КАЗАХСКИЙ НАЦИОНАЛЬНЫЙ УНИВЕРСИТЕТ ИМ. АЛЬ-ФАРАБИ

Алия Нугуманова, Мадина Мансурова

АВТОМАТИЧЕСКОЕ РАСПОЗНАВАНИЕ ТЕРМИНОВ В ТЕКСТАХ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

Монография

Алматы
«Қазақ университеті»
2018

УДК 004(89)
ББК
Н

Рекомендовано к изданию Ученым советом КазНУ им. аль-Фараби (протокол №2 от 29.10.2018)

Рецензенты:
Макашев Е.П., к.ф.-м.н., доцент кафедры информатики КазНУ имени аль-Фараби
Жантасова Ж.З., кандидат т.н., зав. кафедрой компьютерного моделирования и информационных технологий Восточно-Казахстанского государственного университета им. С. Аманжолова

Нугуманова Алия, Мансурова Мадина

Автоматическое распознавание терминов в текстах на естественном языке: монография / Нугуманова Алия, Мансурова Мадина. – Алматы: Қазақ университеті, 2018. – 51 с.

ISBN 978-601-04-3686-2

Автоматическое извлечение терминов из текстов предметной области представляет собой задачу, которая имеет множество приложений. Термины, извлекаемые автоматическим способом, могут использоваться как классификационные признаки для рубрикации документов, как семантические концепты для генерации тезаурусов и онтологий, как опорные понятия для контент-анализа СМИ. Практически во всех задачах, связанных с автоматической обработкой текстов, как то аннотирование, индексирование, классификация, машинный перевод, извлечение знаний и т.д., требуется извлечение терминологии. Для решения указанной задачи разработано большое количество эффективных методов, которые позволяют автоматизировать извлечение терминов из текстов предметной области. Данная монография призвана охарактеризовать и систематизировать эти методы, чтобы определить перспективы и проблемы развития технологий обработки естественного языка. Значимость определяемых проблем обусловлена феноменом информационного взрыва, который переживает современное общество. Для научных работников, специалистов в области обработки естественного языка, преподавателей вузов, докторантов, магистрантов и студентов.

ISBN 978-601-04-3686-2

Введение

Автоматическое распознавание терминов на основе текстов на естественном языке – это одновременно сложная и популярная научная задача, имеющая множество приложений в информационном поиске и инженерии знаний. Решение этой задачи направлено на автоматическое формирование терминологического лексикона какой-либо предметной области. Такая постановка задачи возникла как альтернатива традиционному ручному извлечению терминов, при котором специалист по терминологии сначала составляет список терминов-кандидатов, а затем консультируется с экспертом предметной области для утверждения окончательного лексикона.

Актуальность задачи автоматического распознавания терминов определяется тем, что в условиях быстро меняющегося мира, где постоянно появляются новые технологические отрасли, возникают новые понятия и термины, а объем технической лексики увеличивается экспоненциально, ручное построение и описание терминологии представляет собой трудоемкое предприятие. Поэтому большую теоретическую и практическую ценность представляют методы выделения терминов с помощью программных инструментов.

Извлеченные программными способами списки терминов являются наиболее простыми машиночитаемыми структурами знаний, но несмотря на простоту, их роль в информационном поиске и других практических приложениях трудно переоценить. С помощью машиночитаемых списков терминов можно без привлечения экспертов аннотировать и индексировать документы, устанавливать их тематическую направленность (рубрицировать) и осуществлять поддержку машинного перевода. Также машиночитаемые списки терминов могут использоваться в качестве строительного материала для более сложных структур знаний, например, таксономий и онтологий. В работе [1] отмечается, что машиночитаемые тезаурусы чрезвычайно важны для цифровых

библиотек, они позволяют организовать простую и эффективную навигацию внутри библиотеки и быстрый поиск по ее ресурсам.

В литературе существует множество обозначений для автоматического распознавания терминов: извлечение терминов (term extraction), распознавание терминов (term recognition), идентификация терминов (term identification), обнаружение терминов (term detection), получение терминов (term acquisition) и добыча терминов (term mining). Несмотря на то, что между этими обозначениями имеются небольшие различия, в данной работе мы рассматриваем их как синонимы (см. рисунок 1).

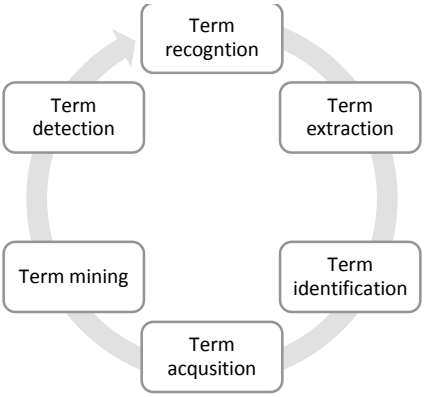


Рисунок 1 – Существующие в литературе обозначения проблемы автоматического распознавания терминов

В работе [2] выделяют 5 последовательных этапов, из которых состоит процесс автоматического распознавания терминов на основе текстов предметной области (см. рисунок 2):

1. Сборка корпуса (corpus collection) – компиляция репрезентативного корпуса текстов предметной области, из которого будут извлекаться термины. Если используются контрастные подходы к извлечению терминов, необходим также корпус текстов общего характера. В зависимости от конкретных методов извлечения терминов, используемых далее, собранные корпусы подвергаются предварительной обработке, такой как лемматизация (приведение слов в нормальную форму), частеречная

разметка (морфологическая разметка слов), фрагментация или синтаксический анализ;

2. Выявление устойчивых лексических единиц (unithood detection) – идентификация лексических элементов, состоящих из нескольких слов, но относящихся к одной понятийной единице;

3. Определение терминологичности (termhood detection) – определение вероятности того, что извлеченное слово или устойчивая лексическая единица являются терминами;

4. Выявление терминологических вариантов (term variants detection) – идентификация различных лингвистических реализаций одного и того же понятия предметной области;

5. Оценка и валидация (evaluation and validation) – процедура оценки качества автоматического выделения терминов по сравнению с ручной работой эксперта предметной области.

В данной монографии представлен обзор существующих методов автоматического распознавания терминов и приведен практический пример, содержащий подробное описание каждого из 5 вышеперечисленных этапов распознавания. В целом, структура монографии выглядит следующим образом. В главе 1 приводятся общие сведения о проблематике автоматического распознавания терминов, даются вводные определения и обсуждаются способы операционализации такого сложного понятия как терминологичность. В главе 2 рассматриваются методы автоматического распознавания терминов, приводится их классификация. В главе 3 рассматривается практический пример автоматического извлечения терминов из учебника “Introduction to Information Retrieval”.

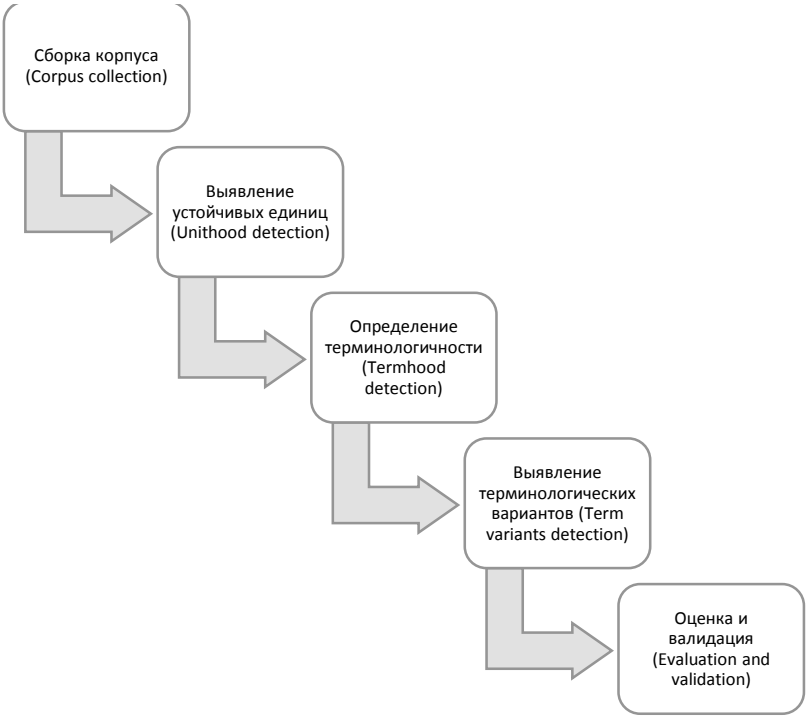


Рисунок 2 – Этапы автоматического распознавания терминов

1. ПОСТАНОВКА ПРОБЛЕМЫ АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ ТЕРМИНОВ

1.1. Понятие термина и его противопоставление со словом. Критерии терминологичности

Любая предметная область состоит из понятий, которые связаны между собой основополагающими для этой области отношениями. Эти понятия представлены в литературе по предметной области в виде терминов. Определение того, какие слова являются для данной предметной области терминами, а какие не являются, представляет сложную задачу, решение которой требует, в первую очередь, формализации требований, предъявляемых к тому, что можно считать терминами. Так, в работе [3] отмечается, что термины – это лингвистические представления понятий в конкретной предметной области, предназначенные для классификации предметных знаний. Другими словами, термин T можно определить как упорядоченную пару (c, t), где c – это понятие предметной области (единица знания), а t – его терминологическая форма (см. рисунок 3).

