*Моделирование и анализ информационных систем.* Т. 23, №6 (2016), с. 826–840
*Modeling and Analysis of Information Systems.* Vol. 23, No 6 (2016), pp. 826–840



**DOI:** 10.18255/1818-1015-2016-6-826-840

УДК 004.912

# Методические аспекты выделения семантических отношений для автоматической генерации специализированных тезаурусов и их оценки

**Лагутина Н. С., Лагутина К. В., Мамедов Э. И., Парамонов И. В.**[1]

*получена 19 октября 2016*

**Аннотация.** Работа посвящена анализу методов автоматической генерации специализированного тезауруса. Основной алгоритм генерации состоит из трех шагов: отбор и предварительная обработка корпуса текстов, формирование множества терминов для включения в тезаурус и выделение связей между терминами тезауруса. Данное исследование сфокусировано на изучении методов выделения семантических связей, для чего авторами был разработан программный стенд, который позволяет протестировать распространенные алгоритмы выделения гиперонимов и синонимов, использующие в своей работе лексико-синтаксические шаблоны, морфо-синтаксические правила, количество информации терминов, тезаурус общего назначения WordNet и расстояние Левенштейна. Для анализа результирующего тезауруса, созданного на стенде, авторами была разработана комплексная оценка, содержащая следующие характеристики качества: точность выделения терминов, точность и полнота выделения синонимических и гиперонимических связей, а также метрики графа тезауруса (количество выделенных терминов, количество семантических связей различных типов, число компонент связности и число вершин в наибольшей компоненте). Предлагаемый набор метрик позволяет оценить качество тезауруса в целом, выявить отдельные недостатки стандартных методов выделения связей и построить более эффективные гибридные методы, генерирующие тезаурус с лучшими характеристиками по сравнению с тезаурусами, генерируемыми при использовании отдельных методов. Для иллюстрации данного факта в статье рассмотрен один из таких гибридных методов. Он комбинирует лучшие стандартные алгоритмы построения гиперонимических и синонимических связей и строит специализированный тезаурус в области медицины с тем же уровнем качества, что и другие методы, но с большим количеством связей между терминами.

**Ключевые слова:** тезаурус, семантические отношения, гибридный метод, комплексная оценка, программный стенд

**Для цитирования:** Лагутина Н. С., Лагутина К. В., Мамедов Э. И., Парамонов И. В., "Методические аспекты выделения семантических отношений для автоматической генерации специализированных тезаурусов и их оценки", *Моделирование и анализ информационных систем*, **23**:6 (2016), 826–840.

**Об авторах:**
Лагутина Надежда Станиславовна, orcid.org/0000-0002-6137-8643, канд. физ.-мат. наук, доцент,
Ярославский государственный университет им. П.Г. Демидова,
ул. Советская, 14, г. Ярославль, 150000 Россия, e-mail: lagutinans@rambler.ru

Лагутина Ксения Владимировна, orcid.org/0000-0002-1742-3240, студент,
Ярославский государственный университет им. П.Г. Демидова,
ул. Советская, 14, г. Ярославль, 150000 Россия, e-mail: lagutinakv@mail.ru

Мамедов Эльдар Интизамович, orcid.org/0000-0003-1997-7084, стажёр-исследователь,
Ярославский государственный университет им. П.Г. Демидова,
ул. Советская, 14, г. Ярославль, 150000 Россия, e-mail: eldarmamedov90@gmail.com

Парамонов Илья Вячеславович, orcid.org/0000-0003-3984-8423, канд. физ.-мат. наук, доцент,
Ярославский государственный университет им. П.Г. Демидова,
ул. Советская, 14, г. Ярославль, 150000 Россия, e-mail: Ilya.Paramonov@fruct.org

**Благодарности:**
[1]Работа выполнена при финансовой поддержке гранта Президента Российской Федерации для государственной поддержки молодых российских ученых (государственный контракт № МК-5456.2016.9).

# Введение

Тезаурус можно определить как словарь терминов на естественном языке, в котором все элементы связаны между собой семантическими отношениями, отражающими основные соотношения понятий в описываемой предметной области знаний [1].

Тезаурусы широко применяются в ряде областей. Одной из таких областей является информационный поиск, включающий в себя поиск по запросам пользователей и индексацию документов. Другая область — анализ текстов, в частности рубрикация и классификация документов, моделирование связности текста, автоматический перевод.

Большое практическое значение имеет информационный поиск в специализированных областях, где он существенно отличается от поиска в коллекциях документов из интернета, так как требует лингвистических и онтологических знаний о предметной области. Одним из ресурсов, содержащих такие знания, является специализированный тезаурус [2].

В настоящий момент существует не слишком много информационно-поисковых тезаурусов предметных областей, поскольку процесс построения специализированных тезаурусов является чрезвычайно трудоемким и финансово затратным, кроме того, результат сильно зависит от квалификации экспертов. Автоматизировать процесс построения такого тезауруса также достаточно сложно. Весь процесс целиком и его отдельные части являются предметом многочисленных исследований.

