r,w') = \\ = -log(P_{MLE}(B)P_{MLE}(A|B)P_{MLE}(C|B)) - \\ - (-log(P_{MLE}(A,B,C)) = \\ = log\frac{||w,r,w'|| \times ||*,r,*||}{||w,r,*|| \times ||*,r,w'||}. \end{aligned}$$

Отметим, что значение *I(w,r,w')* равно количеству взаимной информации (*mutual information*) между $w$ и $w'$ [19].

Пусть $T(w)$ — это множество пар *(r, w')*, при которых $log\frac{||w,r,w'|| \times ||*,r,*||}{||w,r,*|| \times ||*,r,w'||}$ имеет положительное значение. Определим значение сходства (похожести) двух слов *w1* и *w2* с помощью формулы:

$$sim(w_1,w_2) = \frac{\sum\limits_{(r,w)\in T(w_1)\cap T(w_2)} (I(w_1,r,w) + I(w_2,r,w))}{\sum\limits_{(r,w)\in T(w_1)} I(w_1,r,w) + \sum\limits_{(r,w)\in T(w_2)} I(w_2,r,w)}.$$

**Практическая реализация метода.** Был обработан корпус, включающий 64 млн. слов. Из него было извлечено 56,6 миллионов троек, включающих 8,7 миллиона уникальных троек.

Сам корпус был разбит на классы по частям речи. Исследовалось попарно сходство между всеми глаголами, всеми существительными, всеми прилагательными/наречиями по формуле $sim(w_1,w_2)$. Для каждого слова был построен аналог словарной статьи в тезаурусе, включающий упорядоченный набор 200 наиболее похожих слов. Статья в тезаурусе для слова $w$ имела следующий формат:

$$w(pos) : w_1, s_1, w_2, s_2, \ldots, w_N, s_N$$

где *pos* — это часть речи, $w_i$ — это похожее слово, $s_i$ — это значение сходства между $w$ и $w_i$, слова упорядочены по убыванию значения сходства.

Два слова являются *парой взаимных ближайших соседей (RNN от respective nearest neighbors)*, если они являются наиболее похожими словами друг для друга (первыми в списке из двухсот слов). С помощью программы удалось получить 543 пары RNN существительных, 212 пар RNN глаголов, 382 пары RNN прилагательных/наречий в созданном автоматически тезаурусе. В Табл. 5 представлен список каждого 10-го RNN для глаголов.

## Выявление значений слов из текста — кластеризация посредством комитетов

*Д. Ю. Янкевич*

*Таблица 5.* Список пар взаимных ближайших соседей (RNN) глаголов

| **Ранг** | **RNN** | **Значение сходства** |
|---|---|---|
| 1 | *fall rise* | 0,67 |
| 11 | *injure kill* | 0,38 |
| 21 | *concern worry* | 0,34 |
| 31 | *convict sentence* | 0,29 |
| 41 | *limit restrict* | 0,27 |
| 51 | *narrow widen* | 0,26 |
| 61 | *attract draw* | 0,24 |
| 71 | *discourage encourage* | 0,23 |
| 81 | *hit strike* | 0,22 |
| 91 | *distregard ignore* | 0,21 |
| 101 | *overstate understate* | 0,20 |
| 111 | *affirm reaffirm* | 0,18 |
| 121 | *inform notify* | 0,17 |
| 131 | *differ vary* | 0,16 |
| 141 | *scream yell* | 0,15 |
| 151 | *laugh smile* | 0,143 |
| 161 | *compete cope* | 0,136 |
| 171 | *add whisk* | 0,130 |
| 181 | *blossom mature* | 0,12 |
| 191 | *smell taste* | 0,11 |
| 201 | *bark howl* | 0,10 |
| 211 | *black white* | 0,07 |

В статье [48] представлен алгоритм автоматического обнаружения значений слов в тексте, названный *кластеризация посредством комитетов* (Clustering By Committee, далее CBC). Также авторы предлагают методологию оценки для автоматического измерения точности и полноты найденных значений.

**Алгоритм** первоначально находит множество компактных кластеров, называемых комитетами, каждый из которых представляет собой одно из значений определяемого слова. Центр тяжести членов комитета (мера связности с определяемым словом) используется в качестве вектора признаков кластера.

Алгоритм CBC включает три этапа.

*На этапе I для каждого элемента (слова) вычисляется k наиболее похожих слов, строится база данных сходства S.* Сначала весь список относящихся к слову значений сортируется по убыванию значений связи согласно формуле точечной взаимной информации (pointwise mutual information или PMI [40]) (с. 66–68), а затем, с помощью иерархического кластерного анализа по *методу средней связи* [3], вычисляется сходство между всеми элементами кластера попарно. Значение функции PMI между предполагаемым значением слова (контекстом) и элементом (словом) вычисляется следующим образом: пусть $x$ — это рассматриваемый элемент, а $y$ — контекст. *Точечная взаимная информация* между $x$ и $y$ определена как:

$$pmi(x;y) \equiv log\frac{p(x,y)}{p(x)p(y)} = \\ = log\frac{p(x|y)}{p(x)} = log\frac{p(y|x)}{p(y)}.$$

При кластеризации посредством метода средних связей *(average-link clustering)* вычисляется среднее сходство между данным объектом и всеми объектами в кластере, а затем, если найденное *среднее значение сходства* достигает или превосходит некоторый заданный пороговый уровень сходства, объект присоединяется к этому кластеру [3]. Сложность этого алгоритма $O(n^2 \times log(n))$, где $n$ — число кластеризуемых элементов [32].