Таким образом, для того, чтобы ввести новый термин в предметную область, должно существовать понятие, указывающее на этот термин. Чтобы классифицировать термин, необходимо связать и сгруппировать понятие с другими понятиями в предметной области. Как отмечается в работе [3], термины сами по себе не являются единицами знаний, но они относятся к понятиям, которые являются единицами знаний, и их именование включает использование конкретных предметно-зависимых шаблонов. Например, в компьютерных науках новые термины часто образуются путем комбинирования существующих терминов, а в технических отраслях путем использования существующих слов в новых значениях.

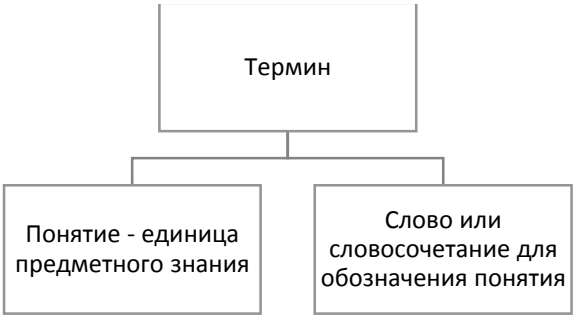


Рисунок 3 – Формальное представление термина

В таблице 1 представлено 3 основных шаблона, с помощью которых образуются новые термины.

Таблица 1

Шаблоны, использующиеся для создания новых терминов

№	Описание шаблона	Пример
1	Использование языковых ресурсов, которые изменяют смысл существующего слова посредством сравнений, метафор, ассоциаций и т.д. В результате появляются многозначные слова.	1) Память (в общеупотребительном значении) – способность сохранять и воспроизводить в сознании прежние впечатления, опыт, а также запас хранящихся в сознании впечатлений. 2) Память (в компьютерных науках) – физическое устройство или среда для хранения данных, используемых в течение определённого времени.
2	Изменение существующих слов за счет таких преобразований, как аффиксация, соединение, сокращение и т.д.	Информатика <- Информатика + Автоматизация. e-learning <- electronic + learning
3	Создание новых терминов за счет заимствований слов из других языков, прямого калькирования слов, придумывания новых слов и т.д.	Цифровизация <- Digitalization Хранилище <- Storage Кубит <- q-bit <- quantum bit

Как отмечается в работе [4], термин – это не просто обозначение научного понятия или предмета профессиональной деятельности, это обозначение, используемое в профессиональной среде. Если убрать включенность в профессиональную деятельность, то следом будут сняты предъявляемые к термину требования однозначности, четкости, точности и т.п. Таким образом, делает вывод автор работы [4], выход за пределы профессиональной деятельности можно считать первым действием по слиянию терминологического слова с лексикой общего языка. Соответственно, можно выделить первый критерий терминологичности – ограниченность области употребления (критерий «где служит»).

Ограниченность области употребления – это знак, где искать термин [4]. На этом критерии построены контрастные методы распознавания терминов, сравнивающие частоту употребления термина-кандидата в текстах предметной области (профессиональной тематики) с частотой употребления в текстах общего языка (см. рисунок 4).

Фактически, речь идет об операционализации (практическом измерении) терминологичности через оценку того, насколько ограниченна область употребления слова. Под операционализацией здесь понимается процесс преобразования теоретической идеи в эмпирическую форму.

Как мы видим, существует устойчивая тенденция к образованию оппозиции терминов со словами общего языка, т.е. противопоставление терминов и «обычных» слов общего языка [4]. По словам автора [4], важным условием опознания слова как термина является наличие в лексике его первого имени, а сам термин зачастую осознается как второе, особое имя предмета, уже имеющего первое имя. Соответственно, можно выделить второй критерий терминологичности – наличие эквивалента в общем языке. Оппозиция «термин – слово общего языка» позволяет разделять научное понятие (специальное, предметное, профессиональное знание) и бытовое понятие (наивное знание), сам предмет и понятия о предмете [4].

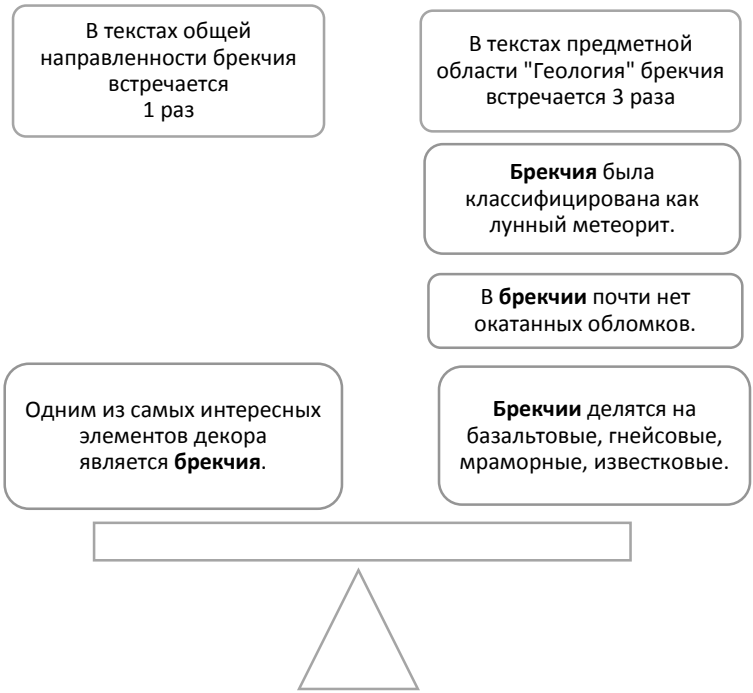


Рисунок 4 – Операционализация терминологичности через критерий ограниченности области применения

Разумеется, не все термины имеют прямые эквиваленты в общем языке, например, термину брекчия соответствует определение «горная порода обломочного происхождения», которое в силу его размера сложно назвать эквивалентом. Поэтому методы, основанные на операционализации терминологичности через поиск слов-эквивалентов, гораздо менее разработаны и менее популярны, чем контрастные методы.

1.2. Точность и полнота автоматического распознавания терминов

Две главные проблемы, связанные с автоматическим извлечением терминов – это ложные обнаружения (их также называют шумом – noise) и ложные пропуски (их также называют тишиной – silence). Как правило, под шумом понимаются высокочастотные слова и словосочетания, которые не являются терминами, но при этом часто употребляются в текстах предметной области, а под тишиной – низкочастотные слова и словосочетания, которые являются терминами, но при этом редко употребляются в текстах предметной области.

Как отмечалось выше, канонический процесс распознавания терминов завершается процессом валидации (проверки). Валидация осуществляется экспертами предметной области в ручном или полуавтоматическом режиме, и позволяет подтвердить или опровергнуть терминологический статус извлекаемых слов. По этой причине авторы работы [3] считают, что правильнее называть извлекаемые слова не терминами, а терминами-кандидатами, из которых в процессе валидации отбираются истинные термины предметной области.

Если изобразить множество прошедших валидацию терминов графически, то оно будет представлять область пересечения двух множеств: множества всех терминов-кандидатов и множества истинных терминов (см. рисунок 5). Шум – это разность между множеством терминов-кандидатов и валидных терминов, а тишина – это разность между множеством истинных терминов и валидных терминов.