В то же время отмечается ряд проблем, которые мешают эффективному использованию тезаурусов в системах компьютерной обработки текста, в частности нехватка отношений между терминами предметной области [3]. Таким образом, в процессе автоматического построения информационно-поисковых тезаурусов важно добиться качественного моделирования рассматриваемой области, при этом существенными характеристиками результата являются параметры связей между отдельными элементами.

Задачей статьи является анализ и обобщение результатов исследований авторов, представленных в предыдущих работах, посвященных отдельным аспектам алгоритмов автоматического построения тезаурусов. В данной работе значительное внимание уделяется методам выделения связей между терминами. Наличие разных типов связей позволяет максимально эффективно отразить знания о соответствующей предметной области и улучшить в дальнейшем качество применения тезауруса в информационных системах. В связи с этим большое значение приобретают характеристики связности между терминами тезауруса, поэтому авторы рассматривают методические аспекты выделения связей между терминами, основанные на применении гибридных методов. Кроме того, авторы рассматривают вопросы комплексной оценки автоматически сгенерированного тезауруса. Такая оценка может послужить основой для построения полностью автоматических алгоритмов создания тезауруса.

# 1. Общая схема построения тезауруса

Формальное определение тезауруса может выглядеть следующим образом:

$$T =< D, R_s, R_h, R_a >$$

Здесь $D$ — множество всех терминов тезауруса. $R_s \subset D \times D$ — отношение синонимии, обладающее свойствами симметричности и транзитивности. В каждом множестве синонимов $d_i : \forall i, j = \ldots n(d_i, d_j) \in R_s$ обычно выделяется один дескриптор, соответствующий понятию предметной области, остальные синонимы называются аскрипторами. $R_h$ — отношение иерархии. Иерархическими являются родо-видовые (гипонимо-гиперонимические) связи, а также связи типа часть-целое. Иерархические отношения обладают свойствами несимметричности и транзитивности. $R_a$ — ассоциативное отношение. Чаще всего ассоциативными отношениями называются отношения между дескрипторами предметной области, не являющиеся иерархическими или синонимическими.

Метод автоматизированного построения тезауруса состоит из нескольких этапов:

1. Отбор и предварительная обработка корпуса текстов.
2. Формирование множества терминов для включения в тезаурус.
3. Выделение связей между терминами тезауруса.

Отбор и предварительная обработка корпуса текстов не является предметом обсуждения в данной работе. Корпус представляет собой сформированную по определенным правилам выборку текстов, относящихся к предметной области.

Задача выделения терминов для тезауруса обычно рассматривается как задача выделения ключевых фраз из корпуса текстов [4]. Поэтому для исследования были выбраны соответствующие алгоритмы, описанные в разделе 2.

Для выделения связей между терминами существуют разные типы алгоритмов, зависящие от типов связей и способов их нахождения. Подробное описание исследуемых в данной работе методов построения связей в тезаурусе находится в разделе 3.

Для оценки качества работы методов выделения терминов или отношений между ними практически во всех научных работах используются три статистические характеристики: точность, полнота и F-мера [5]. Точность — это число правильно выделенных ключевых слов, поделенное на число всех выделенных ключевых слов. Полнота — это число правильно выделенных ключевых слов, поделенное на число всех подходящих ключевых слов. F-мера — это среднее гармоническое точности и полноты, которая позволяет оценить баланс между точностью и полнотой, придавая одинаковый вес обеим метрикам. В данной работе авторы вычисляли точность выделенных терминов, иерархических и синонимических отношений; полноту иерархических и синонимических отношений.

Качество тезауруса также зависит от набора его связей: чем больше семантических отношений в тезаурусе и чем выше его связность, тем лучше результаты дает информационный поиск с его участием [3]. Для оценки данной характеристики введем граф тезауруса $G = (D, R)$, где множество вершин $D$ — это множество всех терминов тезауруса, определенное ранее, а множество ребер $R = R_s \cup R_h \cup R_a$ — это множество всех связей между терминами.

Авторами были выбраны следующие характеристики графа тезауруса, позволяющие оценить набор его связей:

- количество выделенных терминов $|D|$;
- количество отношений различных типов $|R_s|, |R_h|, |R_a|$;
- число компонент связности и вершин в наибольшей компоненте, т. е. метрики, описывающие связность графа тезауруса.

Число выделенных терминов и отношений между ними — самые грубые метрики, по которым можно оценить размер тезауруса. Общее число терминов не должно быть маленьким, число вертикальных и горизонтальных связей должно быть больше числа терминов, так как практически все термины имеют хотя бы один гипероним и несколько синонимов или ассоциаций. Выход за рамки данных правил означает низкое качество результирующего тезауруса.

Оценка связности тезауруса выполняется по примеру предыдущей работы авторов [6] и обосновывается тем фактом, что для успешного применения тезауруса в большинстве случаев решающее значение имеет навигация по связям. В частности, авторы предлагают использовать характеристики графа, определенного выше. В том случае, если тезаурус содержит только небольшие компоненты связности, то он не отражает структуру предметной области. Также граф тезауруса не должен содержать изолированных вершин, так как соответствующие им термины не имеют связей и не могут быть использованы на практике.

