Труды Карельского научного центра РАН
№0. 2010. С. 1–4

*УДК 004.01:006.72 (470.22)*

# ОБЗОР МЕТОДОВ (УТОЧНИТЬ - КАКИХ?) ПО РЕШЕНИЮ ЗАДАЧИ РАЗРЕШЕНИЯ ЛЕКСИЧЕСКОЙ МНОГОЗНАЧНОСТИ

**С. С. Ткач, Е. А. Ярышкина, А. Н. Кирилов**

*Институт прикладных математических исследований Карельского научного центра РАН*

Данный файл является примером статьи для научного издания Труды Карельского научного центра РАН, серия «Математическое моделирование и информационные технологии». В нем содержатся основные используемые переменные и окружения. При подготовке статьи рекомендуется воспользоваться этим примером в качестве шаблона. Данный абзац оформлен в стиле аннотации статьи.

Ключевые слова: труды, шаблон, подготовка статьи.

## S. S. Tkach, E. A. Yaryshkina, A. N. Kirilov. WORD-SENSE DISAMBIGUATION METHODS (SPECIFIC?) REVIEW

This file is an auxiliary example of an article prepared forTransactions of Karelian Research Centre of RAS. It contains most useful environments and variables. While preparing your article, it's recommended to use this text as a template. This paragraph is formatted as an Abstract of the article.

Key words: transactions, template, article.

---

## Введение

Статья, представляемая в научное издание, должна быть оформлена в соответствии с «Правилами для авторов», размещенными на сайте http://transactions.krc.karelia.ru. Данный документ претендует на роль технической документации в помощь авторам. Достаточно подробную информацию по наборе в системе LaTeXможно найти, напр., в работе [2].

## Структура файла в формате LaTeX 2ε

```
\documentclass{article}
\usepackage{krctran}
...
\begin{document}

\procname{...}
\udk{...}
\rustitle{...}
\engtitle{...}
\rusauthor{...}
\engauthor{...}
\organization{...}
\rusabstract{...}
\engabstract{...}
\ruskeywords{...}
\engkeywords{...}

\maketitle

\begin{articletext}
\section{...}
...
\begin{thebibliography}
...
```

*Рис. 1.* Задержки в модели 10-узловой системы, случай тяжелых хвостов

```
\end{thebibliography}
\end{articletext}

\section{СВЕДЕНИЯ ОБ АВТОРЕ:}
\begin{aboutauthors}
...
\end{aboutauthors}
\end{document}
```

Преамбула статьи должна содержать две обязательные команды:

```
\documentclass{article}
\usepackage{krctran}
```

Далее формируется заголовок статьи: выходные данные, код УДК, название статьи, авторы с указанием мест работы, аннотация, ключевые слова. Например, заголовок этого файла сформирован следующими командами:

```
\procname{Труды...\\ \No...}
\udk{УДК...}
\rustitle{Руководство...}
\engtitle{Usage...}
\rusauthor{А.~С.~Румянцев}
\engauthor{A.~S.~Rumyantsev}
\organization{Институт...}
\rusabstract{Данный файл...}
\engabstract{This file...}
\ruskeywords{труды,...}
\engkeywords{transactions,...}
\maketitle
```

**Замечание 1.** Для ручной разбивки на строки названия статьи воспользуйтесь командой `\newline`.

## Обзор статьи «Resolving Ambiguities in Biomedical Text With Unsupervised Clustering Approaches»

В статье [8] 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 так же и обнаружения новых значений для них, ещё не включённых в данную онтологию.

Авторы описали методологию метода решения лексической многозначности без учителя, учитываемые лексические признаки и наборы экспериментальных данных. В качестве оценки эффективности алгоритмов кластеризации текста была предложена F-мера.

Подход для решения поставленной задачи – это разделение контекстов (фрагментов текста), содержащих определенное целевое слово на кластеры, где каждый кластер представляет собой различные значения целевого слова. Каждый кластер состоит из близких по значению контекстов. Задача решается в предположении, что используемое целевое слово в аналогичном контексте будет иметь один и тот же или очень похожий смысл.

Процесс кластеризации продолжается до тех пор, пока не будет найдено предварительно заданное число кластеров. В данной статье выбор шести кластеров 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 [4]. 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 [3].

Экспериментальные данные – это набор NLM WSD [12] (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 [7], SC – Sch?tze [11]) 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 кластерами 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 [10]. 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 [14]. 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.

## Реферат статьи «Word sense disambiguation with very large neural networks extracted from machine readable dictionaries »

Настоящий текст является рефератом статьи [9], в которой описан метод автоматического построения очень больших нейронных сетей (VLNN) с помощью текстов, извлекаемых из машинно-читаемых словарей (MRD), и рассмотрено использование этих сетей в задачах разрешения лексической неоднозначности (WSD).