На II этапе Алгоритм 1 рекурсивно строит набор компактных кластеров, удаленных друг от друга, где элементы каждого кластера образуют комитет. В ходе работы Алгоритма 1 формируется как можно больше комитетов при условии, что каждый вновь созданный комитет не слишком похож на любой из уже существующих комитетов. Если условие нарушается, комитет просто отбрасывается.

На каждом рекурсивном уровне Алгоритм 1 находит компактный набор кластеров (назовем их — комитеты) и определяет оставшиеся элементы, не вошедшие ни в один из

комитетов. Будем говорить, что комитет «покрывает» элемент (или элемент «входит» в комитет), если значение сходства между элементом и центроидом комитета выше некоторой пороговой величины. При следующем рекурсивном вызове алгоритм снова ищет комитеты среди оставшихся элементов. На выходе Алгоритм 1 даёт список всех найденных комитетов.

На шаге 1 Алгоритма 1 поиска комитетов предпочтение отдается большим и компактным кластерам. На шаге 2 кластеры сортируются по значению сходства для последующего выбора лучшего кластера. На шаге 3 кластер сохраняется только в том случае, если его сходство со всеми ранее полученными кластерами ниже установленного порогового значения (в экспериментах было получено значение $\theta_1 = 0.35$). На шаге 4, если не было найдено комитетов на предыдущем шаге, рекурсия останавливается. Оставшиеся и никуда не вошедшие элементы (остатки) определяются на шаге 5. Если таких остатков нет, то алгоритм завершается, иначе — алгоритм вызывается рекурсивно для остатков.

В результате второго этапа построения CBC строятся плотные компактные кластеры (имеющие бóльшее значение *val*, см. шаги 1 и 3 Алгоритма 1), хорошо отличающиеся друг от друга. На третьем этапе все элементы распределяются по этим кластерам, а именно: каждый элемент $e$ присваивается наиболее близкому кластеру, при этом центроид членов комитета используется в качестве вектора характеристик кластера (Алгоритм 2). Центроиды не изменяются, то есть при добавлении элемента в кластер, элемент не добавляется в комитет кластера.

Алгоритм CBC полагается на дистрибутивную гипотезу. Алгоритм CBC разрешает лексическую многозначность, группируя слова согласно сходству их контекстов. Каждому полученному кластеру соответствует одно из значений слова.

**Сравнение с алгоритмом UNICON.** CBC является разновидностью алгоритма UNICON [38], который также строит центроид кластера, используя небольшой набор похожих элементов.

Одним из основных различий между UNICON и CBC является то, что UNICON гарантирует, что различные комитеты не имеют одинаковых элементов, тем не менее, центры тяжести двух комитетов по-прежнему могут быть очень близкими (похожими). В UNICON'е эта проблема решается объединением таких кластеров. В отличие от этого, на II этапе CBC создаются только те комитеты, центры тяжести которых отличны от всех ранее созданных комитетов.

Есть разница и на III этапе CBC. Алгоритм UNICON плохо работает со словами, которые имеют несколько широко используемых (доминирующих) значений. Например, пусть значение «отмычка» является более употребимым для слова «ключ», чем значение «водный источник». Приведем смесь слов-синонимов к разным значениям слова «ключ»: пневмоключ, электроключ, родник, родничок, источник, криница, гидроключ, ключик, тангента, треншальтер, тумблер, знак, контролька, отпирка, виброплекс, шифр. В этом списке 10 значений относятся к значению «отмычка()», 4 к «водный источник()» и 2 к значению «знак()». По этому списку алгоритмом UNICON будут сгенерированы кластеры «отмычка», «шифрование», «происхождение»,«криптография», «кнопка», «водный источник»,«переключатель», «намек». Сходство между словом и полученными кластерами является очень низким, к тому же есть кластеры, содержащие одинаковые слова. С другой стороны, CBC удаляет «пересекающиеся» (общие для двух кластеров) характеристики после того, как присвоит значение кластеру (допустим характеристики, относящиеся к значению «отмычка» слова «ключ» из вектора характеристик «отмычка»). В результате сходство между кластером «водный источник» {родник, родничок, источник, криница} и пересмотренным вектором характеристик кластера «водный источник» становится намного выше. Что, в свою очередь, приводит к тому, что кластеры становятся гораздо точнее.

## Эксперименты

### Сравнительные эксперименты в WSD: роль предпочтений в машинном обучении

*Н. И. Коржицкий*

В работе Рэймонда Муни [43] представлено одно из первых сравнений разных по природе методов WSD на одних и тех же данных. В статье [43] проведена серия экспериментов, в которых сравнивалась способность различ-

**Data**: $(E, S, \theta_1, \theta_2)$, где $E$ — это список элементов, которые будут сгруппированы, база данных сходства $S$ (построена в ходе этапа I), пороги $\theta_1$ и $\theta_2$ (с помощью порога $\theta_1$ сохраняются только те кластеры, которые имеют значения, отличные от ранее обнаруженных, порог $\theta_2$ позволяет обнаружить элементы, не принадлежащие ни одному из кластеров)
**Result**: $C$ — список комитетов
**Step 1:**
**foreach** $e \in E$ **do**
1. Кластер $k$ наиболее «близких» (похожих) элементов $e$ из $S$ с помощью метода средней связи.
2. Для каждого обнаруженного кластера $c$ вычислить следующую оценку: $val = |C| \times avgsim(c)$, где $|C|$ — количество элементов $c$ и $avgsim(c)$ — усредненное сходство между всеми парами элементов кластера $c$.
3. Записать кластер с наивысшей оценкой в список $L$.