Полученный набор характеристик является комплексной оценкой качества полученного узкоспециализированного тезауруса. Подобные оценки редко появляются в современных исследованиях, так как чаще всего оцениваются отдельные этапы или методы генерации тезауруса. Нередко предлагаемые методы улучшают одну характеристику тезауруса, но ухудшают другую, например, увеличение полноты сопровождается уменьшением точности [7]. Комплексная оценка тезауруса как единого целого позволяет построить более эффективные гибридные методы, дающие возможность найти компромиссное решение.

## 2. Методы выделения терминов тезауруса

Для решения задачи автоматического выделения терминов тезауруса из текстов, иначе говоря — ключевых фраз, разработано большое количество подходов. Все существующие алгоритмы состоят из двух шагов: выделение ключевых слов-кандидатов эвристическими методами и выбор релевантных ключевых слов из набора слов-кандидатов при помощи математических методов видов, относящихся к одной из двух категорий: «обучение с учителем» (обучаемые) и «обучение без учителя» (необучаемые).

Для проведения настоящего исследования из этих двух категорий были выбраны два метода: Topical PageRank из категории «обучение без учителя» и Maui — «обучение с учителем».

В работе [6] авторы данной статьи исследуют выбранные стандартные методы для выделения ключевых слов на корпусе текстов о туристических объектах Карелии. Данный корпус из 900 текстов удобен для анализа методов решения задачи построения узкоспециализированного тезауруса в частности тем, что содержит выборку из 200 текстов с выделенными вручную ключевыми словами. Это позволяет

обрабатывать его при помощи алгоритмов как вида «обучение без учителя», так и «обучение с учителем».

В экспериментах оценивалось качество выделенных ключевых слов для каждого алгоритма из четырех стандартных. Для данных экспериментов использовались 200 текстов с выбранными экспертом ключевыми словами, из которых 100 текстов применялись для обучения алгоритмов, а 100 других текстов — для оценки качества. Все алгоритмы запускались на этом наборе, и их результаты сравнивались с ключевыми словами, выделенными экспертом. Релевантность оценивалась по трем метрикам: точности, полноте и F-мере.

Полученные значения характеристик показывают, что алгоритм Maui работает эффективнее, чем остальные алгоритмы: 57.5 % выбранных слов хорошо характеризуют текст, 24.1 % из всех возможных правильных ключевых слов были выбраны, а значение F-меры 34.0 — наибольшее среди всех. Вторым по эффективности является необучаемый алгоритм Topical PageRank: точность — 51.7 %, полнота — 22.0 %.

Несмотря на то, что значения характеристик достаточно хороши по сравнению с результатами других исследований [5], очевидно, что они являются весьма низкими по абсолютным значениям. Можно сделать вывод, что для повышения качества выделения ключевых слов в конкретной предметной области требуется существенная доработка стандартных алгоритмов. Авторами был предложен алгоритм выделения ключевых слов, учитывающий особенности предметной области туризма, а также особенности информационной системы, для которой он разрабатывался [6]. Аналогичный подход может быть использован для повышения качества выделения терминов специализированного тезауруса.

В данной работе авторы не исследуют алгоритмы для выделения терминов, поэтому для выделения ключевых слов во всех экспериментах по выделению семантических отношений был выбран единственный метод TextRank [8] вида «обучение без учителя». Авторы не использовали в данном исследовании обучаемые алгоритмы, так как алгоритмы без учителя не нуждаются в текстах с выбранными вручную ключевыми словами и поэтому представляют разумный компромисс для поставленной задачи максимально автоматизировать построение тезауруса, поскольку показывают результаты лишь немного хуже, чем обучаемые алгоритмы. Хотя одним из лучших необучаемых алгоритмов считается Topical PageRank [9], он может быть успешно заменен на TextRank, так как Topical PageRank выбирает из текстов темы и повторяет для каждой TextRank. В данном исследовании все тексты объединены общей темой, и TextRank достаточно хорошо подходит для поставленной задачи.

## 3. Методы выделения связей между терминами тезауруса

Последним и крайне важным этапом генерации тезауруса является построение отношений между терминами. В данном разделе авторы анализируют наиболее известные и эффективные подходы к выделению различных типов связей, которые затем используются в реализации программного стенда, описанного в подразделе 4.1. Также данные методы комбинируются для создания гибридных методов (подраздел 4.3.).

Методы для определения семантических связей между терминами можно разделить на несколько групп в зависимости от типа выделяемых связей:

- методы определения ассоциативных связей;
- методы определения гипонимо-гиперонимических связей;
- методы определения синонимических связей;
- методы определения нескольких типов связей.

Рассмотрим данные методы подробно.

## 3.1. Методы определения ассоциативных связей

Одним из эффективных методов определения ассоциативных связей является метод LSA (Latent Semantic Analysis) [10]. Его основное предназначение заключается в нахождении взаимосвязей между корпусом документов и терминами, в них встречающимися. Также данный метод можно использовать и для нахождения степени связи между терминами корпуса документов.