Отсюда легко рассчитать полноту и точность извлечения терминов. Обозначим множество терминов-кандидатов через C_t, множество истинных терминов через T_t, множество валидных терминов, как уже отмечалось выше, это $C_t \cap T_t$. Тогда полнота ($Recall$) и точность ($Precision$) будут определяться следующим образом:

$$Recall = \frac{\|C_t \cap T_t\|}{\|T_t\|} \quad (1)$$

$$Precision = \frac{\|C_t \cap T_t\|}{\|C_t\|} \quad (2)$$

Когда множество терминов-кандидатов совпадает со множеством истинных терминов, то и точность, и полнота равны 100%, но такое имеет место крайне редко. Как правило, чем выше полнота, тем ниже точность, и наоборот. Поэтому для оценки качества автоматического распознавания терминов используют усредненный показатель между точностью и полнотой, называемый F-мерой. F-мера представляет собой среднее гармоническое точности и полноты, т.е. определяется следующим образом:

$$F1 = \frac{2Precision*Recall}{\text{Precision+Recall}} \quad (3)$$

Рассмотрим расчет точности и полноты распознавания терминов на следующем примере. Пусть из корпуса текстов по аналитической химии система извлекла, опираясь на частотный принцип, 11 терминов-кандидатов: *активность, гидролиз, кислота, количество, константа скорости, масса, моль, молярная масса, степень, уравнение скорости, электролит*. Из них эксперт посчитал валидными, т.е. действительно относящимися к понятийному аппарату аналитической химии, 9 терминов, исключив термины *количество* и *степень* как слова общей лексики. Полный глоссарий терминов по аналитической химии, построенный вручную, насчитывает 262 термина. Какова точность и полнота извлеченного системой списка терминов? Имеем: $\|C_t\| = 11$, $\|T_t\|$=262, $\|C_t \cap T_t\| = 9$. Тогда точность определяется как 9/11, что в процентном выражении составляет 81,8%, полнота определяется как 9/262, что в процентном выражении составляет 3,4%. Соответственно, F-мера равна 6,5%.

Рисунок 5 – Валидные термины, относящиеся к заданной предметной области

Современные методы распознавания терминов для русских текстов, как правило, обладают точностью и полнотой не выше 50%, хотя в работе [5] приводятся результаты экспериментов, в которых точность составила 40%, а полнота составила 68,6%.

Для повышения точности и полноты извлечения терминов необходимо учитывать возможное варьирование многословных терминов в тексте. Как отмечается в работе [6], «проблема варьирования изучается относительно недавно и заключается в том, что термины при употреблении в тексте могут изменяться по форме (например: *архитектура сети – сетевая архитектура*), обозначая тем не менее одно и то же специальное понятие».

1.3. Лингвистические подходы к распознаванию терминов

Наиболее интуитивно понятным, но, возможно, не самым простым для автоматизации является лингвистический подход к извлечению терминов. Как следует из его названия, он опирается на лингвистические, т.е. языковые признаки терминов и используется в основном для выделения сложных многословных терминов, называемых устойчивыми терминологическими сочетаниями. Как отмечается в работе [6], доля многословных терминов в терминосистеме любой предметной области выше, чем простых

(однословных), поэтому актуальна проблема автоматического выявления таких сложных терминов.

Лингвистический метод использует предположение о том, что терминологические сочетания в конкретном языке строятся в соответствии с определенными морфологическими и лексико-синтаксическими шаблонами [7,8]. Под морфологическими и лексико-синтаксическими шаблонами понимаются модели языковой конструкции, которые отображает ее морфологические, лексические и синтаксические свойства. Например, если заданное слово со своей именной группой употреблено в творительном падеже и предшествует слову *называется*, то скорее всего, это слово является термином (ср. *экспертной системой называется* …). Таким образом, шаблоном здесь является конструкция (Именная группа в тв.п.) + "*называется*" + (?).

В работе [8] рассматриваются 6 групп лексико-синтаксических шаблонов, позволяющих извлекать терминологические сочетания, несколько примеров таких шаблонов представлено в таблице 2. В работах [9,10] рассматриваются 9 видов морфологических шаблонов, используемых для поиска многословных терминов. Описания этих шаблонов представлены в таблице 3.

Таблица 2

Примеры лексико-синтаксических шаблонов для выявления терминов (на основе таблицы из [8])

№	Группа шаблонов	Примеры шаблонов	Примеры терминов и их употреблений
1	Морфо-синтаксические образцы терминов	(сущ.)	• брекчия
		(прил.) + (сущ.)	• обломочная порода
		(сущ.) + (сущ. в род. п.)	• разлом породы
2	Контексты определения авторских терминов	(?) + "будем" "называть" + (?)	• Такие породы *будем называть обломочными породами* -> обломочная порода
		"под" + (?) + "понимается" (?)	• *Под шельфом понимается* область, которая … -> шельф

		(?) + "это" "часть" +(?) (?) + "это" "раздел" +(?) (?) + "это" "свойство" +(?)	• *Подсистема – это часть системы, выделенная по какому-либо признаку* -> подсистема • *Магнитность – это свойство минерала …* -> магнитность
3	Соединения терминов	(?) + "," + (?) + {"и" | "или"} + (?)	• *шины адреса, данных и управления* -> шина адреса, шина данных, шина управления; • *серебро, золото или платина* -> серебро, золото, платина
		"как" + (?) + "," "так" "и" + (?)	• *как тонкий, так и толстый клиент* -> тонкий клиент, толстый клиент
		(?) + "," + (?) +{"," + (?) } "и" "другие" + (?)	• В этом случае образуются *гематит, самородная сера, борная кислота, реальгар, аурипигмент, киноварь и другие минералы* -> гематит, самородная сера …

Таблица 3

Морфологические шаблоны для извлечения многословных терминов русского языка

№	Морфологический шаблон	Пример
1	[сущ.+прил.(Р.п.)+сущ.(Р.п.)]	• технология безбумажного документооборота; • метод опорных векторов.
2	[прил.+прил.+сущ.]	• серверная операционная система; • интеллектуальная транспортная система.
3	[прил.+сущ.+сущ.(Р.п.)]	• автоматическое распознавание терминов; • виртуальная точка доступа.
4	[сущ.+сущ.(Р.п.)+сущ.(Р.п.)]	• сервер баз данных; • метод анализа иерархий.
5	[прил.+сущ.]	• вычислительная система; • информационный поиск.

6	[прич.+сущ.]	• плавающая точка; • удаленный доступ.
7	[сущ.+сущ.(Р.п.)]	• мера близости; • коэффициент сжатия.
8	[сущ.+сущ.(Т.п.)]	• сортировка пузырьком; • сортировка вставками.
9	[сущ.+"-"+сущ.]	• веб-браузер; • веб-сервис.

Анализ работ в области выделения терминологических сочетаний на основе лингвистических подходов демонстрирует, что применение шаблонов повышает точность и полноту автоматического распознавания терминов [10]. Разумеется, среди извлекаемых с помощью шаблонов словосочетаний будут встречаться как истинные термины, так и просто высокочастотные словосочетания (сравните, *обломочная порода* и *такая порода*, *шина данных* и *количество данных*). Дифференциация первых от вторых представляет достаточно сложную задачу, тем не менее, она решаема, если не на теоретическом, то хотя бы на прагматическом уровне [11].

1.4. Классификация подходов к распознаванию терминов

Лингвистические подходы к извлечению терминов часто противопоставляют статистическим, основанным на подсчете частот встречаемости лексических единиц в тексте. При этом предполагается, что ключевые слова и терминологические сочетания будут иметь более высокую частоту встречаемости в текстах предметной области, нежели общеупотребительные слова, разумеется, за исключением служебных слов. В работе [12] отмечается, что подобная бинарная классификация подходов к извлечению терминов является неполной, и приводится более сложная классификация на основе трех определяющих признаков (см. рисунок 6):

- по наличию системы обучения;
- по наличию используемых лингвистических ресурсов;

- по способу определения признаков терминологичности (операционализации).