**end**
**Step 2:**
SortByDecreasingOrder($c(val) \in L$) // Сортировка кластеров в списке $L$ в порядке убывания их оценок $val$.
**Step 3:**
$C = \varnothing$ // Пусть перечень комитетов $C$ будет изначально пустым.
**foreach** $c \in L$ **do**
// в отсортированном по убыванию порядке:
1. Вычислить центр тяжести, усредняя поэлементно значение векторов, и вычислить вектор PMI центроида (так же, как и для отдельных элементов на шаге 1).
2. Если схожесть $c$ и центроида каждого комитета, ранее добавленного к $C$, ниже порогового $\theta_1$, то следует добавить $c$ в $C$.

**end**
**Step 4:**
**if** $C = \varnothing$ **then**
  **return** $C$
**end**
**Step 5:**
$R = \varnothing$ // $R$ — это множество остатков, то есть элементов, не охваченных ни одним из кластеров
**foreach** $e \in E$ **do**
  **if** $sim(e,$ *foreach* $c \in C) < \theta_2$ // *сходство по всем комитетам из* $C$ *меньше* $\theta_2$ // **then**
    $R+ = e$ // то следует добавить $e$ в список остатков $R$.
  **end**
**end**
**Step 6:**
**if** $R = \varnothing$ **then**
  **return** $C$ **else**
    **return** $C \cup Algorithm1(R, S, \theta_1, \theta_2)$
  **end**
**end**

**Algorithm 1:** Этап II. Поиск комитетов

**Result**: Итоговые кластеры с максимальным значением связи *val* между словами (см. вычисление *val* на шаге 1 и 3 Алгоритма 1)
Пусть $C$ — это список кластеров (изначально пустых).
Пусть $S$ — это первые 200 кластеров наиболее похожих на $e$ (база данных сходства $S$ построена в ходе этапа I).
**while** $S \neq \varnothing$ **do**
  пусть $c \in S$ наиболее близкий кластер к $e$
  **if** *сходство* $(e, c) < \sigma$ **then**
    **конец цикла**
  **end**
  **if** *с не схож ни с одним кластером в* $C$ **then**
    присвоить $e$ к $c$;
    удалить из $e$ его характеристики, которые перекрываются с характеристиками $c$;
  **end**
  удалить $c$ из $S$
**end**

**Algorithm 2:** Этап III. Присвоение элементов кластерам: для каждого из элементов $e$ находится наиболее близкий кластер, в который включается $e$

ных обучающихся алгоритмов определять значение слова в зависимости от контекста.

В машинном обучении под термином *bias* (пристрастие, тенденция, предпочтение) понимается любое основание для выбора одного обобщения другому, вместо строгого соответствия примерам [43]. В деревьях принятия решений предпочтение (*bias*) отдается простым деревьям решений, в нейронных сетях — линейным пороговым функциям, а в байесовском классификаторе — функциям, учитывающим условную независимость свойств. Чем лучше «предпочтение» обучающегося алгоритма соответствует характеристикам конкретной задачи, тем лучше будет результат. Большинство обучающихся алгоритмов обладают «предпочтением» наподобие Бритвы Оккама, в таких алгоритмах выбираются гипотезы, которые могут быть представлены меньшим количеством информации на каком-нибудь языке представлений. Однако компактность, с которой (деревья решений, дизъюнктивная нормальная форма, сети с линейным пороговым значением) представляют конкретные функции — может существенно различаться. Поэтому различные «предпочтительные» оценки могут работать лучше или хуже в конкретных задачах. Одной из основных целей в машинном обучении является поиск «предпочтений» с целью решения прикладных практических задач.

Выбор правильного «предпочтения» и обучающегося алгоритма является сложной задачей. Простым подходом является автоматизация выбора метода на основе результатов внутренней перекрестной валидации. Другой подход, который называется мета-обучением (meta-learning), заключается в том, чтобы сформировать набор правил (или аналогичный классификатор), который бы на основании предметных признаков, описывающих задачу, предсказывал бы, когда обучающийся алгоритм будет срабатывать наилучшим образом.

Описанный в [43] эксперимент заключается в определении значения слова *line* (англ. *линия*) среди 6 возможных вариантов (*строка, ряд, дивизия, телефон, веревка, продуктовая линия*). Данные для проведения экспериментов взяты из работы [33].

Для получения обучающей выборки брались предложения со словом *line*, и им в соответствие ставилось одно из 6 значений. Распределение значений неравномерно: включение в список иточников журнала The Wall Street Journal привело к тому, что одно из значений встречалось в 5 раз чаще всех остальных [33].