Метод LSA работает следующим образом: сначала строится матрица термины-на-документы, в которой каждой строке соответствует термин из корпуса документов, каждому столбцу — документ. На пересечении строк и столбцов матрицы записывается число, указывающее, сколько раз соответствующий термин встречается в соответствующем документе. Каждую строку в получившейся матрице можно рассматривать как вектор, характеризующий определенный термин в корпусе и содержащий информацию о встречаемости термина в документах корпуса. После того как матрица построена, для неё вычисляется сингулярное разложение, чтобы получить наилучшее приближение матрицей меньшей размерности. В получившейся матрице каждая строка также характеризует определенный термин корпуса. Для того, чтобы определить степень связи между различными терминами, к соответствующим векторам терминов применяется какая-либо известная мера близости векторов, например, косинусная мера. Чем ближе по выбранной мере оказываются векторы, тем более связанными считаются соответствующие им термины.

## 3.2. Методы определения гипонимо-гиперонимических связей

Для определения гипонимо-гиперонимических связей часто используются лингвистические методы. Наиболее известными из них являются методы, основанные на морфо-синтаксических правилах и лексико-синтаксических шаблонах.

Метод, основанный на морфо-синтаксических правилах [11], определяет гипонимо-гиперонимические отношения в тех случаях, когда один термин является частью второго. Отношения определяются по следующим правилам:

- если первый термин является строкой-суффиксом для второго термина, то первый является гиперонимом, а второй — гипонимом;
- если первый термин является частью фразы второго многословного термина, то первый является гиперонимом, а второй — гипонимом.

Другой лингвистический метод использует известные лексико-синтаксические шаблоны, в которых часто встречаются термины, состоящие в иерархическом отношении [12]. Каждое предложение в исходном тексте проверяется на наличие опреде-

ленного списка известных лексико-синтаксических шаблонов. В случае нахождения одного из шаблонов термины, в зависимости от их положения в шаблоне, определяются как гипонимы и гиперонимы по отношению друг к другу.

## 3.3. Методы определения синонимических связей

Одним из вариантов определения синонимичных терминов является поиск различных форм одного и того же слова. В узкоспециализированных тезаурусах формы слова и синонимы часто обозначаются как один и тот же тип связей (пример — медицинский тезаурус MeSH: https://www.nlm.nih.gov/mesh/). Методы нахождения форм слова можно разделить на несколько типов в зависимости от способа их определения: орфографические, морфологические, лексические, структурные, аббревиатурные [4].

Известным морфо-орфографическим методом определения различных вариантов слова является метод, основанный на расстоянии Левенштейна [13]. Он определяет количество изменений, которые требуются для того, чтобы преобразовать одно слово в другое. Учитываются такие изменения, как удаление символа, добавление символа и замена символа. Затем с учетом длины исходного термина и найденного количества требуемых изменений вычисляется коэффициент схожести терминов. Чем больше данный коэффициент, тем более схожими считаются термины, таким образом расстояние Левенштейна может использоваться для поиска однокоренных слов и лексических вариантов терминов тезауруса.

## 3.4. Методы определения нескольких типов связей

В данном разделе выделены методы, способные определять сразу несколько различных типов связей, в частности синонимические и гипонимо-гиперонимические связи. К таким методам относятся метод, основанный на оценке количества информации терминов, и метод, основанный на использовании тезауруса общего назначения WordNet.

Метод, основанный на оценке количества информации терминов [14], — это статистический метод, определяющий тип связи между терминами, опираясь на такие показатели, как количество информации терминов и схожесть контекстов, в которых встречаются термины. Во-первых, метод вычисляет количество информации для терминов по следующему правилу: чем чаще встречается термин, тем меньше он несет в себе информации. Частота встречаемости слов определяется в некотором большом общем корпусе текстов. Во-вторых, метод определяет термины, которые часто употребляются в схожих контекстах. Далее для каждой пары терминов сравниваются их контекст и количество информации. Если контекст у терминов один и тот же, и количество информации близко по значению, то термины определяются как синонимы. Если контекст одинаковый, а сами термины несут различное количество информации, то они формируют пару гипоним-гипероним. Причем гипонимом в данном случае считается термин, который несет в себе больше информации, а гиперонимом — термин с меньшим количеством информации.

Метод, основанный на использовании WordNet [15], устанавливает связь между терминами создаваемого специализированного тезауруса, если связь такого же типа

была найдена между данными терминами в тезаурусе общего назначения WordNet. Синонимическая связь между терминами устанавливается в том случае, когда данные термины определены как синонимы в WordNet; гипонимо-гиперонимическая связь устанавливается, если между данными терминами в WordNet также определена соответствующая связь.

# 4. Программный стенд для построения и оценки тезауруса

## 4.1. Общий алгоритм работы стенда

Для построения и оценки методов автоматической генерации тезаурусов авторы данной работы разработали программный стенд. Он позволяет автоматически сгенерировать узкоспециализированный тезаурус из неструктурированного корпуса текстов с использованием различных методов из предыдущего раздела или их комбинаций. Также стенд оценивает качество терминов и отношений между ними в результирующем тезаурусе. Общий алгоритм стенда включает в себя следующие шаги:

1. Выбор терминов для тезауруса с использованием алгоритма TextRank.
2. Выделение ассоциативных связей алгоритмом LSA.
3. Выделение гиперонимических связей при помощи статистических и семантических методов, а также их комбинаций.
4. Выделение синонимических связей различными статистическими и семантическими алгоритмами и их комбинациями.
5. Фильтрация терминов, которые не имеют связей.
6. Оценка качества терминов и связей в тезаурусе.

