Моделирование и анализ информационных систем. Т. 24, №6 (2017), с. 772–787
Modeling and Analysis of Information Systems. Vol. 24, No 6 (2017), pp. 772–787

DOI: 10.18255/1818-1015-2017-6-772-787

УДК 004.912

Анализ использования различных типов связей между терминами тезауруса, сгенерированного с помощью гибридных методов, в задачах классификации текстов

Лагутина Н. С., Лагутина К. В., Щитов И. А., Парамонов И. В.[1,2]

получена 16 октября 2017

Аннотация. Цель данной статьи — проанализировать, насколько эффективно могут применяться различные типы тезаурусных связей в задачах классификации текстов. Основой исследования является автоматически сгенерированный тезаурус предметной области, содержащий три типа связей: синонимические, иерархические и ассоциативные. Для генерации тезауруса используется гибридный метод, основанный на нескольких лингвистических и статистических алгоритмах выделения семантических связей и позволяющий создать тезаурус с достаточно большим числом терминов и связей между ними. Авторы рассматривают две задачи: тематическая классификация текстов и классификация больших новостных статей по тональности. Для решения каждой из них авторами были использованы два подхода, каждый из которых дополняет стандартные алгоритмы процедурой, применяющей связи тезауруса для определения семантических особенностей текстов. Подход к тематической классификации включает в себя стандартный алгоритм BM25 вида «обучение без учителя» и процедуру, использующую синонимические и иерархические связи тезауруса предметной области. Подход к классификации по тональности состоит из двух шагов. На первом шаге создается тезаурус, тональные веса терминов которого считаются в зависимости от частоты встречаемости в обучаемой выборке или от веса соседей по тезаурусу. На втором шаге тезаурус применяется для вычисления признаков слов из текстов и классификации текстов методом опорных векторов или наивным байесовским классификатором. В экспериментах с корпусами BBCSport, Reuters, PubMed и корпусом статей об американских иммигрантах авторы варьировали типы связей, которые участвуют в классификации, и степень их использования. Результаты экспериментов позволяют оценить эффективность применения тезаурусных связей для классификации текстов на естественном языке и определить, при каких условиях те или иные связи имеют бóльшую значимость. В частности, наиболее полезными тезаурусными связями оказались синонимические и иерархические, так как они обеспечивает лучшее качество классификации.

Ключевые слова: тезаурус, семантические отношения, тезаурусные связи, тематическая классификация, классификация по тональности

Для цитирования: Лагутина Н. С., Лагутина К. В., Щитов И. А., Парамонов И. В., "Анализ использования различных типов связей между терминами тезауруса, сгенерированного с помощью гибридных методов, в задачах классификации текстов", *Моделирование и анализ информационных систем*, **24**:6 (2017), 772–787.

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, 150003 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, 150003 Россия, e-mail: lagutinakv@mail.ru

Щитов Иван Андреевич, orcid.org/0000-0002-5027-5024, аспирант,
Ярославский государственный университет им. П.Г. Демидова,
ул. Советская, 14, г. Ярославль, 150003 Россия, e-mail: ivan.shchitov@e-werest.org

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, 150003 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).
[2] Авторы благодарят заведующую кафедрой иностранных языков гуманитарных факультетов ЯрГУ, доцента Наталию Николаевну Касаткину за подготовку корпуса статей об американских иммигрантах.

Введение

Автоматическая обработка текстов на естественных языках является неотъемлемой частью современных информационных технологий. Необходимость анализа текстов возникает при решении таких задач, как поиск информации, поиск ответов на вопросы, рубрикация, классификация, аннотирование документов, машинный перевод и многое другое.

Огромное количество информации содержится как в коллекциях документов из Интернета, так и в книгах, статьях, узкоспециализированных информационных базах и т. п. Для качественного и эффективного использования этих ресурсов недостаточно рассматривать текст как набор независимых терминов, требуется учитывать структуру предложений и связь между словами и терминами предметной области. Это означает, что необходима модель, объединяющая знания о языке и особенностях предметной области. Такой моделью может служить тезаурус, включающий в себя термины предметной области и связи между ними.

Изначально тезаурусы создавались для индексирования документов вручную. В сфере автоматизации одно из первых применений тезауруса произошло в области машинного перевода в 1954 году [1]. Однако при автоматической обработке текстов между текстом и результатом работы нет человека-посредника, использующего не только тезаурус или словарь, но и свои знания о предмете исследования. Есть только автоматический процесс и тезаурус, который должен включать как набор общеупотребительных терминов и связей между ними, так и те знания, которые использует эксперт для анализа текста. Таким образом, тезаурус, предназначенный для автоматической обработки текстов, должен содержать значительно больше информации о предметной области, в частности, большее количество терминов и большее количество связей [2].

Расширение базы тезауруса ведет к увеличению и усложнению отношений между понятиями тезауруса. Поэтому, кроме развития и совершенствования методов выделения таких связей, необходимо понимание того, как разные виды связей влияют на эффективность использования тезаурусов в тех или иных задачах автоматической обработки текстов. Во многих современных исследованиях отношения между терминами активно используются, однако практически нигде виды связей не дифференцируются, и, тем более, не анализируется влияние разных типов связей на

качество решения рассматриваемой задачи. Поэтому авторы статьи в рамках своих исследований в области классификации текстов по тональности поставили вопрос о степени влияния связей терминов примененного тезауруса. В данной работе описаны проведенные в этой области эксперименты и их результаты.