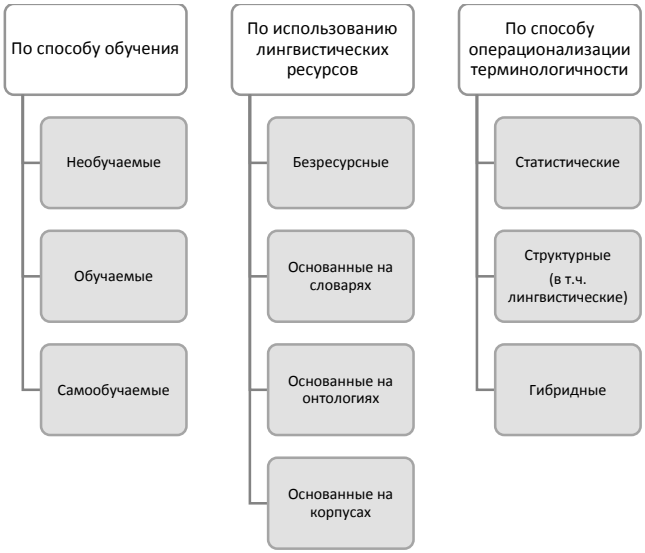


Рисунок 6 – Классификация подходов к извлечению терминов

Как отмечают авторы [12], по наличию системы обучения подходы делятся на три группы: необучаемые, обучаемые и самообучаемые. Обучаемые подходы опираются на технологии машинного обучения, использующих определенные наборы признаков для выделения терминов. Затем на основе этого набора признаков выполняется обучение на тестовых данных, содержащих уже идентифицированные термины. В данной группе подходов самой сложной задачей является определение признакового пространства. В работе [13] все признаки условно делятся на два типа: (i) признаки, которые получают статистические, лингвистические и гибридные знания из входного корпуса, такие как, например, мера TF-IDF и POS-разметка слова, и (ii) признаки, ко-

торые получают эти знания из метрик, которые используют знания из других корпусов помимо входного корпуса. Корпуса, которые принадлежат другой предметной области, отличающейся от входной области называются контрастными, а которые не принадлежат к какой-либо конкретной предметной области называются общими. В частности, из общего корпуса авторы берут такой признак как Freq GC – относительную частоту термина-кандидата в этом корпусе и делают предположение, что термин-кандидат не должен слишком часто встречаться в общем корпусе.

По наличию используемых лингвистических ресурсов подходы делятся на 4 группы: не использующие ресурсов, использующие словари, использующие онтологии, использующие корпуса текстов. Последние в свою очередь делятся на 2 подгруппы: использующие размеченные корпуса и использующие неразмеченные корпуса. Так в работе [14] в качестве лингвистического ресурса используется Википедия, а в качестве основного признака терминологичности используется признак «вероятность быть гиперссылкой в Википедии». По словам авторов, «значение этого признака будет близко к нулю для слов и словосочетаний, являющихся частью общей лексики, то есть не принадлежащих какой-либо предметной области».

По способу описания признаков терминологичности подходы делятся на 3 группы: опирающиеся на статистические признаки терминологичности, опирающие на структурные признаки терминологичности (на лингвистические шаблоны или графовые представления), и гибридные.

Если в основе лингвистических и статистических методов лежат достаточно интуитивные представления (в первом случае – о наличии лексико-синтаксических маркеров в окружении термина, во втором случае – о повышенной частоте употребления термина в текстах предметной области), то в основе графовых методов лежит математический аппарат теории графов. Графовые методы представляют текст как граф, вершинами которого являются слова или словосочетания, а ребрами отношения между ними. Отношения могут определяться различными способами: например, выражать совместную встречаемость в одном предло-

жении или окне текста заданного размера, семантическую близость. Среди всех вершин графа на основе какой-либо меры центральности вычисляются самые авторитетные вершины, они и отбираются как ключевые слова.

Одним из наиболее популярных вариантов графового подхода к извлечению терминов и ключевых слов является алгоритм TextRank [15]. Создатели алгоритма были вдохновлены формулой PageRank, используемой Гуглом для ранжирования веб-страниц. Пусть V_i – это текущая веб-страница, а $In(V_i)$ – это множество страниц V_j, которые ссылаются на нее. Тогда ранг (значимость) веб-страницы V_i определяется по следующей формуле:

$$PR(V_i) = (1 - \mathrm{d}) + d * \sum_{j \in In(V_i)} \frac{PR(V_j)}{|Out(V_j)|} \quad (4),$$

где d – это коэффициент затухания, означающий вероятность того, что пользователь зашедший на данную страницу перейдет по одной из выходящих ссылок, т.е. ссылок, содержащихся на этой странице (в классической формуле этот коэффициент равен 0.85); $|Out(V_j)|$ – это количество выходящих ссылок на каждой из рассматриваемых страниц V_j.

Для извлечения ключевых слов формула PageRank использовалась следующим образом:

1. Исходная коллекция текстов токенизировалась (делилась на лексические единицы, т.е. слова) и размечалась с помощью частеречной разметки (PoS).
2. Ко всем выделенным лексическим единицам применялся синтаксический фильтр (т.е. в качестве вершин были оставлены только существительные и глаголы).
3. Две вершины соединялись ребром, только если соответствующие им лексические единицы встречались в пределах окна из N слов. В результате получался невзвешенный неориентированный граф.
4. Для полученного графа запускался алгоритм PageRank, чтобы ранжировать слова по их значимости.
5. Выбирались только слова из верхней части списка ранжирования, т.е. с самым высоким рангом.

6. Смежные ключевые слова сворачивались в ключевое словосочетание.

1.5. Модельный пример извлечения ключевых слов и терминов

В данном разделе мы используем экосистему R (язык и прилагающиеся к нему библиотеки обработки естественного языка) для автоматического извлечения ключевых слов и терминов из культовой статьи Алана Тьюринга «Вычислительные машины и разум», написанную в далеком 1950-м году [16]. Статья Тьюринга, посвященная вопросу «могут ли машины мыслить», была загружена в русском переводе в виде корпуса, состоящего из 7 текстовых документов (файлов) по числу основных разделов статьи (см. таблицу 4).

Алан Тьюринг использовал в своей статье более 2000 уникальных слов, в том числе слово *машина* 297 раз. Это второе по частоте использования слово в статье, после предлога *в*. Следующее часто используемое слово в статье, если не принимать во внимание служебные слова, такие как предлоги и некоторые связующие глаголы, это слово *вопрос*, которое употреблялось автором 85 раз, за ним следует слово *человек*, которое употреблялось 83 раза. Все три указанных слова *машина*, *вопрос* и *человек* можно назвать ключевыми для данной статьи, но только одно из них условно можно считать термином, это слово *машина*. Таким образом, на наглядном примере можно увидеть разницу между ключевыми словами и терминами. Ключевые слова отражают содержание документа, в то время как термины – содержание предметной области.

Таблица 4

Состав корпуса «Вычислительные машины и разум»

№	Документ	Заголовок	Количество токенов, включая слова и знаки препинания
1	01.txt	I. Игра в имитацию	569
2	02.txt	II. Критика новой постановки проблемы	614
3	03.txt	III. Машины, привлекаемые к игре	496
4	04.txt	IV. Цифровые вычислительные машины	1487
5	05.txt	V. Универсальность цифровых вычислительных машин	1240
6	06.txt	VI. Противоположные точки зрения по основному вопросу	6800
7	07.txt	VII. Обучающиеся машины	3165

Для обработки представленного корпуса мы использовали библиотеку UDPipe Natural Language Processing [17], представляющую собой разработку ученых и специалистов в области компьютерной лингвистики из Карлов Университета, Чехия. Как отмечают сами разработчики, автоматическая обработка больших текстов на естественном языке часто представляет повторяющиеся задачи для многих языков: даже для самых продвинутых задач, тексты сначала должны пройти предварительные этапы обработки: от токенизации до парсинга. В ответ на это, авторы представили очень простой в использовании инструмент, состоящий из одного бинарного файла и одной модели (для каждого языка), который решает указанные задачи предобработки без необходимости использования каких-либо внешних данных.

Таким образом, UDPipe – это конвейер, который способен выполнить 1) токенизацию; 2) морфологический анализ и лемматизацию; 3) частеречную разметку (POS-tagging); и 4) анализ зависимостей для большого набора языков, в том числе, английского, немецкого, французского, чешского, китайского, русского, турецкого, хинди, казахского и даже языка африкаанс.

Кроме того, конвейер легко обучается на тренировочных данных в формате CoNLL-U (в некоторых случаях также на необработанных корпусах) и требует минимальных лингвистических знаний со стороны пользователей. UDPipe доступен как библиотека для C ++, Python, Perl, Java, C #, а также как веб-сервис.

Также существует библиотека udpipe для R, разработанная сторонними разработчиками, которую мы и будем использовать в данном модельном примере.