В работе [39] было установлено, что наиболее эффективными при решении задачи WSD являются алгоритмы на основе дерева решений (decision tree). Данный класс методов обходил по точности и скорости работы класс нейронных сетей. Другие исследования [44] показали, что класс методов индуктивного логического программирования (inductive logic programming) справляется с задачей разрешения лексической многозначности слова лучше алгоритмов на основе дерева решений.

В серии экспериментов в [43] сравнивались следующие методы: байесовский классификатор, перцептрон, C4.5, метод k-ближайших соседей и модификации алгоритма FOIL: PFOIL-DLIST, PROIL-DNF, PFOIL-CNF.

После проведения сравнительных экспериментов, заключавшихся в обучении и определении значения слова *line,* было выяснено, что

*Таблица 6.* Шесть значений слова *line* из Английского Викисловаря и Русского Викисловаря

| **ключевое слово** | **перевод** | **толкование на английском (Английский Викисловарь)** | **толкование на русском (Русский Викисловарь)** |
|---|---|---|---|
| text | строка | A small amount of text | ряд слов, букв или иных знаков, написанных или напечатанных в одну линию |
| formation | ряд | A more-or-less straight sequence of people, objects, etc. | несколько объектов, расположенных в линию или следующих один за другим |
| division | дивизия | A formation, usually made up of two or three brigades | тактическое воинское соединение |
| phone | телефон | The wire connecting one telegraphic station with another, a telephone or internet cable | то же, что телефонный номер |
| cord | веревка | A rope, cord, string, or thread, of any thickness | гибкое и длинное изделие, — чаще всего сплетенное или свитое из льняных (или пеньковых, полимерных и т. п.) волокон или прядей |
| product | продуктовая линия | The products or services sold by a business, or by extension, the business itself | совокупность однородной продукции единого назначения |

байесовский классификатор и перцептрон работают точнее других рассмотренных методов.

Эксперименты проводились с разными размерами обучающей выборки для того, чтобы выяснить, какого рода зависимость имеет место между точностью определения значения и размером выборки. На Рис. 5c отображена зависимость точности работы алгоритмов от размера выборки. При увеличении размера обучающей выборки сначала происходит резкий рост точности, последующий прирост точности становится незначительным.

Эксперименты учитывали не только точность определения значения, но и требовательность алгоритма к ресурсам в процессе обучения и работы. На Рис. 5a можно увидеть зависимость времени обучения от размера выборки. Самыми быстрообучаемыми оказались байесовский классификатор и перцептрон, а самыми медленными — нормальные формы (Рис. 5a).

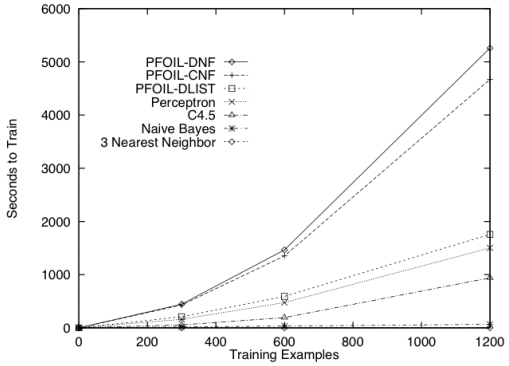
На Рис. 5b представлена зависимость времени работы алгоритмов от размера обучающей выборки. Время работы алгоритмов дает другую картину: байесовский классификатор и перцептрон работают долго при максимальном размере обучающей выборки, в то время как остальные методы решают WSD задачу за постоянное время (Рис. 5b).

## Заключение

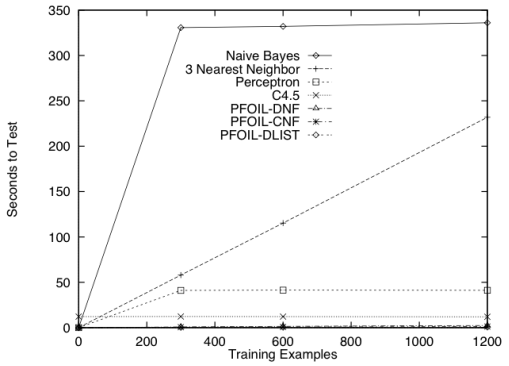
Разрешение лексической многозначности - это задача выбора между разными значениями слов и словосочетаний в словаре в зависимости от контекста. Задача разрешения лексической многозначности является открытой проблемой, то есть крайне интересной и привлекательной с научной точки зрения.

В статье представлен краткий обзор методов и алгоритмов, применяемых для разрешения лексической многозначности. Эти методы используют различный математический и алгоритмический аппарат для решения WSD-задачи: нейронные сети, адаптивный алгоритм улучшения точности обучения AdaBoost, построение лексических цепочек, методы на основе применения теоремы Байеса и методы кластеризации контекстных векторов и семантически близких слов. Работу завершает исследование, в котором сравниваются время обучения, время работы и результаты работы разных алгоритмов решения WSD-задачи.