1. Тезаурусы и виды связей

Тезаурус представляет собой словарь, включающий в себя термины или понятия, организованный таким образом, чтобы установить явные отношения между этими терминами [3]. Термин тезауруса — это слово или словосочетание, обычно в форме существительного или именной группы, являющееся точным обозначением определённого понятия какой-либо области знания. Опишем более подробно различные виды отношений между терминами тезауруса. Одним из основных является отношение синонимии — тождественность или близость значения слов, а также морфем, синтаксических конструкций, словосочетаний. В классических тезаурусах каждая группа синонимов формирует отдельную единицу словаря — синсет. Среди синонимов выбирается дескриптор — термин, который представляет отдельное понятие предметной области. Другие термины из синонимического ряда, включённые в синсет, называются аскрипторами [4].

В соответствии со стандартом основными типами отношений между синсетами тезауруса являются следующие:

- род — вид;
- часть — целое;
- причина — следствие;
- сырьё — продукт;
- процесс — объект;
- процесс — субъект;
- свойство — носитель свойства;
- функциональное сходство;
- административная иерархия;
- антонимия.

В тезаурусах, используемых для автоматической обработки текста, все эти отношения, как правило, делятся на два типа: иерархические и ассоциативные, где ассоциативная связь обозначает наличие связи между понятиями, отличающейся от синонимической и иерархической. Таким образом, тезаурус представляет собой модель предметной области, состоящую из терминов и синонимических, иерархических и ассоциативных связей между ними.

2. Обзор смежных работ

Влияние тезаурусных связей на качество анализа текста редко анализируется в научной литературе. Часть исследований на данную тему касается задачи индексирования документов и близкой к ней проблемы тематического моделирования, другие области обработки текстов практически не затрагиваются. Авторы проанализировали несколько работ, посвященных изучению структуры тезауруса и ее значения для решения практических задач.

В работе [5] анализируется влияние структуры тезауруса на качество предметного индексирования. Авторы разработали метод индексирования, который включает в себя алгоритм случайного блуждания. Данный алгоритм обрабатывает различные семантические связи с различными вероятностями, и, таким образом, связи из тезауруса влияют на результат в большей или меньшей степени. Авторы экспериментировали с четырьмя ручными тезаурусами: AGROVOC, HEP, NALT и MeSH — и вероятностями для иерархических и ассоциативных связей. Полученная средняя точность 19.01% достаточно низкая. Она достигается несколькими возможными комбинациями параметров метода: для трех тезаурусов необходимо задать высокую вероятность для ассоциаций и низкую для остальных связей. С тезаурусом HEP возникает противоположная ситуация, гиперонимы влияют сильнее других связей.

Авторы статьи [6] также применяют тезаурус для индексирования документов. Они разработали полуавтоматический инструмент DigiDoc MetaEdit, который позволяет пользователю соотносить термины из тезауруса с HTML-документами. Инструмент составляет набор возможных ключевых слов в зависимости от частоты появления и релевантности термина и дополняет этот набор тезаурусными синонимами, гипонимами, гиперонимами и ассоциациями. При этом параметры алгоритма задаются пользователем.

Эксперименты проводились со 100 испанскими журнальными статьями с портала BiD и тезаурусом TLIS (Thesaurus on Library and Information Science). Авторы сравнивали эффективность автоматической части инструмента для случаев, когда он применяет различные виды тезаурусных связей. Лучшая полнота 73% достигается при использовании всех видов связей. Использование только горизонтальных связей позволяет достичь только 64%, иерархических — 58%. Алгоритм без тезауруса демонстрирует самую низкую полноту — 49%.

Н. Лукашевич с соавторами [7] предлагают метод решения задачи тематического моделирования, который дополняет наборы близких тем терминами из тезауруса, созданного вручную. Авторы провели несколько экспериментов с алгоритмами без тезауруса, с синонимами и с синонимами и гиперонимами, используя для измерения качества метрику уникальности ядра. Она описывает, насколько различные темы, определяемые группами связанных тезаурусных терминов, отличаются друг от друга. В результате комбинация двух связей обеспечивает лучшее значение данной метрики: 0.4–0.7, тогда как в случае алгоритма без тезауруса эта характеристика равна 0.3–0.5.

Данные исследования демонстрируют, что различные способы использования связей из тезауруса существенно влияют на результат анализа текста на естественном языке. К сожалению, значимость тезаурусных связей в данной области изучена

недостаточно, в том числе для задач классификации текстов по темам или тональности. В методах решения этих задач обычно применяется один тип связей или все связи унифицируются и обрабатываются одинаково как ассоциации.