Шаг 1. Подключение библиотеки и загрузка модели для определенного языка. Модель предварительно должна быть скачана в рабочую директорию со страницы http://ufal.mff.cuni.cz/udpipe#language_models

```
library(udpipe)
library(utf8)
library(readtext)
dl.rus <- udpipe_load_model (file="russian-syntagrus-ud-2.3-181115.udpipe")
```

Можно использовать другой вариант загружать файл напрямую с помощью команды. Тогда будет использоваться стандартная модель для русского языка, но она ограниченная.

```
dl.file <- udpipe_download_model (language="russian")
dl.rus <- udpipe_load_model(file = dl.file$file_model)
```

Шаг 2. Загрузка и аннотация с помощью модели корпуса текстов. Файлы, образующие корпус, обязательно должны быть в кодировке UTF8. Если они в другой кодировке, предварительно нужно провести конвертацию и только затем выполнить аннотацию. Аннотированный набор данных представляет собой таблицу, строки которой соответствуют выделенным токенам, а столбцы – признакам выделенных токенов, например, морфологическим: часть речи, падеж, число, род и т.д. (см. рисунок 7).

```
x<-readtext(paste0(getwd(), "/turing_rus/*.txt"),
    docvarsfrom = "filenames")
x$text <- enc2utf8(x$text)
y <- udpipe(x,object=dl.rus)
```

В нашем примере аннотированный корпус состоит из 14371 токенов (слова, знаки препинания и другие обособленные единицы текста), каждый из которых описывается 18 атрибутами (признаками).

y

Filter

token_id	token	lemma	upos	xpos	feats	head_token_id	dep_rel
1	Я	я	PRON	NA	Case=Nom|Number=Sing|Person=1	2	nsubj
2	собираюсь	собираться	VERB	NA	Aspect=Imp|Mood=Ind|Number=Sing|Person=1|Tense=...	0	root
3	рассмотреть	рассмотреть	VERB	NA	Aspect=Perf|VerbForm=Inf|Voice=Act	2	xcomp
4	вопрос	вопрос	NOUN	NA	Animacy=Inan|Case=Acc|Gender=Masc|Number=Sing	9	obj
5	:	:	PUNCT	NA	NA	4	punct
6	могут	мочь	VERB	NA	Aspect=Imp|Mood=Ind|Number=Plur|Person=3|Tense=P...	3	ccomp
7	ли	ли	PART	NA	NA	6	advmod
8	машины	машина	NOUN	NA	Animacy=Inan|Case=Nom|Gender=Fem|Number=Plur	6	nsubj
9	мыслить	мыслить	VERB	NA	Aspect=Imp|VerbForm=Inf|Voice=Act	6	xcomp
10	.	.	PUNCT	NA	NA	2	punct
1	Но	но	CCONJ	NA	NA	4	cc
2	для	для	ADP	NA	NA	3	case
3	этого	это	PRON	NA	Animacy=Inan|Case=Gen|Gender=Neut|Number=Sing	4	obl
4	нужно	нужный	ADJ	NA	Degree=Pos|Gender=Neut|Number=Sing|Variant=Short	0	root

Рисунок 7 – Аннотированный корпус текстов

Шаг 3. Просмотр корпусной статистики. Библиотека udpipe позволяет просматривать статистику аннотированного корпуса, например, количество существительных, прилагательных в корпусе, частоту существительных и т.д. (см. рисунки 8-10). Несмотря на то, что эта задача не относится напрямую к извлечению терминов, такая статистика бывает полезна для языковых исследований и выявления авторского стиля и жанра произведения.

```
stats <- txt_freq(y$upos)
stats$key <- factor(stats$key, levels = rev(stats$key))
barchart(key ~ freq, data = stats, col = "cadetblue",
   main = "UPOS (Universal Parts of Speech)\n frequency of occurrence",
xlab = "Freq")

stats <- subset(y, upos %in% c("NOUN"))
stats <- txt_freq(stats$lemma)
stats$key <- factor(stats$key, levels = rev(stats$key))
```

```
barchart(key ~ freq, data = head(stats, 20), col = "cadetblue",
    main = "Most occurring NOUNS", xlab = "Freq")

stats <- subset(y, upos %in% c("VERB"))
stats <- txt_freq(stats$lemma)
stats$key <- factor(stats$key, levels = rev(stats$key))
barchart(key ~ freq, data = head(stats, 20), col = "cadetblue",
    main = "Most occurring VERBS", xlab = "Freq")
```

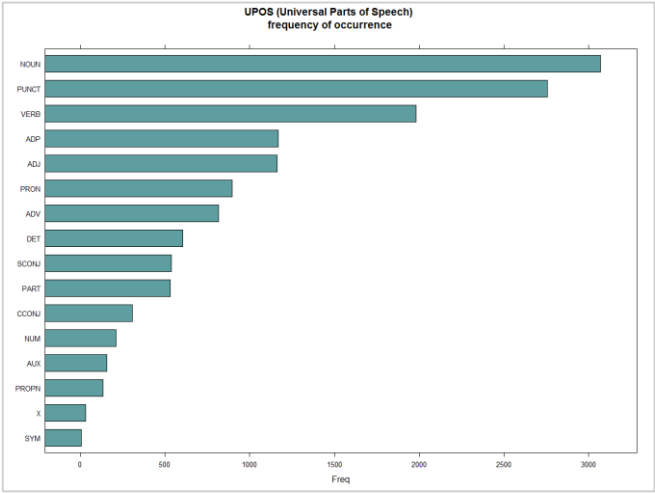


Рисунок 8 – Корпусная статистика в разрезе частей речи

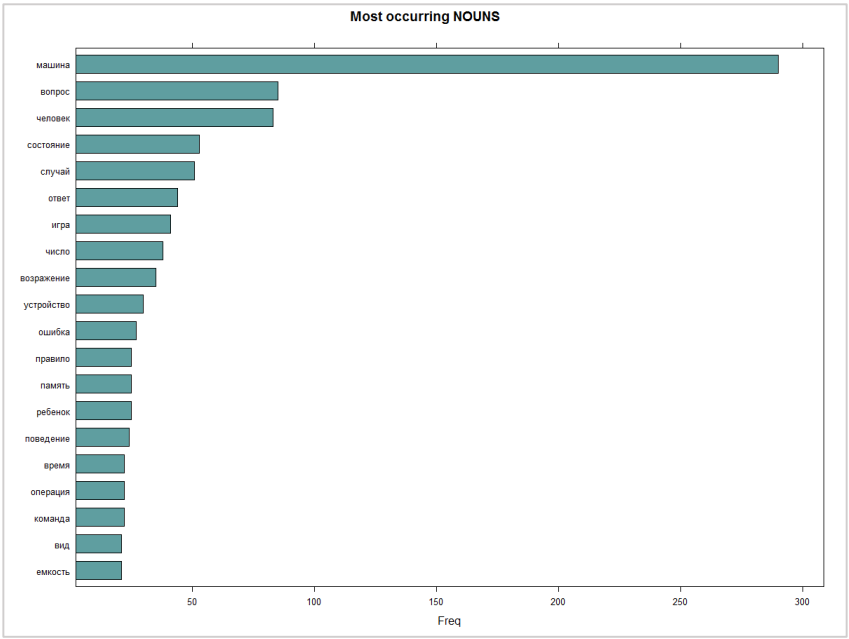


Рисунок 9 – Корпусная статистика в разрезе существительных

Как уже отмечалось выше, на самом деле, большая часть терминов, вероятно, будет представлена в виде сложных, т.е. многословных устойчивых выражений. Извлечь выражения, образованные из нескольких слов, можно либо выделяя слова, следующие друг за другом, либо выделяя высокую совместную встречаемость слов в одном предложении или в одном окне. Оба этих подход могут быть реализованы с использованием библиотеки udpipe. При этом, к искомым выражениям можно дополнительно применить морфологический шаблон, в котором указать, что будут отобраны существительные и прилагательные.

```
stats <- keywords_collocation(x = y, term = "token", group = c("doc_id",
"paragraph_id", "sentence_id"), ngram_max = 4)
## Co-occurrences in the same sentence, only nouns or adjectives
stats <- cooccurrence(x = subset(y, upos %in% c("NOUN", "ADJ")),
```

```
 term = "lemma", group = c("doc_id", "paragraph_id", "sentence_id"))

## How frequent do words follow one another
stats <- cooccurrence(x = y$lemma,
        relevant = y$upos %in% c("NOUN", "ADJ"))

## How frequent do words follow one another even if skip 2 words
stats <- cooccurrence(x = y$lemma,
        relevant = y$upos %in% c("NOUN", "ADJ"), skipgram = 2)
head(stats)
```