В дальнейшем будем называть слова, смысл которых требуется установить целевыми словами.

Широко известен метод Леска [6] использования информации из MRD для задачи WSD. Суть этого метода состоит в вычислении так называемой «степени пересечения», т.е. количества общих слов в словарных определениях слов из контекста (“окна”) условного размера, содержащего целевое слово. Основной недостаток метода Леска – зависимость от словарной статьи, т.е. 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 (строения) VLNN.

Использование нейронных сетей для WSD было предложено в работах [5, 13]. В рассматриваемой статье для построения VLNN использован словарь Collins English Dictionary.

Топология сети. Целевое слово представлено узлом, соединенным активирующими связями со смысловыми узлами, представляющими все возможные смыслы слова, имеющиеся в словарных статьях. Каждый смысловой узел, в свою очередь, соединен активирующими связями с узлами, представляющими слова в словарной статье, соответствующей определению данного смысла. Процесс соединения повторяется многократно, создавая большую 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 смыслы данного слова, соединены запрещающими (подавляющими) связями.

Алгоритм функционирования сети. При запуске сети первыми активируются узлы входного слова, которое кодируется согласно принятому правилу. 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 смыслов получают обратные сигналы от узлов, соединенных с ними. Узлы конкурирующих смыслов 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-смыслов, одновременно уменьшая активацию узлов соответствующих неправильным смыслам. После нескольких десятков циклов сеть стабилизируется в состоянии, в котором активированы только узлы-смыслы с наиболее активированными связями с узлами-словами. В статье не указан алгоритм настройки, т.е. обучения сети. Видимо, используется метод встречного распространения (back propagation).

## Библиография

Библиографические ссылки принято оформлять в виде [номер], в отличие от ранее принятых [Автор, год] (см., напр., [1]). Источник, процитированный выше, был набран командой

```
\bibitem{Trans}
\textit{Борисов~Г.~А.,
Тихомирова~Т.~А.}
Характеристики и свойства потерь
энергии и мощности на пределах
энергетического хозяйства региона //
Труды Карельского научного центра
Российской академии наук. 2010.
\No 3. С.~4--10.
```

## Теоремоподобные окружения

Для теорем, утверждений и пр. необходимо использовать соответствующие окружения. Например:

**Утверждение 1.** *В предложенной модели системы обслуживания при $\rho = ES/ET < 1$ условие $ES^{\alpha+1} < \infty$ является достаточным для конечности момента порядка $\alpha$ времени ожидания в системе, $ED^{\alpha} < \infty$.*

*Доказательство.* Очевидно. □

В данном случае было использовано окружение `\begin{State}...\end{State}`. Для набора доказательства использовалось окружение `\begin{proof}...\end{proof}`. Доступные автору теоремоподобные окружения перечислены в Таблице 1.

*Таблица 1.* Теоремоподобные окружения

| | |
|---|---|
| `Theorem` | Теорема |
| `Lemma` | Лемма |
| `State` | Утверждение |
| `Corollary` | Следствие |
| `Axiom` | Аксиома |
| `Definition` | Определение |
| `Example` | Пример |
| `Remark` | Замечание |

Для определений, примеров и замечаний используется прямое написание.

**Пример.** Например, как в этом примере.

Соответствующие версии окружений «со звездой» также работают. Пример выше был набран такой командой:

```
\begin{Example*}
Например, как в этом примере.
\end{Example*}
```

*Таблица 2.* Таблица, демонстрирующая возможность размещения на всю ширину страницы

| Первая колонка | вторая колонка | третья колонка | четвертая колонка | пятая колонна |
|---|---|---|---|---|

## Рисунки и таблицы

Рисунки и таблицы могут вставляться как на всю ширину страницы, так и на ширину колонки. Желательно использовать рисунки формата pdf. Для конвертации из формата eps можно воспользоваться утилитой epstopdf. Так, например, Рис. 2 был вставлен на ширину колонки командой


*Рис. 2.* Задержки в модели 10-узловой системы

```
\begin{figure}[H]
\includegraphics[keepaspectratio=true,
 width=0.9\columnwidth]{delay_80.pdf}
\caption{Задержки в модели
 10-узловой системы}
\label{fig1}
\end{figure}
```

## Размещение на всю ширину страницы

В стилевом файле предусмотрена возможность размещения формул, рисунков и таблиц на ширину страницы. Для этого размещаемый элемент необходимо заключить между командами `\bfullwidth` и `\efullwidth`.