В работе [8] решается задача классификации текстов по темам при помощи тезауруса. Тезаурус OpenOffice бразильского варианта португальского языка является вспомогательным инструментом для соотнесения текстов и терминов из онтологии. Алгоритм ищет для каждого слова из текста ближайшие связанные термины из тезауруса и отбирает те из них, которые содержатся также и в онтологии. На основе этих данных работает предложенный авторами классификатор. Точность и полнота результата вырастают на 7–10% (до 60–70%) по сравнению с методом опорных векторов без тезауруса. На следующем этапе исследования тезаурус и онтология дополняются терминами и связями вручную при помощи экспертов, а также авторы изменяют меру близости термина и текста. Это позволяет достичь одних из лучших результатов по области: 96% точности и полноты.

В статье [9] исследуется проблема классификации отзывов пользователей Интернета на позитивные и негативные. Авторы предлагают классификатор, основанный на применении автоматически созданного тезауруса. Тезаурус строится по всему корпусу текстов, на котором впоследствии будет обучаться и работать классификатор. Слова маркируются на положительные и отрицательные в зависимости от тональности текстов, в которых они чаще встречаются. Термины, появляющиеся в одном предложении, считаются связанными ассоциативной связью, причем вес данной связи также зависит от частоты встречаемости терминов в одном и том же тексте. Далее вычисляются веса пар термин–отзыв в зависимости от количества и веса связей в тезаурусе между данным термином и другими словами отзыва. Эти характеристики формируют векторы отзывов, которые в итоге поступают на вход классификатору, использующему метод максимальной энтропии. В результате получены следующие результаты для 800 позитивных и 800 негативных отзывов с Amazon: точность без тезауруса 62–72%, точность с тезаурусом 72–87% — одна из лучших для данной задачи. Предлагаемый метод позиционируется как независимый от предметной области, т.е. его потенциально можно использовать для текстов отзывов по любой теме.

Таким образом, несмотря на то, что большинство исследователей не дифференцируют семантические связи различных типов между терминами тезауруса для классификации текстов, данная идея была успешно применена для некоторых задач анализа текстов на естественном языке, поэтому ее можно попытаться распространить и на сферу классификации.

3. Классификация текстов по темам с применением тезауруса

Тематическая классификация текстов — это задача разделения корпуса текстов на несколько классов (возможно пересекающихся), каждый из которых обозначает основную тему текста. Для целей данного исследования авторы используют алгоритм тематической классификации, который дополняет существующий алгоритм

BM25 [10] процедурой, применяющей специализированный тезаурус, разработанный авторами.

3.1. Генерация тезауруса

Специализированный тезаурус генерируется полностью автоматически на корпусе текстов заданной предметной области, которые затем будут классифицироваться. Это позволяет создать модель предметной области, описывающую основные темы данной области и связи между ними. Подробная характеристика разработанного гибридного алгоритма и получающегося в результате тезауруса дана в предыдущей статье авторов [11]. Данный алгоритм объединяет в себе несколько существующих статистических и лингвистических алгоритмов выделения тезаурусных связей, за счет этого он позволяет построить значительное количество различных семантических связей между терминами. Краткое описание шагов алгоритма выглядит следующим образом:

1. Выделение терминов алгоритмом TextRank [12].
2. Построение семантических связей между терминами (ассоциативных, синонимических и гипонимо-гиперонимических) гибридным методом.
3. Фильтрация терминов, не имеющих связей.

В результате работы алгоритма создается тезаурус с большим набором терминов и различных связей между ними. Также основным достоинством алгоритма является то, что он работает полностью автоматически и не требует вмешательства эксперта ни на одном этапе, поэтому тезаурус генерируется достаточно быстро. Сравнение с существующим тезаурусом, построенным вручную, показало, что автоматический тезаурус показывает хорошие точность и полноту терминов и связей, поэтому он может успешно применяться для обработки текстовой информации [11].

3.2. Классификация текстов

Построенный тезаурус применяется авторами для тематической классификации текстов на естественном языке. Автоматическая классификация осуществляется при помощи стандартного алгоритма BM25 вида «обучение без учителя», который дополнительно учитывает связи между терминами тезауруса. Входными данными для алгоритма являются корпус текстов, не размеченных по классам, список классов и автоматический тезаурус.

Алгоритм состоит из следующих шагов:

1. Выделение всех существительных и прилагательных из текстов и вычисление частот их встречаемости.
2. Создание для каждого класса списка терминов-соседей по тезаурусу.
3. Ранжирование пар текст–класс алгоритмом BM25 в зависимости от частоты встречаемости терминов класса и их соседей.

Сначала алгоритм выбирает из текстов отдельные термины, строит для каждого текста инвертированный индекс и считает частоту встречаемости каждого слова.

Затем для терминов класса находятся соседи по тезаурусу первого порядка, т.е. термины, имеющие с ними прямую связь — синонимическую, гипонимическую или гиперонимическую. Ассоциации не обрабатываются, так как они часто обозначают довольно слабую семантическую связь и даже могут связывать слова из разных тем, например, «тромб» и «сердечный клапан». С синонимическими и иерархическими связями ситуация обратная: они в большинстве случаев отражают связи между терминами из одной и той же темы, так что лучше подходят для тематического сопоставления текстов и классов.