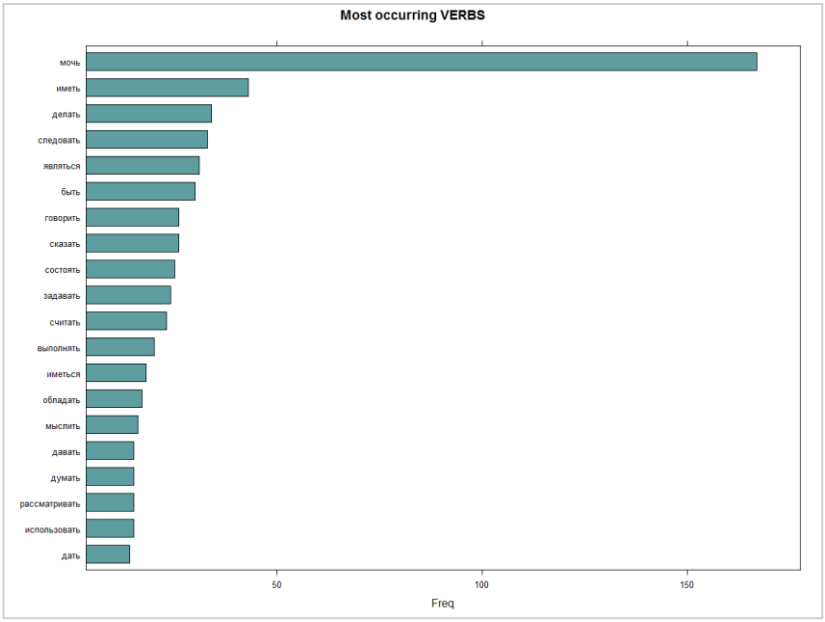


Рисунок 10 – Корпусная статистика в разрезе глаголов

Устойчивые выражения, т.е. коллокации, можно легко визуализировать в виде графа (см. рисунок 11).

```
library(igraph)
library(ggraph)
library(ggplot2)
wordnetwork <- head(stats, 10)
wordnetwork <- graph_from_data_frame(wordnetwork)
ggraph(wordnetwork, layout = "fr") +
 geom_edge_link(aes(width = cooc, edge_alpha = cooc), edge_colour = "red") +
 geom_node_text(aes(label = name), col = "darkgreen", size = 4) +
 theme_graph(base_family = "Arial Narrow") +
 theme(legend.position = "none") +
 labs(title = "Cooccurrences within 3 words distance", subtitle = "Nouns & Adjective")
```

Можно также использовать встроенный метод RAKE – один из самых простых и быстрых способов извлечения ключевых слов (см. рисунок 12).

```
stats <- keywords_rake(x = y, term = "lemma", group = "doc_id",
          relevant = y$upos %in% c("NOUN", "ADJ"))
stats$key <- factor(stats$keyword, levels = rev(stats$keyword))
barchart(key ~ rake, data = head(subset(stats, freq > 3), 20), col = "red",
    main = "Keywords identified by RAKE", xlab = "Rake")
```

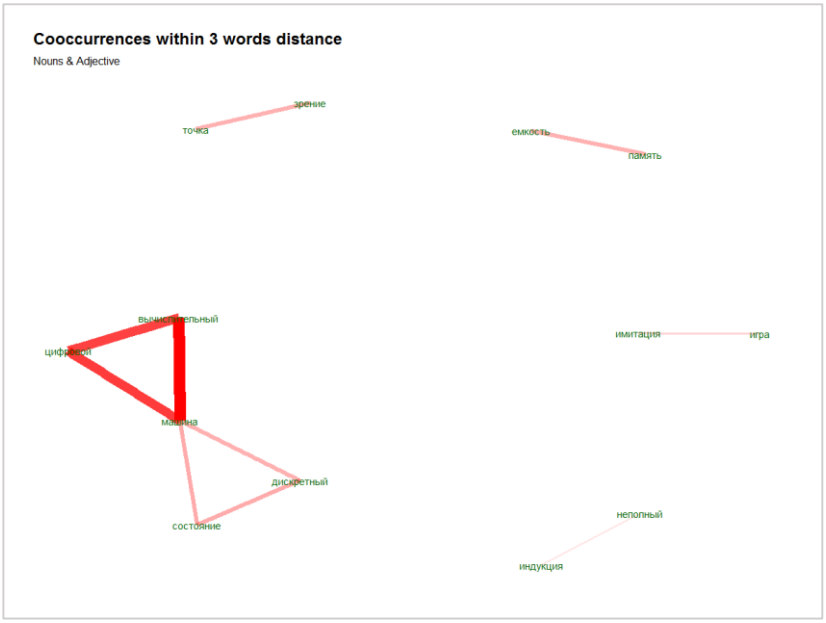


Рисунок 11 – Граф устойчивых выражений

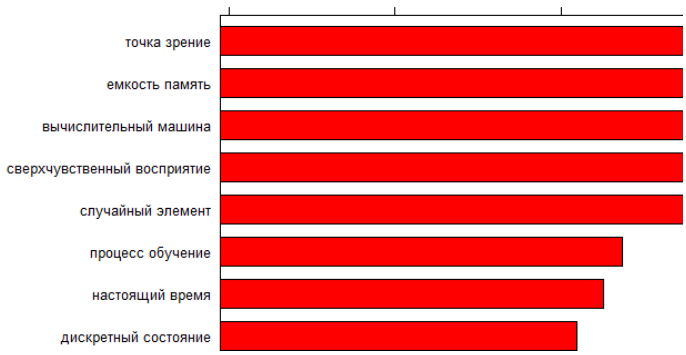


Рисунок 12 – Ключевые слова и устойчивые выражения, извлеченные методом RAKE

Более сложные коллокации можно извлечь с помощью паттернов регулярных выражений (см. рисунок 13). Например, с помощью паттерна "(A|N)*N(P+D*(A|N)*N)*" можно извлечь такие выражения как *игра в имитацию* или *машина с дискретным состоянием*. Тем не менее, и здесь мы видим, что по-прежнему высока доля общеупотребительных выражений, например, таких как *точка зрения*.

```
y$phrase_tag <- as_phrasemachine(y$upos, type = "upos")
stats <- keywords_phrases(x = y$phrase_tag, term = tolower(y$token),
          pattern = "(A|N)*N(P+D*(A|N)*N)*",
          is_regex = TRUE, detailed = FALSE)
stats <- subset(stats, ngram > 1 & freq > 2)

stats$key <- factor(stats$keyword, levels = rev(stats$keyword))
barchart(key ~ freq, data = head(stats, 20), col = "magenta",
   main = "Keywords - simple noun phrases", xlab = "Frequency")
```

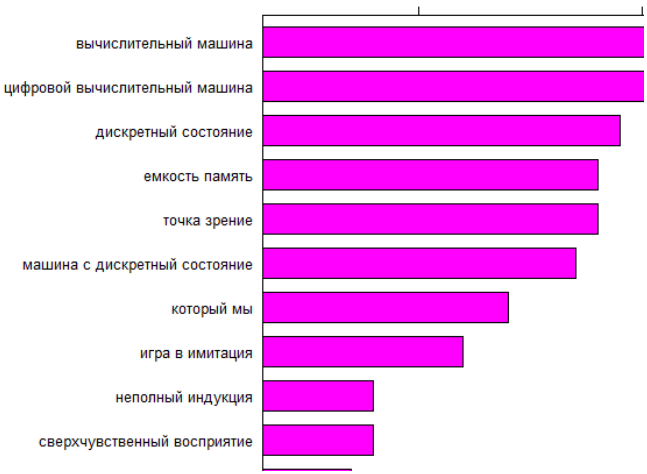


Рисунок 13 – Ключевые слова и устойчивые выражения, извлеченные с помощью паттернов регулярных выражений

И наконец, извлечение терминов с помощью рассмотренного в предыдущем разделе метода TextRank может быть реализовано с помощью одноименной библиотеки и визуализировано с помощью библиотеки wordcloud (см. рисунок 15).

```
library(textrank)
stats <- textrank_keywords(y$lemma,
          relevant = y$upos %in% c("NOUN", "ADJ"),
          ngram_max = 8, sep = " ")
stats <- subset(stats$keywords, ngram > 1 & freq >= 3)

library(wordcloud)
wordcloud(words = stats$keyword, freq = stats$freq, colors = brewer.pal(5,"Pastel1"))
```

Как мы видим, задача извлечения ключевых слов только с помощью использования лингвистических и статистических подходов решается достаточно успешно, чего нельзя сказать о задаче извлечения терминов. Термины – как словесные выражения понятий предметной области – должны рассматриваться именно в контексте предметной области и соответственно, для их извлечения нужны внешние источники знаний, например, контрастные, сравнивая с которыми можно утверждать, данный термин принадлежит только данной предметной области и не встречается в контрастных корпусах. Именно контрастным методам извлечения терминов будет посвящена следующая глава монографии.