Для выделения иерархических связей исследуются четыре метода, описанные в разделе 2: морфо-синтаксические правила; лексико-синтаксические шаблоны; метод, использующий тезаурус общего назначения WordNet; метод на основе измерения количества информации, которое несёт термин. Два последних метода также используются и для выделения синонимов, наряду с методом, использующим расстояние Левенштейна. В разных экспериментах используются разные методы и их комбинации.

После выделения связей фильтруются термины, которые остались без связей, так как они бесполезны в подавляющем большинстве сценариев использования тезаурусов.

Результатом работы алгоритма стенда является узкоспециализированный тезаурус, содержащий термины и синонимические и гипонимо-гиперонимические связи между терминами, описывающий структуру предметной области, заданной конкретным корпусом текстов. Оценка тезауруса выполняется с использованием метрик, описанных в разделе 1.

## 4.2. Эксперименты на стенде: стандартные методы

Для экспериментов по автоматическому построению тезаурусов использовалась хорошо известная коллекция медицинских текстов MEDLINE (http://ir.dcs.gla.ac.uk/resources/testcollections/medl/). Коллекция MEDLINE содержит 1033 статьи из медицинских журналов и предназначена для оценки методов информационного поиска. В наших экспериментах данная коллекция использовалась как корпус текстов, из которого выделяются термины и связи между ними для автоматического построения тезаурусов в области медицины.

При использовании метода, основанного на оценке количества информации [14], для определения частоты встречаемости слов дополнительно использовался медицинский словарь Стедмана (http://stedmansonline.com/public/LearnMore.aspx?resourceID=Medical). Для определения терминов, встречающихся в схожих контекстах, использовались термины, связанные отношением ассоциации, найденные методом LSA на соответствующем шаге работы стенда.

Для оценки качества автоматически построенных тезаурусов выделенные термины и связи сравнивались с терминами и связями эталонного тезауруса в указанной предметной области. В качестве эталонного тезауруса использовался биомедицинский тезаурус MeSH (https://www.nlm.nih.gov/mesh/). Для оценки качества, в частности, оценивалась полнота и точность выделенных терминов и связей.

Большой интерес в экспериментах по автоматическому построению тезаурусов вызывает качество различных методов для выделения гипонимо-гиперонимических и синонимических связей, и то, насколько их качество отличается друг от друга. Поэтому в первую очередь мы провели эксперименты с тезаурусами, в которых для определения связей использовался какой-либо один из четырёх перечисленных выше стандартных методов. Для определения синонимов во всех указанных экспериментах использовался метод, основанный на расстоянии Левенштейна. Кроме того, в экспериментах с методом, основанном на использовании тезауруса WordNet, и методом, использующим количество информации терминов, эти же самые методы использовались и для выделения синонимов.

В результате проведенных экспериментов на стенде были вычислены некоторые статистические характеристики построенных тезаурусов, позволяющие оценить их общую структуру (см. первые четыре строки в таблице 1). Как видно из результатов, наибольшее количество выделенных терминов, наибольшее количество найденных гипонимо-гиперонимических связей и наибольшая связность графа относится к методу, основанному на использовании WordNet. Наибольшее количество синонимов было найдено с помощью метода, основанного на количестве информации терминов.

Далее мы оценили качество построенных тезаурусов в сравнении с эталонным тезаурусом MeSH. В частности, были вычислены точность и полнота выделения терминов и отношений. Результаты записаны в таблице 2.

Точность выделения терминов примерно одинаковая для всех методов и составляет около 38.5 %. Метод, использующий расстояние Левенштейна для выделения синонимов, показал высокую точность на уровне 76–89 %, но при этом он даёт довольно низкую полноту на уровне 11–12 %. Использование WordNet для нахождения синонимов также дало неплохие результаты. Несмотря на то, что точность уменьшилась до 57.4 %, этот показатель остается на достаточно хорошем уровне. Полно-

Таблица 1: Характеристики автоматически построенных тезаурусов

Table 1: Characteristics of automatically generated thesauri

| Методы выделения | | Количество выделенных | | | | Количество | |
|---|---|---|---|---|---|---|---|
| гиперонимов | синонимов | терминов | гиперонимов | синонимов | ассоциаций | компон. связности | вершин в наиб. комп. |
| Extraction methods for | | Quantity of extracted | | | | Quantity of | |
| hypernyms | synonyms | terms | hypernyms | synonyms | associations | connected comp. | vertices in max. comp. |
| Лекс | Лев | 2167 | 102 | 63 | 4918 | 86 | 1936 |
| Морф | Лев | 2090 | 248 | 46 | 4731 | 76 | 1909 |
| КолИнф | КолИнф, Лев | 2105 | 346 | 753 | 3883 | 84 | 1881 |
| WordNet | WordNet, Лев | 2406 | 2107 | 290 | 4375 | 35 | 2318 |
| Гибрид | WordNet, Лев | 2411 | 1570 | 312 | 4842 | 36 | 2323 |

Лекс — лексико-синтаксические шаблоны.
Морф — морфо-синтаксические правила.
КолИнф — метод, основанный на оценке количества информации.
WordNet — метод, основанный на использовании WordNet.
Лев — метод, использующий расстояние Левенштейна.
Гибрид — совместное применение методов Морф, КолИнф и WordNet.