На последнем шаге применяется BM25 для пар текст–класс, чтобы ранжировать тексты в зависимости от терминов классов и их терминов-соседей по тезаурусу. Данный алгоритм популярен при решении различных задач анализа текстов, в том числе классификации [13]. Также он не требует настройки дополнительных параметров и обучения на существующих корпусах данных, поэтому он легко дополняется обработкой информации из тезауруса, так что хорошо подходит для целей данного исследования.

4. Классификация текстов по тональности с применением тезауруса

Проблема классификации текстов по тональности подразумевает разделение корпуса текстов на два или более класса в зависимости от тональности текста в целом: на положительные или отрицательные, или на положительные, отрицательные и нейтральные. В данной работе рассматривается задача классификации больших новостных статей на два класса с использованием автоматически сгенерированного тезауруса.

Предлагаемый авторами подход состоит из следующих двух шагов: создание тонального тезауруса и классификация текстов. На первом шаге полностью автоматически создается тезаурус, на втором статьи классифицируются при помощи метода опорных векторов и наивного байесовского классификатора, причем векторы признаков составляются на основе весов терминов в тезаурусе. Оба шага получают на вход корпус необработанных новостных статей, предварительно разделенный на обучающую и тестовую выборки. Обучающая выборка изначально была промаркирована экспертами-филологами.

4.1. Генерация тезауруса

В качестве терминов тонального тезауруса выбираются все слова, которые могут иметь положительную или отрицательную семантику: существительные, прилагательные, глаголы и наречия. Связи между ними строятся по алгоритму, описанному в предыдущей главе. Результатом данного алгоритма является специализированный тезаурус с большим количеством синонимических, ассоциативных и гипонимо-гиперонимических связей. Большое количество связей может позволить вычислить

более точную числовую характеристику тональности термина, так как тональность в данном случае будет зависеть сразу от нескольких соседей термина по тезаурусу.

На последнем этапе генерации терминам сопоставляются тональные веса, обозначающие тональность: от -1 до 0 для отрицательных и от 0 до 1 для положительных терминов. Сначала считаются веса тех терминов, которые встречаются в обучающей выборке текстов. Поскольку тексты в данной выборке уже промаркированы по тональности, её можно распространить и на термины.

Авторы предположили, что положительные термины чаще встречаются в положительных текстах, а отрицательные — в отрицательных, поэтому использовали для веса термина следующую формулу: $w = (p-n)/(p+n)$, где p обозначает, сколько раз термин появился в положительных текстах, n — в отрицательных. Данная формула позволяет назначить тональности из диапазона от -1 до 1, описанного выше.

Далее необходимо вычислить веса терминов, которые не встречаются в обучающей выборке, но появляются в тестовой. Авторы предположили, что термины, имеющие общую тезаурусную связь, имеют и близкую тональность. Поэтому тональность можно вычислить, опираясь на веса соседей по тезаурусу. Для реализации данной идеи каждому типу связи был назначен свой коэффициент для преобразования тональности, который варьировался от 0.1 до 1.0.

Алгоритм вычисления весов по тезаурусным связям следующий. Если у термина есть синонимы, маркированные по тональности, берется их средний вес и умножается на коэффициент связи, результат записывается как вес термина. Если у термина таких синонимов нет, но есть гиперонимы, аналогично поступают с ними. И если у термина имеются только ассоциации, среднее значение, умноженное на коэффициент, считается для них. Так как связей в тезаурусе достаточно много, у каждого термина будет хотя бы один маркированный сосед.

Таким образом, сгенерированный тезаурус содержит набор терминов из предметной области, семантические связи между ними и тональные веса всех терминов.

4.2. Классификация текстов

Тональность терминов созданного тезауруса используется для вычисления векторов признаков, используемых далее в алгоритме классификации. Каждому тексту ставится в соответствие числовой вектор, размер которого равен количеству терминов в тезаурусе. Каждый элемент вектора соответствует конкретному тезаурусному термину и считается как произведение $w \cdot F$, где w — это вес термина, а F — некоторая стандартная статистическая характеристика пары термин—текст, зависящая от частоты встречаемости термина в корпусе или тексте и/или от его веса в тезаурусе. Авторы использовали четыре различные статистические характеристики: TF*IDF, коэффициент Джини, расстояние Кульбака—Лейблера, взаимная информация и критерий χ^2 [14].

На последнем этапе векторы признаков подаются на вход одному из стандартных бинарных классификаторов. Многие исследования показывают, что лучшими классификаторами для вычисления тональности текстов являются алгоритмы машинного обучения [15]. Среди данных алгоритмов одними из самых эффективных

считаются метод опорных векторов и наивный байесовский классификатор [16], поэтому они были выбраны авторами для реализации предлагаемого подхода.

5. Корпуса текстов и методика проведения экспериментов

Предложенные авторами алгоритмы классификации тестировались на нескольких корпусах англоязычных текстов из различных областей знаний:

- Корпус PubMed (https://www.nlm.nih.gov/databases/download/pubmed_medline.html) содержит 1000 медицинских статей из 63 классов, общее количество слов — 154 850.

- Корпус Reuters (http://www.daviddlewis.com/resources/testcollections/reuters21578/) содержит 1534 экономические статьи из 15 классов, общее количество слов — 294 813.

