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, И. А. Сихонина

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, I. A. Sikhonina. 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 2_ε

```
\documentclass{article}
\usepackage{krctran}
...
\begin{document}
\procname{...}
\udk{...}
\rustitle{...}
\engtitle{...}
\rusauthor{...}
\engauthor{...}
\organization{...}
\rusabstract{...}
\engabstract{...}
\ruskeywords{...}
\engkeywords{...}

\maketitle

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

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

```
\begin{thebibliography}
...
\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`.

WSD-методы с учителем

«Разрешение лексической многозначности методом ансамбля Байесовских классификаторов»

А. Л. Чухарев, Т. В. Каушинис

В статье [15] 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, представленной как задача обучения с учителем, применяют статистические методы и методы машинного обучения к размеченному корпусу. В таких методах словам корпуса, для которых указано значение, соответствует набор лингвистических свойств. Алгоритм обучения строит модель классификатора значений по этим лингвистическим особенностям.

Автор статьи [15] предлагает подход, основанный на объединении ряда простых классификаторов в ансамбль, который разрешает многозначность с помощью голосования простым большинством голосов. Педерсен утверждает [15], что, во-первых, рассмотрение более сложного набора языковых особенностей или более сложных алгоритмов обычно не улучшает точность разрешения по сравнению с простыми языковыми особенностями, используемых в алгоритме обучения с учителем. Во-вторых, совместная встречаемость слов и словосочетаний имеет большее влияние на точность разрешения, оперирование такой лингвистической информацией как: часть речи или наличие отношений действие-объект.

В рассматриваемой статье [15] 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. Он 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 определённой 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 взаимозависи-

мыми 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 путём присвоения им по умолчанию очень маленького значения. После оценки всех параметров модель считается обученной и может быть использована в качестве классификатора.

Контекст в [15] 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 [15] 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 [15] 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 [15] 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 Байес (1,3) относится к диапазону категории (узкий, средний) поскольку он основан на окне из одного слова слева и окне из трёх 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 [12], [7]. В статье приводится информация о частоте использования шести значений для каждого из этих слов (Табл. 1, Табл. 2).

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

sense	count
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
total	4148

Таблица 2. Число употреблений слова *interest* (столбец *count*) для шести наиболее часто встречаемых значений (значения из словаря Longman Dictionary of Contemporary English, столбец *sense*). Этот набор данных был получен в 1994 году Брюсом и Виебе [7] путём указания значений для всех вхождений слова *interest* в корпус ACL/DCI Wall Street Journal

sense	count
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
total	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 [15] было получено, что ансамбль классификаторов с голосованием простым большинством даёт более высокую точность, чем взвешенное голосование. Например, для слова *interest* при голосовании простым большинством точность составила 89%, а взвешенное голосование дало только 83%.

«WSD на основе нейронных сетей, построенных по данным машино-читаемых словарей»

А. Н. Кирилов

Настоящий текст является рефератом статьи [20], 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 [13] 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 [9, 24]. 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).

WSD-методы без учителя

«Различение значений слов на основе векторов свойств, расширенных словарными толкованиями»

А. Спирикова

Амрута Пурандаре и Тед Педерсен в 2004 году разработали "Алгоритм различения значений на основе контекстных векторов"(Context vector sense discrimination) [17]. 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 [17].

Различение значений слов (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, которая говорит о том, что: лингвистические единицы, встречающиеся в схожих контекстах, имеют близкие значения. Следует различать понятия *различение значений слов* и *разрешение лексической многозначности*. При *различении*

значений слов нет никаких предопределённых 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 [17] предназначен для работы при недостаточном объёме текстовых данных, при этом вектор свойств расширяется данными, извлечёнными 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 (здесь тексты Wall Street Journal и Британского национального корпуса).

Вектор свойств (строка матрицы) содержит информацию о совместной встречаемости данного слова с другими. Было решено в [17], что слова «встречаются», если они находятся в тексте на расстоянии не более пяти словопозиций (то есть между ними находится не более трёх 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 [11], [10], [26]. 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 таблице 3 представлены примеры толкований и содержательные слова для восьми значений слова «история» из Русского Викисловаря.