Рисунок 14 – Ключевые слова и устойчивые выражения, извлеченные с помощью метода TextRank

2. МЕТОДЫ АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ ТЕРМИНОВ

2.1. Контрастные методы автоматического распознавания терминов

Контрастный подход – это общее название для методов, выявляющих термины с позиции их разной встречаемости внутри и за пределами предметной области [18,19,20]. Все эти методы объединяет общая идея об определении терминологичности слов на основе сравнения их распределений в двух коллекциях: целевой (предметной) и альтернативной. В качестве альтернативной коллекции может использоваться либо контрастная коллекция, т.е. сформированная из текстов другой предметной области, либо общая коллекция, т.е. сформированная из текстов, не относящихся ни к какой предметной области [21].

В числе одной из первых работ, посвященных контрастному извлечению терминов, можно назвать [22]. Ее авторы для оценки терминологичности вводят в обиход новую, интуитивно понятную меру, получившую название странность (англ. Weirdness). Странность вычисляется для каждого термина-кандидата и представляет собой отношение частоты его употребления в целевой коллекции к частоте употребления в общей коллекции. Поскольку коллекции чаще всего не сбалансированы по размеру, то используются относительные частоты.

Для обычных слов формула странности возвращает значения близкие к 1, а для терминов – значения, намного превышающие 1, т.к. в этом случае знаменатель формулы близок к 0:

$$Weirdness = \frac{F_{SL}/N_{SL}}{F_{GL}/N_{GL}} = \frac{F_{SL} \cdot N_{GL}}{F_{GL} \cdot N_{SL}} \quad (5),$$

где F_{SL}, F_{GL} – это частоты употреблений слова в целевой (SL) и общей (GL) коллекциях соответственно; N_{SL}, N_{GL} – это количества всех слов в целевой и общей коллекциях соответственно.

В своих более поздних работах, например, в [23], авторы представляют модифицированный вариант формулы (1), т.к. исходная формула, по их словам, проявляет сингулярность, когда знаменатель обращается в 0. Это происходит в тех случаях, когда частота употребления слова в общей коллекции равна 0, что в результате приводит к бесконечности. Модифицированная, сглаженная формула странности отличается от исходной прибавлением 1 к частоте употребления слова в общей коллекции текстов:

$$Weirdness = \frac{F_{SL} \cdot N_{GL}}{(1+F_{GL}) \cdot N_{SL}} \quad (6)$$

Дальнейшее сравнение работ [22] и [23], показывает, что в первой работе авторы формируют список терминов только на основе высоких значений странности, а во второй они уже используют комбинацию высокой странности с высокой частотой. Тем самым они пытаются избавиться от «странных» слов, случайно оказавшихся в целевой коллекции, т.е. не относящихся к предметной области. Такой подход гарантирует более высокую точность охвата терминов, но при этом, как мы уже отмечали, страдает полнота, т.к. редкие термины выпадают из рассмотрения.

В [24] идея контрастной оценки терминов формируется в виде не одного, а двух утверждений. Во-первых, реже употребляемые в целевой коллекции слова должны иметь более низкую оценку. Во-вторых, чаще употребляемые в целевой коллекции слова должны имеют более высокую оценку, но с оговоркой, что они не встречаются часто в контрастной коллекции или в ограниченном наборе текстов целевой коллекции. Авторы операционализируют эти утверждения в виде метрики, которую называют релевантностью (relevance):

$$Relevance = \frac{1}{log_2\left(2+\frac{f_{SL} \cdot N_{SL}^t}{f_{GL}}\right)} \quad (7),$$

где f_{SL} и f_{GL} – это относительные частоты употреблений слова t в целевой и контрастной коллекциях соответственно; N_{SL}^{t} – относительное число текстов целевой коллекции, в которых встречается данное слово. Приведенная метрика хорошо справляется с извлечением репрезентативных терминов, но искусственно занижает оценку редких терминов, что, как мы уже отмечали, негативно влияет на полноту охвата терминов.

В [25] предлагается несколько иной способ оценки терминологичности на базе контрастного подхода. Способ берет за основу известную формулу взвешивания слов TF-IDF, согласно которой вес слова в документе тем выше, чем выше частота его использования в этом документе и чем ниже его разброс по всей коллекции. В новом варианте формулы, который авторы называют «term frequency – inverse domain frequency», оценивается вес слова не в документе, а в целевой коллекции. Согласно новой формуле вес слова тем выше, чем выше относительная частота его использования в целевой коллекции и чем ниже его относительный разброс по всем коллекциям:

$$TF \cdot IDF = TF(t, D) \cdot IDF(t) = \frac{n_{t,D}}{\sum_k n_{k,D}} \cdot log\left(\frac{|TS|}{|\{d: t \in d\}|}\right) \qquad (8),$$

где $n_{t,D}$ – это число вхождений слова t в целевую коллекцию D, $\sum_k n_{k,D}$ – это сумма вхождений всех слов в целевую коллекцию D, $|TS|$ – это количество документов во всех используемых коллекциях, $|\{d: t \in d\}|$ – это количество всех документов, в которые слово t входит хотя бы один раз. Таким образом, авторы считают терминами все слова с высокой концентрацией в пределах узкого подмножества документов. Для определенной части терминов это, безусловно, справедливый подход, но для редких терминов он малопригоден.

Авторы [26] также предлагают оценивать терминологичность слов на базе формулы TF-IDF. Собственный вариант этой формулы они называют контрастным весом (contrastive weight) и опре-

деляют его как меру, которая тем выше, чем выше частота употребления слова в целевой коллекции и чем ниже относительная частота его употребления в контрастных коллекциях:

$$Contrastive\ Weight = TF(t, D) \cdot IDF(t) = \log\left(f_t{}^D\right) \cdot \log\left(\frac{F_{TC}}{\sum_j f_t^j}\right) \qquad (9),$$

где $f_t{}^D$ – частота употребления слова в целевой коллекции, $\sum_j f_t^j$ – сумма частот всех употреблений слова в контрастных коллекциях, $F_{TC} = \sum_{i,j} f_i^j$ – сумма частот употреблений всех слов во всех коллекциях, включая целевую. Как отмечают сами авторы, контрастный вес значительно лучше оценивает терминологичность слов, чем чистые частоты, однако общая эффективность метода, определенная с помощью F-меры, по их словам, не бросается в глаза.

В [27] формула контрастного веса подвергается критической оценке. Как отмечают авторы указанной работы, контрастный вес и подобные ему метрики на самом деле оценивают не принадлежность терминов предметной области, а их распространенность. Чтобы исправить указанный недостаток, авторы предлагают оценивать терминологичность на базе двух показателей: меры преобладания слова в целевой коллекции DP (англ. domain prevalence) и меры тяготения слова к целевой коллекции DT (англ. domain tendency). Высокое значение DP указывает на преобладание слова в целевой коллекции по сравнению с другими словами. Высокое значение DT указывает на преобладание слова в целевой коллекции по сравнению с контрастной коллекцией. Формула для расчета DP по сути является сглаженным вариантом формулы контрастного веса (9):

$$DP(t) = log_{10}(f_t{}^D + 10) \cdot log_{10}\left(\frac{F_{TC}}{f_t{}^D + f_t{}^{\bar{D}}} + 10\right) \qquad (10),$$

где $f_t{}^D$ и $f_t{}^{\bar{D}}$ – частоты употреблений данного слова в целевой и контрастной коллекциях соответственно, $F_{TC} = \sum_j f_j^D + \sum_j f_j{}^{\bar{D}}$ –

суммы частот употреблений всех терминов-кандидатов в целевой и контрастной коллекциях соответственно.