- Корпус BBCSport (http://mlg.ucd.ie/datasets/bbc.html) содержит 737 новостных спортивных статей из 5 классов, общее количество слов — 253 667.

- Корпус статей об американских иммигрантах из газет The New York Times, The New York Post и The Los Angeles Times. Он содержит 56 статей, из них 34 положительных и 22 отрицательных, общее количество слов — 37 669. Корпус был создан и размечен по тональности экспертами-филологами.

Первые три корпуса текстов использовались для тематической классификации, последний — для классификации по тональности. Также четвертый корпус был разделен на обучающую и тестовую выборки, по 17 положительных и 11 отрицательных статей в каждой.

Оба алгоритма классификации реализованы на языке программирования Python с использованием библиотеки NLTK (http://www.nltk.org). Данная библиотека содержит реализации стандартных алгоритмов обработки текстов и машинного обучения, в том числе наивного байесовского классификатора и метода опорных векторов.

Процедура оценки результатов классификации также была написана на языке Python. Авторы применили для оценки стандартные метрики качества: точность, полноту, F-меру и долю правильных ответов [17].

Следует отметить, что для тематической классификации были выбраны микроусредненные метрики, которые считают полноту и точность одновременно по всем классам, так как классов у текстов достаточно много и подсчет отдельных значений метрик для каждого класса усложнит анализ результата. Задача второго типа, классификация по тональности, подразумевает деление всего на два класса, поэтому точность, полнота и F-мера считались отдельно для положительных и отрицательных текстов.

6. Результаты экспериментов

Авторы поставили несколько экспериментов с разработанными алгоритмами и при этом варьировали их параметры: подключали различные типы связей по отдельности и в комбинации. Для алгоритма тематической классификации также изменялся порог фильтрации: если результат BM25 был меньше определенного порога (0, 2 или 4), пара класс–текст отвергалась. Для алгоритма классификации по тональности рассматривались различные коэффициенты связей от 0.1 до 1.0 с шагом 0.1. В данном случае коэффициенты для разных типов связей использовались, чтобы проанализировать изменение степени влияния каждого типа связи в отдельности.

Таблица 1. Тематическая классификация корпуса BBCSport
Table 1. Topical classification of the BBCSport corpus

Связи	Порог фильтрации	P	R	F	A
нет	4	0.835	0.096	0.173	0.815
синонимы	4	0.828	0.098	0.175	0.815
гипонимы	4	0.790	0.107	0.189	0.816
гиперонимы	0	0.403	0.479	0.438	0.754
синонимы и гипонимы	4	0.784	0.109	0.191	0.816
синонимы и гиперонимы	0	0.399	0.482	0.437	0.751

Таблица 2. Тематическая классификация корпуса Reuters
Table 2. Topical classification of the Reuters corpus

Связи	Порог фильтрации	P	R	F	A
нет	4	0.738	0.246	0.369	0.933
синонимы	2	0.519	0.559	0.538	0.924
синонимы	4	0.749	0.276	0.404	0.935
все	0	0.329	0.746	0.457	0.859

Таблица 3. Тематическая классификация корпуса PubMed
Table 3. Topical classification of the PubMed corpus

Связи	Порог фильтрации	P	R	F	A
нет	4	0.234	0.065	0.101	0.943
синонимы	4	0.216	0.072	0.109	0.941
гиперонимы	2	0.169	0.169	0.169	0.917
все	0	0.125	0.201	0.154	0.890

В таблицах 1, 2, 3 представлены лучшие результаты классификации по темам для каждого корпуса текстов. Символы P, R и F обозначают точность, полноту и F-меру соответственно. A — это доля правильных ответов (от accuracy). «Нет»

в первом столбце значит, что данный эксперимент проводился с алгоритмом без связей, а «все» — что подключались все связи: гипонимы, гиперонимы и синонимы.

Из результатов классификации новостей BBCSport видно, что лучшую точность 0.835, но самую низкую полноту 0.096 обеспечивает алгоритм, который не использует тезаурус. Лучшая полнота 0.482 и вместе с тем худшая точность 0.399 у алгоритма, использующего синонимы и гиперонимы и вместе с тем не использующего фильтрацию. Примечательно, что его F-мера 0.437 отличается от лучшей всего на 0.1, которая является результатом алгоритма с гиперонимами. Лучшая доля правильных ответов 0.816 у алгоритмов с синонимами и гипонимами, т.е. они дают больше всего правильных ответов, но F-мера у всех них низкая: 0.175–0.191. Таким образом, лучшие характеристики для корпуса BBCSport достигаются за счет использования иерархических связей.

Для корпуса Reuters самыми значимыми связями оказались синонимы. Именно алгоритмы с ними обеспечивают лучшую F-меру 0.538 и лучшую долю правильных ответов 0.935.

Результаты классификации PubMed получились низкими для всех комбинаций тезаурусных связей. Алгоритм выдал слишком мало пар класс–текст, поэтому точность и полнота не превышают 0.25. Доля правильных ответов высока за счет того, что она учитывает, сколько текстов было верно не соотнесено с неподходящими классами, и за счет большого количества классов (63) это число достаточно большое.