Таблица 3. 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, чёрным – новые слова из толкований, которыми будет расширен вектор.

№	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	отдалённое 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* (таблица 3) 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 нём 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 трудностях и несовершенстве метода. 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.

«Разрешение многозначности в биомедицинских текстах с помощью методов кластеризации без учителя»

В статье [19] 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 [23] (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 [16], SC – Sch?tze [22]) 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 [21]. 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 [25]. 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.

WSD-методы, основанные на знаниях

«Построение сочетаемостных ограничений на основе Байесовских сетей для разрешения многозначности»

И. А. Сихонина

В статье [8] 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 “xxxx” – 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 [14]) – это обычно база данных, в которой слова сгруппированы в классы.

Статистический компонент состоит из пар предикат-аргумент, извлечённых 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 [8] семантическим классом называется синсет (группа синонимов) 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) [18], Абни и Лайт (1999) [5]) было обнаружено, что главная трудность в таком тривиальном алгоритме – это наличие неоднозначных слов в обучающих данных. В тех же работах ([18], [5]) 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. Любой переменной A с родителями $B1, \ldots, Bn$ соответствует таблица условных вероятностей (conditional probability table, далее CPT).

Например, построим SP для глагола *ползать* и сеть на рисунке 2 будет базой знаний.

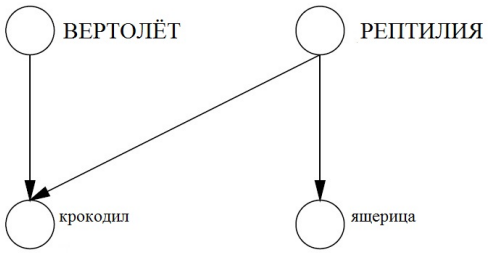

Рис. 2. Байесовская сеть для многозначного существительного *крокодил*

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 4. 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 таблицы 4 условных вероятностей (CPT), учтём 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 таблицы 4 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 [8] 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.

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 α 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 5.

Таблица 5. 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 6. 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. 3 был вставлен на ширину колонки командой

Рис. 3. 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 6):

```
\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. *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.

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

7. *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, pages 139–146.

8. *Ciaramita, M. and Johnson, M.*Explaining away ambiguity: Learning verb selectional preference with Bayesian networks. 2000.

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

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

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

12. *C. Leacock, G. Towell, and E. Voorhees.*Corpus-based statistical sense resolution. In Proceedings of the ARPA Workshop on Human Language Technology, pages 260–265, March. 1993.

13. *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.

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

15. *Pedersen, T.*A Simple Approach to Building Ensembles of Naive Bayesian Classi?ers for Word Sense Disambiguation// Department of Computer Science, University of Minnesota Duluth.–2000.

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

17. *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. – С. 123-130.

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

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

20. *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.

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

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

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

24. *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.

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

26. *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, pages 515-524, McLean, VA, 2002.

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

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

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

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

Сихонина Ирина Александровна
Математический факультет
Петрозаводский государственный университет
пр-кт Ленина, 33, Петрозаводск, Республика Карелия
тел.: +7(8142) 71-10-78
syawenka@mail.ru

Sikhonina, Irina
Student
Faculty of Mathematics
Petrozavodsk State University
Prospect Lenina, 33, Petrozavodsk, Republic of Karelia
tel.: +7(8142) 71-10-78
syawenka@mail.ru

Ткач Станислав Сергеевич
Математический факультет
Петрозаводский государственный университет
пр-кт Ленина, 33, Петрозаводск, Республика Карелия
тел.: +7 (8142) 71-10-78
tkachkras@gmail.com

Tkach, Stanislav
Student
Faculty of Mathematics
Petrozavodsk State University
Prospect Lenina, 33, Petrozavodsk, Republic of Karelia
tel.: +7 (8142) 71-10-78
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