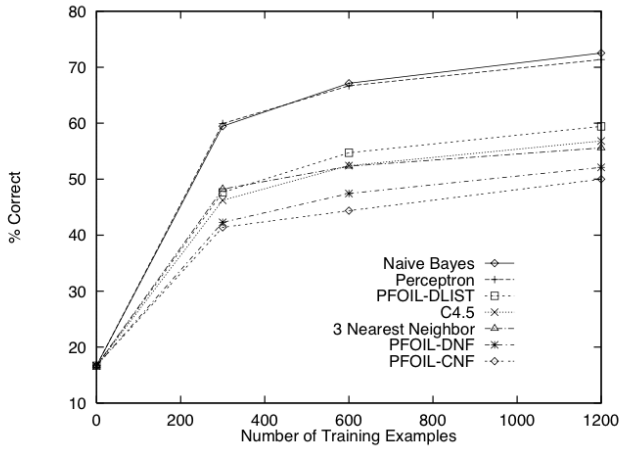
Работа Старковой В.Г. поддержана грантом РГНФ (проект № 15-04-12029), работа Кириллова А.Н. и Чирковой Ю.В. поддержана




(a) Зависимость времени затраченного на обучение алгоритмов от размера обучающей выборки




(b) Зависимость времени работы алгоритмов от размера обучающей выборки




(c) Рост точности решения WSD-задачи для разных алгоритмов при увеличении размера обучающей выборки

*Рис. 5.* Сравнение времени обучения, времени работы и результатов работы алгоритмов PFOIL-DLIST, PROIL-DNF, PFOIL-CNF, C4.5, Naive Bayes — наивный баейсовский классификатор; Perceptron — персептрон; 3 Nearest Neighbor — метод 3-х ближайших соседей при определении значения слова *line* [43]

грантом РГНФ (проект № 15-04-12006). Работа Крижановского А. А. выполнена при частичной финансовой поддержке Программы фундаментальных исследований Секции литературы и языка ОИФН РАН «Язык и информационные технологии» 2015-2017 (проект «Корпус вепсского языка: разработка и формирование морфологической базы электронного ресурса»).

## Литература

1. *Аверин А. Н.* Разработка сервиса поиска биграмм // Труды международной конференции «Корпусная лингвистика–2006. СПб., С.Петерб. ун-та., 2006. С. 5 - 15.

2. *Епрев А. С.* Применение контекстных векторов в классификации текстовых документов // "Журнал радиоэлектроники". 2010. N 10. URL: http://jre.cplire.ru/iso/oct10/1/text.html (дата обращения: ).

3. *Ким Дж. О., Мьюллер Ч. У., Клекка У. Р.* Факторный, дискриминантный и кластерный анализ / «Финансы и статистика», Москва, Россия. 1989. Стр.172.

4. *Лукашевич Н. В.* Тезаурусы в задачах информационного поиска / Издательство МГУ, 2011. 495 с.

5. *Марманис Х., Бабенко Д.* Алгоритмы интеллектуального Интернета. Передовые методики сбора, анализа и обработки данных. – Пер. с англ. – СПб.: Символ-Плюс, 2011. 480 с.

6. *Паклин Н. Б., Орешков В. И.*Бизнес-аналитика: от данных к знаниям: Учебное пособие. 2-е изд., испр. — СПб.: Питер, 2013. — 704 с.

7. *Турдаков Д. Ю.* Методы и программные средства разрешения лексической многозначности терминов на основе сетей документов: дис. . . . канд. физико-математических наук: Москва, 2010. 138 с.

8. *Abney S. and Light M.* Hiding a semantic hierarchy in a markov model. In Proceedings of the Workshop on Unsupervised Learning in Natural Language Processing, ACL. 1999.

9. *Azzini A., da Costa Pereira C., Dragoni M. and Tettamanzi A. G. B.* Evolving Neural Networks for Word Sense Disambiguation // 8-th International conference on hybrid intelligent systems. Spain. Barcelona, 2008. P. 332–337. doi: 10.1109/HIS.2008.88

10. *Barzilay R. and Elhadad M.* Using lexical chains for text summarization // In Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization (Madrid, Spain). 1997. P. 10–17.

11. *Berry M., Do T., O'Brien G., Krishna V. and Varadhan S.* SVDPACK (version 1.0) user's guide. Technical Report CS-93-194, University of Tennessee at Knoxville, Computer Science Department, April 1993.

12. *L. Breiman.* Arcing classifiers. The Annals of Statistics. Vol 26 (3), 1998, pp. 801–849.

13. *Bruce R. and Wiebe J.* Word-sense disambiguation using decomposable models // In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, 1994. P. 139–146. doi: 10.3115/981732.981752

14. *Carpuat M. and Wu D.* Evaluating the word sense disambiguation performance of statistical machine translation. In Proceedings of the second international joint conference on natural language processing (IJCNLP), 2005, pp. 122–127. URL: http://www.aclweb.org/anthology/I05-2021

15. *Ciaramita M. and Johnson M.* Explaining away ambiguity: Learning verb selectional preference with Bayesian networks. In Proceedings of the 18th conference on Computational linguistics, vol. 1, 2000, pp. 187–193.

16. *Cottrell G. W. and Small S. L.* A connectionist scheme for modelling word sense disambiguation // Cognition and brain theory. 1983. № 6. P. 89–120.

17. *Cottrell G. W.* A connectionist approach to word sense disambiguation / Pitman, London, 1989.

18. *Duong D. T.* Automated text summarization. Graduation Thesis. Hanoi University. 2011.

19. *Hindle D.* Noun classification from predicate-argument structures // In Proceedings of ACL-90, Pittsburg, Pennsylvania, June, 1990. P. 268-275.