та выделения синонимов увеличилась до 13.9 %. Метод, основанный на количестве информации, показал весьма слабые результаты по нахождению синонимов — точность и полнота не превышают уровня 15.5 %.

Качество выделения гипонимо-гиперонимических связей для всех методов оказалось довольно невысоким. Самым точным оказался метод, основанный на морфо-синтаксических правилах. Его точность составляет 23.1 %, что в несколько раз больше, чем точность всех остальных методов. Наибольшую полноту (23.8 %) показал метод, основанный на использовании WordNet, в то время как остальные методы показали значительно меньшее значение этого показателя. Примечательно, что в наших экспериментах метод, основанный на лексико-синтаксических шаблонах, показал наихудшие результаты и не нашел ни одной гипонимо-гиперонимической связи из тезауруса MeSH.

По результатам экспериментов со стандартными методами можно сделать несколько выводов. Метод, использующий WordNet, находит много правильных отношений, но его главный недостаток заключается в том, что тезаурус общего назначения не содержит в себе всех терминов из специализированной предметной области, и многие взаимоотношения не могут быть найдены таким способом. Методы, использующие морфо-синтаксические правила и лексико-синтаксические шаблоны, позволяют найти отношения, явно определенные в тексте, но их использование становится очень ограниченным в неструктурированных тестах и не позволяет найти множество других взамосвязанных терминов. Статистический метод находит большое количество различных отношений, но в то же время этот метод гораздо чаще, чем остальные, находит неверные отношения.

Таблица 2: Точность и полнота выделения терминов и семантических отношений при автоматическом построении тезауруса (в сравнении с тезаурусом MeSH)

Table 2: Precision and recall of extraction of terms and relations for case of automatic thesaurus generation (compared to the MeSH thesaurus)

| Методы выдел. гиперонимов | Методы выдел. синонимов | Все термины | Синонимы | | Гиперонимы | |
|---|---|---|---|---|---|---|
| | | Точность | Точность | Полнота | Точность | Полнота |
| Extr. meth. for hypernyms | Extr. meth. for synonyms | All terms Precision | Synonyms Precision | Recall | Hypernyms Precision | Recall |
| Лекс | Лев | 39.0 | 75.8 | 11.2 | 0.0 | 0.0 |
| Морф | Лев | 38.2 | 88.5 | 11.7 | 23.1 | 13.0 |
| КолИнф | КолИнф, Лев | 38.5 | 12.8 | 15.2 | 7.1 | 8.3 |
| WordNet | WordNet, Лев | 38.4 | 57.4 | 13.9 | 6.9 | 23.8 |
| Гибрид | WordNet, Лев | 38.4 | 57.4 | 15.7 | 8.3 | 17.9 |

Лекс — лексико-синтаксические шаблоны.
Морф — морфо-синтаксические правила.
КолИнф — метод, основанный на оценке количества информации.
WordNet — метод, основанный на использовании WordNet.
Лев — метод, использующий расстояние Левенштейна.
Гибрид — совместное применение методов Морф, КолИнф и WordNet.

## 4.3. Эксперименты на стенде: гибридный метод

Результаты экспериментов подтвердили, что стандартные методы хороши при выделении определенных типов отношений, а не для всех отношений сразу, что привело нас к идее гибридного метода, заключающегося в использовании комбинации стандартных методов выделения отношений.

В данном исследовании мы реализовали гибридный метод, объединяющий в себе большинство стандартных методов. По результатам экспериментов, практически все алгоритмы достаточно полезны при выделении различных типов связей за исключением лишь нескольких. В частности, мы выяснили, что метод, основанный на лексико-синтаксических шаблонах, показал отрицательные результаты в нахождении гипонимо-гиперонимических связей и не выделил ни одной правильной связи. Также метод, основанный на количестве информации, привел к значительному ухудшению точности в задаче выделения синонимов. Данные обстоятельства привели нас к идее того, чтобы не включать указанные шаги в гибридный метод. В результате работа гибридного метода была построена по следующему алгоритму:

1. Выделение терминов тезауруса.
2. Определение ассоциативных связей методом LSA.
3. Определение гипонимо-гиперонимических связей с помощью метода, основанного на использовании WordNet.
4. Определение гипонимо-гиперонимических связей с помощью метода, основанного на морфо-синтаксических правилах.
5. Определение синонимических связей с помощью метода, основанного на расстоянии Левенштейна.
6. Определение синонимических связей с помощью метода, основанного на использовании WordNet.

7. Определение гипонимо-гиперонимических связей с помощью метода, основанного на количестве информации терминов.
8. Фильтрация терминов, которые не имеют связей.

Следует отметить, что если для пары терминов на некотором этапе была выделена связь, то на следующих шагах алгоритма возможность выделения связи другого типа для этой же пары исключается.

Для построенного гибридного метода мы провели точно такие же эксперименты, как и для стандартных методов. Результаты экспериментов показаны в последней строке таблиц 1 и 2.