Качество классификации по тональности выше тематической в абсолютных числах. Ее результаты представлены в таблицах 4–10. P, R и F обозначают точность, полноту и F-меру, нижние индексы neg и pos — отрицательный и положительный классы текстов. Слова Hyp, Syn и $Assoc$ подразумевают использование в эксперименте гиперонимов, синонимов и ассоциаций.

Таблица 4. Классификация по тональности с методом опорных векторов и коэффициентом Джини
Table 4. Sentiment classification with SVM and Gini Index

Связи	Коэффициенты	A	P_{neg}	R_{neg}	F_{neg}	P_{pos}	R_{pos}	F_{pos}
нет	0	0.643	0.571	0.364	0.444	0.667	0.824	0.737
Hyp	0.1, 0.3, 0.7, 0.9	0.750	1.000	0.364	0.533	0.708	1.000	0.829
Syn и $Assoc$	1.0 и 0.6	0.714	0.800	0.364	0.500	0.696	0.941	0.800

Таблица 5. Классификация по тональности с наивным байесовским классификатором и коэффициентом Джини
Table 5. Sentiment classification with Naive Bayes and Gini Index

Связи	Коэффициенты	A	P_{neg}	R_{neg}	F_{neg}	P_{pos}	R_{pos}	F_{pos}
нет	0	0.607	0.500	0.273	0.353	0.636	0.824	0.718
Syn	0.5	0.714	0.800	0.364	0.500	0.696	0.941	0.800
$Assoc$	0.6	0.679	0.667	0.364	0.471	0.682	0.882	0.769
Syn и Hyp	1.0 и 1.0	0.679	0.667	0.364	0.471	0.682	0.882	0.769

Таблица 6. Классификация по тональности с методом опорных векторов и расстоянием Кульбака—Лейблера
Table 6. Sentiment classification with SVM and information gain

Связи	Коэффициенты	A	P_{neg}	R_{neg}	F_{neg}	P_{pos}	R_{pos}	F_{pos}
нет	0	0.571	0.444	0.364	0.400	0.632	0.706	0.667
Syn и *Hyp*	1.0 и 0.5, 0.6, 0.9	0.750	0.833	0.455	0.588	0.727	0.941	0.821
Syn и *Assoc*	1.0 и 0.5	0.714	0.714	0.455	0.556	0.714	0.882	0.789

Таблица 7. Классификация по тональности с наивным байесовским классификатором и расстоянием Кульбака—Лейблера
Table 7. Sentiment classification with Naive Bayes and information gain

Связи	Коэффициенты	A	P_{neg}	R_{neg}	F_{neg}	P_{pos}	R_{pos}	F_{pos}
нет	0	0.643	0.667	0.182	0.286	0.640	0.941	0.762
Syn	1.0	0.714	0.800	0.364	0.500	0.696	0.941	0.800
Syn и *Hyp*	1.0 и 0.1, 0.9	0.750	1.000	0.364	0.533	0.708	1.000	0.829
Syn и *Assoc*	1.0 и 0.3	0.714	0.800	0.364	0.500	0.696	0.941	0.800

Таблица 8. Классификация по тональности с методом опорных векторов и взаимной информацией
Table 8. Sentiment classification with SVM and mutual information

Связи	Коэффициенты	A	P_{neg}	R_{neg}	F_{neg}	P_{pos}	R_{pos}	F_{pos}
нет	0	0.607	0.500	0.364	0.421	0.650	0.765	0.703
Syn	0.1–0.9	0.643	0.571	0.364	0.444	0.667	0.824	0.737
Hyp	0.5	0.714	0.800	0.364	0.500	0.696	0.941	0.800
Syn и *Hyp*	1.0 и 0.4, 0.8	0.643	0.571	0.364	0.444	0.667	0.824	0.737

Таблица 9. Классификация по тональности с наивным байесовским классификатором и взаимной информацией
Table 9. Sentiment classification with Naive Bayes and mutual information

Связи	Коэффициенты	A	P_{neg}	R_{neg}	F_{neg}	P_{pos}	R_{pos}	F_{pos}
нет	0	0.607	0.500	0.273	0.353	0.636	0.824	0.718
Hyp	1.0	0.750	0.833	0.455	0.588	0.727	0.941	0.821
Syn и *Hyp*	1.0 и 0.5, 0.8	0.750	0.750	0.545	0.632	0.750	0.882	0.811

Каждая таблица описывает результаты классификации статей об американских иммигрантах для пар классификатор–характеристика веса термина. В таблицы не вошли пары с TF*IDF и пара байесовский классификатор — χ^2, так как их результаты оказались ниже остальных и при этом не зависели от способа использования тезауруса. Отметим, что во всех случаях результаты для алгоритмов с тезуарусом

Таблица 10. Классификация по тональности с методом опорных векторов и χ^2
Table 10. Sentiment classification with SVM and χ^2

Связи	Коэффициенты	A	P_{neg}	R_{neg}	F_{neg}	P_{pos}	R_{pos}	F_{pos}
нет	0	0.643	0.571	0.364	0.444	0.667	0.824	0.737
Hyp	0.2, 0.6	0.679	0.667	0.364	0.471	0.682	0.882	0.769
Syn и *Hyp*	1.0 и 0.4	0.679	0.667	0.364	0.471	0.682	0.882	0.769
Syn и *Hyp*	1.0 и 0.8	0.679	0.625	0.455	0.526	0.700	0.824	0.757
Syn и *Assoc*	1.0 и 0.1–0.9	0.679	0.625	0.455	0.526	0.700	0.824	0.757

оказались выше результатов алгоритма без него, а в большинстве случаев — значительно выше.