20. *Escudero G., Màrquez L. and Rigau G.* Using LazyBoosting for word sense disambiguation. In Proceedings of the Second International Workshop on evaluating Word Sense Disambiguation Systems. Toulouse, France, 2001, pp. 71–74.

21. *Escudero G., Màrquez L. and Rigau G.* Boosting Applied to Word Sense Disambiguation. In Proceedings of the 12th European Conference on Machine Learning, ECML. Barcelona, Catalonia. 2000.

22. *Freund Y., Schapire R. E.* A Short Introduction to Boosting // AT&T Labs Research, Shannon Laboratory. 1999.

23. *Freund Y., Schapire R. E.* Game theory, on-line prediction and boosting // In Proceedings of the Ninth Annual Conference on Computational Learning Theory, 1996. P. 325-332.

24. *Freund Y., Schapire R. E.* A decision-theoretic generalization of on-line learning and an application to boosting // Journal of Computer and System Sciences. 1997. P. 119–139. doi: 10.1006/jcss.1997.1504

25. *Halliday M. and Hasan R.* Cohesion in English / London: Longman. 1976.

26. *Harris Z.* Distributional structure / In: Katz, J. J. (ed.) The Philosophy of Linguistics. New York: Oxford University Press. 1985. P. 26–47

27. *Hearst M.* Multi-paragraph segmentation of expository text // In Proceedings of the 32th Annual Meeting of the Association for Computational Linguistics, 9–16. Las Cruces, New Mexico: Association for Computational Linguistics. 1994. doi: 10.3115/981732.981734

28. *Hinton G. E., McClelland J. L., Rumelhart D. E.* Distributed representations // In Parallel Processing: explorations in the microstructure of cognition. MIT Press, Cambridge, MA, 1986. P. 5–44.

29. *Hirst G. and St-Onge D.* Lexical chains as representations of context for the detection and correction of malapropisms. WordNet: An electronic lexical database, 1998. P. 305-332.

30. *Hoey M.* Patterns of Lexis in Text / Oxford: Oxford University Press. 1991.

31. *Jain A. and Dubes R.* Algorithms for Clustering Data / Prentice-Hall, Inc., Upper Saddle River, NJ, 1988.

32. *Jain A., Murthy M. and Flynn P.* Data clustering: a review // ACM Computing Surveys, 31(3):264-323, September 1999. doi: 10.1145/331499.331504

33. *Leacock C., Towell G. and VoorheesE.* Corpus-based statistical sense resolution // In Proceedings of the ARPA Workshop on Human Language Technology, March. 1993, P. 260–265.

34. *Lesk M.* Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone // Proceedings of the 5th SIGDOC. New York. 1986. P. 24–26. doi: 10.1145/318723.318728

35. *Lin D.* Automatic Retrieval and Clustering of Similar Words // Proceedings of the 17th international conference on Computational linguistics-Vol. 2. – Department of Computer Science University of Manitoba Winnipeg, Manitoba, Canada, 1998, P. 768-774. doi: 10.3115/980432.980696

36. *Lin D.* Principle-based parsing without overgeneration // In Proceedings of ACL-93, Columbus, Ohio, 1993, P. 112-120. doi: 10.3115/981574.981590

37. *Lin D.* Using syntactic dependency as local context to resolve word sense ambiguity // In Proceedings of ACL/EACL-97, Madrid, Spain, July, 1997, P. 64-71. doi: 10.3115/979617.979626

38. *Lin D. and Pantel P.* Induction of semantic classes from natural language text // In Proceedings of SIGKDD-01. San Francisco, CA. 2001. P. 317–322. doi: 10.1145/502512.502558

39. *Ling Charles X., Marinov M.* Answering the connectionist challenge: A symbolic model of learning the past tenses of English verbs / Cognition, Elsevier, 1993.

40. *Manning C. D. and Schütze H.* Foundations of Statistical Natural Language Processing / MIT Press. 1999.

41. *Merz C. J. and Murphy P. M.* UCI repository of machine learning databases, 1998. URL: www.ics.uci.edu/ mlearn/MLRepository.html (дата обращения: 24.04.2015).

42. *Miller G.* Wordnet: An on-line lexical database // International Journal of Lexicography, 3(4). 1990.

43. *Mooney R. J.* Comparative Experiments on Disambiguating Word Senses: An Illustration of the Role of Bias in Machine Learning / Department of Computer Sceinces, University of Texas, Austin, TX 78712-1188, 1996.

44. *Mooney R. J., Califf M. E.* Induction of First-Order Decision Lists: Results on Learning the Past Tense of English Verbs / Department of Computer Sceinces, University of Texas, Austin, TX 78712-1188, 1995.

45. *Morris J. and Hirst G.* Lexical cohesion computed by thesaural relations as an indicator of the structure of text // Computational Linguistics 17(1):21–43. 1991.

46. *Navigli R.* Word sense disambiguation: A survey. ACM Computing Surveys (CSUR) 41, no. 2 (2009): 10. doi: 10.1145/1459352.1459355

47. *Nida Eugene A.* Componential Analysis of Meaning / The Hague, Mouton. 1975.

48. *Pantel P., Lin D.* Discovering Word Senses from Text / University of Alberta. Department of Computing Science Edmonton, Alberta T6H 2E1 Canada, 2002. doi: 10.1145/775047.775138

49. *Pedersen T.* A Simple Approach to Building Ensembles of Naive Bayesian Classifers for Word Sense Disambiguation / Department of Computer Science, University of Minnesota Duluth. 2000.