Из результатов данных экспериментов видно, что по сравнению со всеми другими методами гибридный метод выделяет наибольшее количество терминов и отношений между ними. Кроме того, тезаурус, построенный с помощью данного метода, имеет высокую связность.

В сравнении с тезаурусом MeSH было замечено, что гибридный метод находит больше совпавших терминов и отношений, чем стандартные методы, но количество несовпавших терминов и отношений также возросло. Точность выделения гипонимо-гиперонимических связей упала до 8.3 %, но полнота в среднем стала выше и оказалась равной 17.9 %. Точность выделения синонимов оказалась на одном уровне с точностью метода, основанного на использовании WordNet, и равняется 57.4 %. Полнота выделения синонимов возросла до 15.7 %, что является лучшим результатом среди всех тестируемых методов. Точность выделения терминов осталась такой же, как и прежде.

## Заключение

В данной работе авторы проанализировали несколько методов выделения семантических отношений между терминами тезауруса, которые могут быть использованы в алгоритме полностью автоматической генерации узкоспециализированного тезауруса. Авторами был предложен стенд для подобной генерации и оценки тезауруса, использующий лучшие существующие статистические и семантические алгоритмы выделения терминов и отношений между ними.

На основе анализа результатов экспериментов можно выделить некоторые методические аспекты автоматической генерации тезауруса. Для полностью автоматического выделения терминов тезауруса лучше всего использовать алгоритмы выделения ключевых слов вида «обучение без учителя», так как они работают достаточно эффективно и при этом не требуют работы эксперта. В выделении гиперонимов достаточно хороши два метода, основанные на морфо-синтаксических правилах и тезаурусе WordNet. Первый работает с лучшей точностью, но выделяет мало отношений. Второй работает с лучшей полнотой, но не может выделять связи между специфическими терминами, которые отсутствуют в тезаурусах общего назначения. В выделении синонимов наилучшие результаты показывает алгоритм, считающий расстояние Левенштейна, однако он находит небольшое число отношений и не может выделять связи между неоднокоренными синонимами.

Для более эффективного анализа результатов авторами была предложена комплексная оценка итогового тезауруса, включающая в себя, кроме широко используе-

мых статистических мер точности и полноты, сравнивающих полученный тезаурус с эталонным, метрики связности графа тезауруса. По результатам всех экспериментов лучшую точность выделения иерархических связей показал морфо-синтаксический метод, однако он выделяет существенно меньше связей, чем статистический метод или метод, основанный на использовании WordNet. Лучшей точностью выделения синонимических связей обладает метод, использующий расстояние Левенштейна. Также в задаче выделения синонимов хорошие результаты показал метод, использующий WordNet. Отсюда можно сделать вывод, что существующие методы эффективны в извлечении определенных типов отношений, но недостаточно хороши для построения тезауруса в целом.

Для решения данной проблемы авторами была предложена идея гибридных методов, представляющих собой различные комбинации существующих алгоритмов выделения синонимических и иерархических отношений между терминами. Эксперименты показали: преимущество гибридных методов при выделении обоих типов связей заключается в том, что у них лучшая полнота, т. е. они находят больше правильных отношений, а также данные методы выделяют больше терминов и строят большее количество связей, тем самым повышая связность итогового тезауруса. Таким образом, если для поставленной задачи важна точность выделенных связей, то лучше использовать отдельные стандартные методы. Для обеспечения высокой связности тезауруса и хорошей полноты выделения связей более пригодными являются гибридные методы.

Хотя обнаруженный эффект повышения качества построения тезауруса может зависеть от используемого корпуса текстов или специфических особенностей предметной области, идея гибридных методов выглядит многообещающей в перспективе автоматической генерации узкоспециализированных тезаурусов, что является актуальной задачей в связи с необходимостью минимизировать работу эксперта над тезаурусом. Дальнейшие направления исследований в этой области могут состоять в изучении особенностей гибридных методов, создании подходов для автоматической оценки конкретных методов на конкретных корпусах документов и последующем синтезе гибридных методов для обеспечения надежных результатов.

## Список литературы / References

[1] Aitchison J., Gilchrist A. and Bawden D., *Thesaurus construction and use: a practical manual*, Psychology Press, 2000, 230 pp.

[2] Лукашевич Н. В., Добров Б. В., “Проектирование лингвистических онтологий для информационных систем в широких предметных областях”, *Онтология проектирования*, **5**:1(15) (2015), 47–69; [Loukachevitch N. V., Dobrov B. V., “Developing Linguistic Ontologies in Broad Domains”, *Ontology of Designing*, **5**:1(15) (2015), 47–69, (in Russian).]

[3] Лукашевич Н. В., *Тезаурусы в задачах информационного поиска*, Издательство МГУ, М., 2011, 512 с.; [Lukashevich N. V., *Tezaurusy v zadachah informacionnogo poiska*, Izdatelstvo MGU, Moskow, 2011, 512 pp., (in Russian).]

[4] Astrakhantsev N. A., Turdakov D. Yu., “Automatic construction and enrichment of informal ontologies: A survey”, *Programming and computer software*, **39**:1 (2013), 34–42.

[5] Hasan K. S., Ng V., “Automatic Keyphrase Extraction: A Survey of the State of the Art”, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 2014, 1262–1273.