В сочетании с методом опорных векторов и коэффициентом Джини (таблица 4) лучший результат по всем характеристикам дало использование гиперонимов, причем коэффициент гиперонимов практически не играет роли: он может быть 0.1, 0.3, 0.7 или 0.9. Для сочетания байесовский классификатор — коэффициент Джини (таблица 5) то же можно сказать про синонимы с коэффициентом 0.5.

При экспериментах с комбинацией метода опорных векторов и расстояния Кульбака—Лейблера (таблица 6) снова выделяется конкретный случай с синонимами и гиперонимами, который обеспечивает лучшие результаты. Следует отметить, что случай с синонимами и ассоциациями незначительно хуже, почти все его метрики отличаются от лучших не больше чем на 5%. И для комбинации байесовский классификатор — расстояние Кульбака—Лейблера (таблица 7) лучшие результаты обеспечиваются именно синонимами с коэффициентом 1.0.

Для комбинаций метод опорных векторов — взаимная информация (таблица 8) и байесовский классификатор — взаимная информация (таблица 9) являются значимыми и синонимы, и гиперонимы, так как именно они и их пара позволяют достичь самых высоких значений метрик качества 0.75–0.95. Примечательно, что коэффициент синонимов был равен 1.0, а коэффициент гиперонимов колебался от 0.4 до 1.0.

При экспериментах с комбинацией метода опорных векторов и χ^2 (таблица 10) было обнаружено, что тип используемой тезаурусной связи не играет большой роли: хорошие результаты были обеспечены как гиперонимами и их парой с синонимами, так и парой синонимы — ассоциации, причем коэффициент ассоциаций оказался не значимым.

Подводя итоги всех экспериментов, авторы обнаружили следующие закономерности. Использование тезаурусных связей существенно увеличивает качество результата по всем характеристикам по сравнению с алгоритмами без тезауруса. Самыми значимыми связями в тезаурусе оказались синонимические и иерархические, так как в решениях обеих задач классификации именно они обеспечили лучшие результаты: F-меру 0.5–0.8 и долю правильных ответов 0.75–0.9. Ассоциативные связи не сделали существенного вклада в качество результата, как видно практически из всех экспериментов с разными параметрами.

Другая закономерность заключается в зависимости способа использования тезауруса от предметной области текстов. Эксперименты показали, что для новостных

текстов BBCSport более качественный результат дают иерархические связи, а для экономических текстов Reuters — синонимические. Для больших газетных статей об иммигрантах оказались важными сразу два типа связей: синонимы и гиперонимы, причем ни одну из них нельзя выделить как наиболее значимую.

Заключение

Исследование результатов классификации текстов из разных предметных областей позволяет сделать не только общий вывод о наиболее существенном влиянии синонимических и иерархических связей по сравнению с ассоциативными, но и выделить более детальные закономерности. Наиболее важным с точки зрения авторов является то, что в ходе классификации коротких новостных текстов существенное влияние оказал один тип связей, в то время как при классификации статей большого объема наилучшие результаты оказались при использовании всех связей между терминами тезауруса. Это может быть обусловлено тем, что полноценные публицистические тексты на естественном языке используют гораздо более богатую лексику и структуру предложений, в отличие от коротких новостных публикаций. Поэтому для анализа таких текстов необходим тезаурус, обладающий большим количеством терминов и связей между ними.

Другой важный вывод заключается в том, что влияние разных типов связей между терминами зависит от конкретной предметной области. Это наблюдение нуждается в дополнительном тщательном исследовании. Еще одно направление для дальнейшей работы заключается в детализации ассоциативных связей и анализе их влияния при обработке больших текстов, как публицистических, так и научных, художественных и т.д. В любом случае, исследования показали, что связи между терминами являются важнейшей частью тезауруса как модели предметной области и что тщательный выбор способа их использования позволяет более качественно решать задачи автоматической обработки текста.

Список литературы / References

[1] Masterman M., “Semantic message detection for machine translation, using an interlingua”, *Proc. 1961 International Conf. on Machine Translation*, 1961, 438–475.

[2] Loukachevitch N., Dobrov B., “The Sociopolitical Thesaurus as a resource for automatic document processing in Russian”, *Terminology*, **21**:2 (2015), 237–262.

[3] Aitchison J., Clarke S.D., “The thesaurus: a historical viewpoint, with a look to the future”, *Cataloging and classification quarterly*, **37**:3–4 (2004), 5–21.

[4] 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).]

[5] Willis C., Losee R., “A random walk on an ontology: Using thesaurus structure for automatic subject indexing”, *Journal of the American Society for Information Science and Technology*, **64**:7 (2013), 1330–1344.