Пример формулы на всю ширину страницы:

$$Y = A_1x + A_2x^2 + \ldots + A_nx^n. \tag{1}$$

Набран пример следующим образом:

```
\bfullwidth
\begin{equation}
Y=A_1x+A_2x^2+\ldots +A_nx^n.
\end{equation}
\efullwidth
```

Пример размещения таблицы на ширину страницы (см. таблицу 2):

```
\bfullwidth
\centering
\begin{table}[H]
\begin{tabular}{|c|c|c|c|c|}
\hline
Первая колонка & ...\\
\hline
\end{tabular}
\label{tab_width}
\end{table}
\efullwidth
```

Рис. 1 демонстрирует возможности вставки по всей ширине страницы. Это было достигнуто при помощи команды

```
\bfullwidth
\begin{figure}
\includegraphics[height=100mm,...}
\caption{Задержки в модели...}
\label{fig5}
\end{figure}
\efullwidth
```

Следует обратить внимание на то, что вышеуказанная команда вставит рисунок не ближе,

чем на следующей странице сверху, а не сразу на месте указания команды.

### Сведения об авторах

После основного текста оформляются сведения об авторах. Используется окружение `\begin{aboutauthors}...\end{aboutauthors}`. При этом следует обратить внимание, что работа ведется в двухколоночном режиме, поэтому необходимо вручную указать разрыв колонки для отделения сведений на русском и английском языках. Например:

```
\begin{aboutauthors}
\authorsname{Румянцев Александр...}
аспирант\\
...
\columnbreak
\authorsname{Rumyantsev, Alexander}
...
\end{aboutauthors}
```

### Заключение

Компиляцию исходного файла желательно выполнять с помощью макроса `pdflatex`.

В работе рассмотрены основные технические аспекты подготовки статьи для сборника Трудов Карельского научного центра РАН. Предложения и пожелания по доработке стилевого файла, а также текста этого документа принимаются по электронному адресу, указанному в разделе «Сведения об авторах».

### Литература


1. *Борисов Г. А., Тихомирова Т. А.* Характеристики и свойства потерь энергии и мощности на пределах энергетического хозяйства региона // Труды Карельского научного центра Российской академии наук. 2010. №3. С. 4–10.

2. *Львовский С. М.* Набор и верстка в системе LaTeX. М., 2003. 448 с.

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

4. *Енрев, А. С.*Применение контекстных векторов в классификации текстовых документов. 2010. http://jre.cplire.ru/iso/oct10/1/text.html

5. *COTTRELL, G. W. and SMALL, S. L.*A connectionist scheme for modelling word sense disambiguation – Cognition and brain theory. 1983. № 6. P. 89–120.

6. *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.

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

8. *Savova, G.* Resolving ambiguities in biomedical text with unsupervised clustering approaches. University of Minnesota Supercomputing Institute Research Report, 2005.

9. *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.

10. *SenseClusters*http://senseclusters.sourceforge.net

11. *Schutze, H.*Automatic Word Sense Discrimination. Computational Linguistics, vol. 24, number 1., 1998.

12. *U*MLS Terminology Services (UTS). http://umlsks.nlm.nih.gov/kss/servlet/Turbine/template

13. *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.

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


## СВЕДЕНИЯ ОБ АВТОРЕ:


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

**Румянцев Александр Сергеевич**
аспирант
Институт прикладных математических исследований КарНЦ РАН
ул. Пушкинская, 11, Петрозаводск, Республика Карелия, Россия, 185910
эл. почта: ar0@krc.karelia.ru
тел.: (8142) 763370

**Kirilov, Alexander**
Doctor (DSc) of Physics and Mathematics
Assistant Professor
Institute of Applied Mathematical Research, Karelian Research Centre, Russian Academy of Science
11 Pushkinskaya St., 185910 Petrozavodsk, Karelia, Russia
e-mail: kirillov@krc.karelia.ru
tel.: (8142) 766312

**Rumyantsev, Alexander**
Institute of Applied Mathematical Research, Karelian Research Centre, Russian Academy of Science
11 Pushkinskaya St., 185910 Petrozavodsk, Karelia, Russia
e-mail: ar0@krc.karelia.ru
tel.: (8142) 763370

**Ткач Станислав Сергеевич**
Математический факультет
Петрозаводский государственный университет
пр-кт Ленина, 33, Петрозаводск, Республика Карелия
тел.: (8953) 547-70-43
tkachkras@gmail.com

**Tkach, Stanislav**
Student
Faculty of Mathematics
Petrozavodsk State University
Prospect Lenina, 33, Petrozavodsk, Republic of Karelia
tel.: (8953) 547-70-43
tkachkras@gmail.com

**Ярышкина Екатерина Александровна**
Студентка
Математический факультет
Петрозаводский государственный университет
пр-кт Ленина, 33, Петрозаводск, Республика Карелия
+7 (8142) 71-10-78
kate.rysh@gmail.com

**Yaryshkina, Ekaterina**
Student
Faculty of Mathematics
Petrozavodsk State University
Prospect Lenina, 33, Petrozavodsk, Republic of Karelia
+7 (8142) 71-10-78
kate.rysh@gmail.com