[6] Paramonov I. et al., "Thesaurus-Based Method of Increasing Text-via-Keyphrase Graph Connectivity During Keyphrase Extraction for e-Tourism Applications", *International Conference on Knowledge Engineering and the Semantic Web*, 2016, 129–141.

[7] Yang D., Powers D.Ṁ., "Automatic thesaurus construction", *Proceedings of the thirty-first Australasian conference on Computer science*, **74** (2008), 147–156.

[8] Mihalcea R., Tarau P., "TextRank: Bringing order into texts", *Proceedings of EMNLP*, 2004, 404–411.

[9] Liu Z. et al., "Automatic keyphrase extraction via topic decomposition", *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 2010, 366–376.

[10] Wiemer-Hastings P., Wiemer-Hastings K., Graesser A., "Latent semantic analysis", *Proceedings of the 16th international joint conference on Artificial intelligence*, 2004, 1–14.

[11] Lefever E., Van de Kauter M., Hoste V., "Evaluation of automatic hypernym extraction from technical corpora in English and Dutch", *9th International Conference on Language Resources and Evaluation (LREC)*, 2014, 490–497.

[12] Oakes M. P., "Using Hearst's Rules for the Automatic Acquisition of Hyponyms for Mining a Pharmaceutical Corpus", *RANLP Text Mining Workshop*, **5** (2005), 63–67.

[13] Noh S., Kim S., Jung C., "A Lightweight Program Similarity Detection Model using XML and Levenshtein Distance", *FECS*, 2006, 3–9.

[14] Мозжерина Е. С., "Автоматическое построение онтологии по коллекции текстовых документов", *Электронные библиотеки: Перспективные Методы и Технологии, Электронные коллекции–RCDL*, 2011, 293–298; [Mozzherina E. S., "Ehlektronnye biblioteki: Perspektivnye Metody i Tekhnologii, Ehlektronnye kollekcii–RCDL", 2011, 293–298, (in Russian).]

[15] Mittelu V. B., "Automatic Extraction of Patterns Displaying Hyponym-Hypernym Co-Occurrence from Corpora", *Proceedings of First Central European Student Conference in Linguistics*, 2006, 21, 8 pp.

---

**Lagutina N. S., Lagutina K. V., Mamedov E. I., Paramonov I. V.**, "Methodological Aspects of Semantic Relationship Extraction for Automatic Thesaurus Generation", *Modeling and Analysis of Information Systems*, **23**:6 (2016), 826–840.

**DOI:** 10.18255/1818-1015-2016-6-826-840

**Abstract.** The paper is devoted to analysis of methods for automatic generation of a specialized thesaurus. The main algorithm of generation consists of three stages: selection and preprocessing of a text corpus, recognition of thesaurus terms, and extraction of relations among terms. Our work is focused on exploring methods for semantic relation extraction. We developed a test bench that allow to test well-known algorithms for extraction of synonyms and hypernyms. These algorithms are based on different relation extraction techniques: lexico-syntactic patterns, morpho-syntactic rules, measurement of term information quantity, general-purpose thesaurus WordNet, and Levenstein distance. For analysis of the result thesaurus we proposed a complex assessment that includes the following metrics: precision of extracted terms, precision and recall of hierarchical and synonym relations, and characteristics of the thesaurus graph (the number of extracted terms and semantic relationships of different types, the number of connected components, and the number of vertices in the largest component). The proposed set of metrics allows to evaluate the quality of the thesaurus as a whole, reveal some drawbacks of standard relation extraction methods, and create more efficient hybrid methods that can generate thesauri with better characteristics than thesauri generated by using separate methods. In order to illustrate this fact, one of such hybrid methods is considered in the paper. It combines the best standard algorithms for hypernym and synonym extraction and generates a specialized medical thesaurus. The hybrid method leaves the thesaurus quality on the same level and finds more relations between terms than well-known algorithms.

**Keywords:** thesaurus, semantic relations, hybrid method, complex assessment, test bench

**About the authors:**
Nadezhda S. Lagutina, orcid.org/0000-0002-6137-8643, PhD,
P.G. Demidov Yaroslavl State University,
14 Sovetskaya str., Yaroslavl 150000, Russia, e-mail: lagutinans@rambler.ru

Ksenia V. Lagutina, orcid.org/0000-0002-1742-3240, student,
P.G. Demidov Yaroslavl State University,
14 Sovetskaya str., Yaroslavl 150000, Russia, e-mail: lagutinakv@mail.ru

Eldar I. Mamedov, orcid.org/0000-0003-1997-7084, intern researcher,
P.G. Demidov Yaroslavl State University,
14 Sovetskaya str., Yaroslavl 150000, Russia, e-mail: eldarmamedov90@gmail.com

Ilya V. Paramonov, orcid.org/0000-0003-3984-8423, PhD,
P.G. Demidov Yaroslavl State University,
14 Sovetskaya str., Yaroslavl 150000, Russia, e-mail: Ilya.Paramonov@fruct.org

**Acknowledgments:**
This work was supported by the grant of the President of Russian Federation for state support of young Russian scientists (project MK-5456.2016.9).