50. *Pedersen T. and Bruce R.* Distinguishing word senses in untagged text / Proc.EMNLP.Providence, RI, 1997.

51. *Purandare A. and Pedersen T.* Improving word sense discrimination with gloss augmented feature vectors // Workshop on Lexical Resources for the Web and Word Sense Disambiguation. 2004. P. 123-130.

52. *Quinlan J. R.* C4.5: Programs for Machine Learning / Morgan Kaufmann, 1993.

53. *Resnik P.* Selectional preference and sense disambiguation / In Proceedings of the ANLP-97 Workshop: Tagging Text with Lexical Semantics: Why, What, and How? 1997.

54. *Savova G., Pedersen T., Purandare A., Kulkarni A.* Resolving ambiguities in biomedical text with unsupervised clustering approaches /

University of Minnesota Supercomputing Institute Research Report, 2005.

55. *Schapire R. E. and Singer Y.* Improved Boosting Algorithms Using Confidence-rated Predictions. Machine Learning, vol. 37(3), 1999, pp. 297–336.

56. *Schapire R. E. and Singer Y.* Improved boosting a predictions // In Proceedings of the Eleventh Annual Confere Theory, 1998. P. 80–91.

57. *Schapire R. E.* Using output codes to boost multiclass learning problems. In Machine Learning // Proceedings of the Fourteenth International Conference, 1997. P. 313–321.

58. *Schütze H.* Automatic Word Sense Discrimination // Computational Linguistics, vol. 24, number 1., 1998.

59. *SenseClusters.*
URL: http://senseclusters.sourceforge.net (дата обращения: 24.04.2015).

60. *UMLS Terminology Services (UTS).* URL: http://umlsks.nlm.nih.gov/kss/servlet/Turbine/template (дата обращения: 22.04.2015)

61. *Veronis J. and Ide N.* Word sense disambiguation with very large neural networks extracted from machine readable dictionaries // Proceedings of the 13th International Conference on Computational Linguistics. Helsinki. 1990. P. 389–394. doi: 10.3115/997939.998006

62. *Waltz D. L. and Pollack J. B.* Massively parallel parsing: a strongly interactive model of natural language interpretation // Cognitive science. 1985. № 9. P. 51–74. doi: 10.1207/s15516709cog0901_4

63. *Weeber M., Mork J., Aronson A.* Developing a test collection for biomedical word sense disambiguation / Proc. AMIA., 2001.

64. *Zhao Y. and Karypis G.* Evaluation of hierarchical clustering algorithms for document datasets // In Proceedings of the 11th International Conference on Information and Knowledge Management, McLean, VA, 2002. P. 515-524. doi: 10.1145/584792.584877

## СВЕДЕНИЯ ОБ АВТОРЕ:

**Каушинис Татьяна Викторовна**
Студентка
Математический факультет
Петрозаводский государственный университет
пр-кт Ленина, 33, Петрозаводск, Республика Карелия
тел.: (8142) 711078
эл. почта: merilstreet@mail.ru

**Кириллов Александр Николаевич**
Доктор физико-математических наук
доцент
Институт прикладных математических исследований Карельского научного центра РАН
ул. Пушкинская, 11, Петрозаводск, Республика Карелия, Россия, 185910
эл. почта: kirillov@krc.karelia.ru
тел.: (8142) 766312

**Коржицкий Никита Иванович**
Студент
Математический факультет
Петрозаводский государственный университет
пр-кт Ленина, 33, Петрозаводск, Республика Карелия
тел.: (8142) 711078
эл. почта: nikita@nikita.tv

**Крижановский Андрей Анатольевич**
Кандитат технических наук
Институт прикладных математических исследований Карельского научного центра РАН
ул. Пушкинская, 11, Петрозаводск, Республика Карелия, Россия, 185910
эл. почта: andew.krizhanovsky@gmail.com
тел.: (8142) 766312

**Пилинович Александр Владимирович**
Студент
Математический факультет
Петрозаводский государственный университет
пр-кт Ленина, 33, Петрозаводск, Республика Карелия
тел.: (8142) 711078
эл. почта: alexander.pilinovich@yandex.ru

**Kaushinis, Tatiana**
Petrozavodsk State University
33, Lenin Str., 185910, Petrozavodsk, Republic of Karelia, Russia
tel.: (8142) 711078
e-mail: merilstreet@mail.ru

**Kirillov, Alexander**
Institute of Applied Mathematical Research, Karelian Research Centre, Russian Academy of Sciences
11, Pushkinskaya St., 185910 Petrozavodsk, Karelia, Russia
e-mail: kirillov@krc.karelia.ru
tel.: (8142) 766312

**Korzhitsky, Nikita**
Petrozavodsk State University
33, Lenin Str., 185910, Petrozavodsk, Republic of Karelia, Russia
tel.: (8142) 711078
e-mail: nikita@nikita.tv

**Krizhanovsky, Andrew**
Institute of Applied Mathematical Research, Karelian Research Centre, Russian Academy of Sciences
11, Pushkinskaya St., 185910 Petrozavodsk, Karelia, Russia
e-mail: andew.krizhanovsky@gmail.com
tel.: (8142) 766312