[6] Vállez M., Pedraza-Jiménez R., Codina L., Blanco S., Rovira C, “A semi-automatic indexing system based on embedded information in HTML documents”, *Library Hi Tech*, **33**:2 (2015), 195–210.

[7] Loukachevitch N., Nokel M., Ivanov K., *Combining Thesaurus Knowledge and Probabilistic Topic Models*, 2017, https://arxiv.org/abs/1707.09816.

[8] Sanchez-Pi N., Martí L. Garcia A. C. B., "Improving ontology-based text classification: An occupational health and security application", *Journal of Applied Logic*, **17** (2016), 48–58.

[9] Bollegala D., Weir D., Carroll J., "Cross-domain sentiment classification using a sentiment sensitive thesaurus", *IEEE transactions on knowledge and data engineering*, **25**:8 (2013), 1719–1731.

[10] Sparck Jones K., Walker S., Robertson S.E., "A probabilistic model of information retrieval: development and comparative experiments: Part 2", *Information Processing and Management*, **36**:6 (2000), 809–840.

[11] 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; [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, (in Russian).]

[12] Mihalcea R., Tarau P., "TextRank: Bringing order into texts", *Proceedings of Empirical Methods in Natural Language Processing – EMNLP*, ACL, Barcelona, Spain, 2004, 404–411.

[13] Trieschnigg D., Pezik P., Lee V., De Jong F., Kraaij W., Rebholz-Schuhmann D., "MeSH Up: effective MeSH text classification for improved document retrieval", *Bioinformatics*, **25**:11 (2009), 1412–1418.

[14] Aggarwal C., Zhai C., "A survey of text classification algorithms", *Mining text data*, Springer-Verlag, New York, 2012, 163–222.

[15] Grimmer J., Stewart B., "Text as data: The promise and pitfalls of automatic content analysis methods for political texts", *Political analysis*, **21**:3 (2013), 267–297.

[16] Ravi K., Ravi V., "A survey on opinion mining and sentiment analysis: tasks, approaches and applications", *Knowledge-Based Systems*, **89** (2015), 14–46.

[17] Junker M., Hoch R., Dengel A., "On the evaluation of document analysis components by recall, precision, and accuracy", *Proceedings of the Fifth International Conference on Document Analysis and Recognition*, IEEE, 1999, 713–716.

Lagutina N. S., Lagutina K. V., Shchitov I. A., Paramonov I. V., "Analysis of Influence of Different Relations Types on the Quality of Thesaurus Application to Text Classification Problems", *Modeling and Analysis of Information Systems*, **24**:6 (2017), 772–787.

DOI: 10.18255/1818-1015-2017-6-772-787

Abstract. The main purpose of the article is to analyze how effectively different types of thesaurus relations can be used for solutions of text classification tasks. The basis of the study is an automatically generated thesaurus of a subject area, that contains three types of relations: synonymous, hierarchical and associative. To generate the thesaurus the authors use a hybrid method based on several linguistic and statistical algorithms for extraction of semantic relations. The method allows to create a thesaurus with a sufficiently large number of terms and relations among them. The authors consider two problems: topical text classification and sentiment classification of large newspaper articles. To solve them, the authors developed two approaches that complement standard algorithms with a procedure that take into account thesaurus relations to determine semantic features of texts. The approach to topical classification includes the standard unsupervised BM25 algorithm and the procedure, that take into account synonymous and hierarchical relations of the thesaurus of the subject area. The approach to sentiment classification consists of two steps. At the first step, a thesaurus is created, whose terms

weight polarities are calculated depending on the term occurrences in the training set or on the weights of related thesaurus terms. At the second step, the thesaurus is used to compute the features of words from texts and to classify texts by the algorithm SVM or Naive Bayes. In experiments with text corpora BBCSport, Reuters, PubMed and the corpus of articles about American immigrants, the authors varied the types of thesaurus relations that are involved in the classification and the degree of their use. The results of the experiments make it possible to evaluate the efficiency of the application of thesaurus relations for classification of raw texts and to determine under what conditions certain relationships affect more or less. In particular, the most useful thesaurus connections are synonymous and hierarchical, as they provide a better quality of classification.

Keywords: thesaurus, semantic relations, thesaurus relations, topical classification, sentiment classification

On the authors:
Nadezhda S. Lagutina, orcid.org/0000-0002-6137-8643, PhD, associate professor
P.G. Demidov Yaroslavl State University,
14 Sovetskaya str., Yaroslavl, 150003 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, 150003 Russia, e-mail: lagutinakv@mail.ru

Ivan A. Shchitov, orcid.org/0000-0002-5027-5024, graduate student,
P.G. Demidov Yaroslavl State University,
14 Sovetskaya str., Yaroslavl, 150003 Russia, e-mail: ivan.shchitov@e-werest.org

Ilya V. Paramonov, orcid.org/0000-0003-3984-8423, PhD, associate professor
P.G. Demidov Yaroslavl State University,
14 Sovetskaya str., Yaroslavl, 150003 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).
[2] The authors would like to thank Natalia Kasatkina, associate professor, head of the Department of Foreign Languages of Humanities in P.G. Demidov Yaroslavl State University, for preparation of the corpus of articles about American immigrants.