**Pilinovich, Aleksander**
Petrozavodsk State University
33, Lenin Str., 185910, Petrozavodsk, Republic of Karelia, Russia
tel.: (8142) 711078
e-mail: alexander.pilinovich@yandex.ru

**Сихонина Ирина Александровна**
Студентка
Математический факультет
Петрозаводский государственный университет
пр-кт Ленина, 33, Петрозаводск, Республика Карелия
тел.: (8142) 711078
эл. почта: syawenka@mail.ru

**Sikhonina, Irina**
Petrozavodsk State University
33, Lenin Str., 185910, Petrozavodsk, Republic of Karelia, Russia
tel.: (8142) 711078
e-mail: syawenka@mail.ru

**Спиркова Анна Михайловна**
Студентка
Математический факультет
Петрозаводский государственный университет
пр-кт Ленина, 33, Петрозаводск, Республика Карелия
тел.: (8142) 711078
эл. почта: annspirkova@gmail.com

**Spirkova, Anna**
Petrozavodsk State University
33, Lenin Str., 185910, Petrozavodsk, Republic of Karelia, Russia
tel.: (8142) 711078
e-mail: annspirkova@gmail.com

**Старкова Валентина Геннадьевна**
Старший инженер-программист
Институт прикладных математических исследований КарНЦ РАН
ул. Пушкинская, 11, Петрозаводск, Республика Карелия, Россия, 185910
тел.: (8142) 766312
эл. почта: stark_val@mail.ru

**Starkova, Valentina**
Institute of Applied Mathematical Research, Karelian Research Centre, Russian Academy of Sciences
11, Pushkinskaya St., 185910 Petrozavodsk, Karelia, Russia
tel.: (8142) 766312
e-mail: stark_val@mail.ru

**Степкина Татьяна Владимировна**
Студентка
Математический факультет
Петрозаводский государственный университет
пр-кт Ленина, 33, Петрозаводск, Республика Карелия
тел.: (8142) 711078
эл. почта: hogdp@mail.ru

**Stepkina, Tatiana**
Petrozavodsk State University
33, Lenin Str., 185910, Petrozavodsk, Republic of Karelia, Russia
tel.: (8142) 711078
e-mail: hogdp@mail.ru

**Ткач Станислав Сергеевич**
Студент
Математический факультет
Петрозаводский государственный университет
пр-кт Ленина, 33, Петрозаводск, Республика Карелия
тел.: (8142) 711078
эл. почта: tkachkras@gmail.com

**Tkach, Stanislav**
Petrozavodsk State University
33, Lenin Str., 185910, Petrozavodsk, Republic of Karelia, Russia
tel.: (8142) 711078
e-mail: tkachkras@gmail.com

**Чиркова Юлия Васильевна**
Кандидат физико-математических наук
Институт прикладных математических исследований КарНЦ РАН
ул. Пушкинская, 11, Петрозаводск, Республика Карелия, Россия, 185910
тел.: (8142) 766312 эл. почта: julia@krc.karelia.ru

**Chirkova, Julia**
Institute of Applied Mathematical Research, Karelian Research Centre, Russian Academy of Sciences
11 Pushkinskaya St., 185910 Petrozavodsk, Karelia, Russia
tel.: (8142) 766312 e-mail: julia@krc.karelia.ru

**Чухарев Алексей Леонидович**
Старший инженер-программист
Институт прикладных математических исследований КарНЦ РАН
ул. Пушкинская, 11, Петрозаводск, Республика Карелия, Россия, 185910
тел.: (8142) 766312 эл. почта: chuharev@krc.karelia.ru

**Chuharev, Alexey**
Institute of Applied Mathematical Research, Karelian Research Centre, Russian Academy of Sciences
11, Pushkinskaya St., 185910 Petrozavodsk, Karelia, Russia
tel.: (8142) 766312 e-mail: chuharev@krc.karelia.ru

**Шорец Дарья Сергеевна**
Студентка
Математический факультет
Петрозаводский государственный университет
пр-кт Ленина, 33, Петрозаводск, Республика Карелия
тел: (8142) 711078
эл. почта: da_sha1078@mail.ru

**Shorets, Daria**
Petrozavodsk State University
33, Lenin Str., 185910, Petrozavodsk, Republic of Karelia, Russia
tel.: (8142) 711078
e-mail: da_sha1078@mail.ru

**Ярышкина Екатерина Александровна**
Студентка
Математический факультет
Петрозаводский государственный университет
пр-кт Ленина, 33, Петрозаводск, Республика Карелия
тел:(8142) 711078
эл. почта: kate.rysh@gmail.com

**Янкевич Дарья Юрьевна**
Студентка
Математический факультет
Петрозаводский государственный университет
пр-кт Ленина, 33, Петрозаводск, Республика Карелия
тел: (8142) 711078
эл. почта: dyankevic@gmail.com

**Yaryshkina, Ekaterina**
Petrozavodsk State University
33, Lenin Str., 185910, Petrozavodsk, Republic of Karelia, Russia
tel.: (8142) 711078
e-mail: kate.rysh@gmail.com

**Yankevich, Daria**
Petrozavodsk State University
33, Lenin Str., 185910, Petrozavodsk, Republic of Karelia, Russia
tel.: (8142) 711078
e-mail: dyankevic@gmail.com