Формула для расчета DT является сглаженным вариантом формулы странности (5), т.е. штрафует слова, которые часто встречаются в контрастной коллекции:

$$DT(t) = log_2\left(\frac{f_t^{D}+1}{f_t^{\bar{D}}+1} + 1\right) \quad (11)$$

Меры DP и DT объединяются в один общий показатель, названный дискриминационным весом DW (discriminative weight). По мнению авторов, этот показатель обладает высокой дифференцирующей способностью:

$$DW(t) = DP(t) \cdot DT(t) \quad (12)$$

Следует отметить, что показатели DT и DP сильно коррелируют друг с другом. Например, в наших экспериментах значения корреляции этих показателей составили от 0,71 до 0,82. Чтобы понять природу корреляции, мы разделили все термины-кандидаты на 4 непересекающиеся группы в зависимости от значений DT и DP: 1) значения DT и DP ниже среднего; 2) значение DT ниже среднего, а значение DP не ниже среднего; 3) значения DT не ниже среднего, а значения DP ниже среднего; 4) значения DT и DP не ниже среднего. И экспертные оценки, и оценки на основе формулы (8) показали один и тот же результат: терминами за небольшим исключением могут быть признаны только кандидаты групп 3 и 4, что соответствует высоким значениям показателя DT (см. рисунок 15). Данный результат свидетельствует о высокой информативности показателя DT и об избыточности показателя DP.

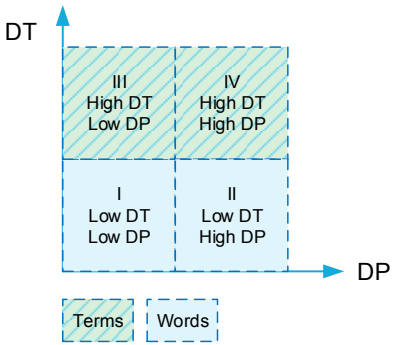


Рисунок 15 – Сравнение критериев DP и DT

При этом оговорка «за небольшим исключением» не случайна. Валидация предложенного способа оценки терминов демонстрирует, что в зоне низких значений терминологичности среди обычных слов попадаются в виде небольшого исключения термины, которые мы называем редкими. Это те термины, которые встречаются 1-2 раза в целевой коллекции и ни разу не встречаются в контрастной коллекции. Извлечение таких терминов требует применения более тонких инструментов дифференциации.

Использование сразу нескольких показателей для оценки терминологичности отличает не только работу [27]. В [28] для этой цели используются сразу 3 показателя: мера пертинентности DR (англ. domain pertinence), мера согласованности DC (англ. Domain consensus) и лексическая когезия LC (англ. Lexical cohesion), предназначенная для оценки когезии многословных терминов.

В результате итоговая оценка терминологичности слова t в целевой коллекции Di складывается из линейной комбинации трех перечисленных мер:

$$w(t, Di) = \alpha \cdot DR + \beta \cdot DC + \gamma \cdot LC \qquad (13),$$

где α, β, γ – это калибровочные параметры, по умолчанию $\alpha = \beta = \gamma = 1/3$.

Мера пертинентности DR представляет собой меру странности, обобщенную на случай множества контрастных коллекций.

Она определяется как отношение частоты слова в целевой коллекции к его наибольшей частоте во всех существующих контрастных коллекциях:

$$DR(t, Di) = \frac{freq(t,Di)}{\underbrace{\max}_{j}(freq(t,Dj)} \quad (14)$$

Мера согласованности DC позволяет учитывать распределение слов в отдельных документах. Она определяется через нормированные частоты ϕ_k встречаемости слова t в документах целевой коллекции Di, и тем выше, чем равномерней распределено слово в этих документах:

$$DC(t, Di) = -\sum_{d_k \in Di} \phi_k log\, \phi_k \quad (15)$$

Вводя меру согласованности, авторы обосновывают ее значимость тем, что термины, которые часто встречаются в большом количестве документов целевой коллекции, должны оцениваться выше, чем термины, которые часто встречаются в ограниченном количестве документов. Данное утверждение является полностью антагонистическим эвристике, использованной в [18]. Этот интересный факт, иллюстрирующий существующую противоречивость в выборе критериев терминологичности, отмечают также авторы [29].

В [30] терминологичность определяется не через отношение частот употребления слов в целевой и контрастной коллекциях, а через разность их рангов. Под рангом слова в коллекции понимается его позиция в списке, составленном из всех слов коллекции и отсортированном по возрастанию частот их употреблений. Слова, которые отсутствуют в данной коллекции, имеют ранг 0. Ранг 1 соответствует самому редкому слову коллекции.

Таким образом, индекс терминологичности слова t, выраженный через разность его рангов в целевой D и контрастной G коллекциях, имеет вид:

$$thd(t, D) = \frac{rank(t,D)}{|V(D)|} - \frac{rank(t,G)}{|V(G)|} \quad (16),$$

где $V(D)$ и $V(G)$ – это словари соответствующих коллекций. Индекс принимает значения от -1 до 1, значение 1 соответствует случаю, когда слово имеет высший ранг в целевой коллекции и нулевой ранг в контрастной. Ранжирование слов по убыванию индекса позволяет вывести наверх наиболее репрезентативные слова целевой коллекции, среди которых имеются и термины. Однако, как отмечают авторы, предложенное ранжирование не позволяет выделить только термины.

Последняя работа, которую мы хотим отметить в этом ряду, это [31]. Она также развивает идею штрафов и вознаграждений, заложенную в базовой конструкции формулы TF-IDF, и предлагает новый вариант этой формулы, получивший название “term frequency – disjoint corpora frequency”. В качестве вознаграждения используется абсолютная частота употребления слова в целевой коллекции, а в качестве штрафа – произведение абсолютных частот употреблений слова в контрастных коллекциях:

$$TF \cdot DCF = \frac{{f_t}^D}{\prod_{g \in G} 1 + log(1 + {f_t}^g)} \quad (17),$$

где ${f_t}^D$ и ${f_t}^g$ – частоты употреблений данного слова t в целевой и контрастной коллекциях соответственно, G – множество всех контрастных коллекций.

Авторы доказывают на серии экспериментов, что их формула является лучшей по точности извлечения терминов по сравнению с оценками, предложенными в [25,30] и ряде других работ. Они оправдывают использование произведения в знаменателе формулы тем фактом, что штраф должен расти в геометрической прогрессии за каждое употребление слова в очередной контрастной коллекции. По мнению авторов, терминологичность слов, которые употребляются небольшое число раз в большом количестве контрастных коллекций, должна оцениваться ниже, чем слов, которые используются много раз, но в небольшом количестве контрастных коллекций. В случае, когда имеется одна контрастная коллекция, результаты формулы смещаются в сторону высокочастотных терминов.

Таким образом, в данном обзоре мы рассмотрели 8 наиболее интересных контрастных методов операционализации оценки терминологичности (см. таблицу 5).

Таблица 5

Наиболее значимые способы операционализации контрастного подхода к извлечению терминов

№	Авторы и ссылка на источник	Название метрики или индикатора	Год
1	Ahmad et al [22, 23]	Weirdness	1999, 2005
2	Peñas A. et al [24]	Relevance	2001
3	Kim et al [25]	Term frequency- inverse domain frequency	2009
4	Basili et al [26]	Contrastive weight	2001
5	Wong et al [27]	Domain prevalence, Domain tendency	2007
6	Sclano et al [28]	Domain pertinence, Domain consensus	2007
7	Kit C., Liu X [30]	Termhood index	2008
8	Lopes et al [31]	Term frequency- disjoint corpora frequency	2016

Все эти методы являются эвристическими, т.е. основанными на предположениях относительно характера распределения терминов в целевой и контрастной коллекциях. Сравнительный анализ этих утверждений показывает, что имеют место как совпадения позиций разных авторов, так и серьезные расхождения, что свидетельствует о наличии нерешенных проблем в данной области.

2.2. Критерии хи-квадрат, информационная выгода и взаимная информация

В данном разделе мы рассмотрим 3 популярных контрастных критерия: меру взаимной информации, информационную выгоду и критерий хи-квадрат [32-34]. Прежде чем описывать эти критерии, введем общие для их вычисления обозначения, как показано в таблице 6.

Таблица 6

Обозначения, используемые для критериев отбора признаков

Символ	Расшифровка
TS	Все множество документов
A	количество документов позитивного множества, содержащих термин *t*
B	количество документов позитивного множества, **не** содержащих термин *t*
C	количество документов негативного множества, содержащих термин *t*
D	количество документов негативного множества, **не** содержащих термин *t*

В статистике взаимная информация представляет собой функцию двух случайных величин, которая предназначена для описания количества информации, содержащейся в одной случайной величине относительно другой. В нашем случае меру взаимной информации можно интерпретировать как количество информации, которое вносит присутствие термина в данной категории для правильной классификации документов.

$$MI(t,c) = \log_2 \frac{A*|TS|}{(A+C)*(A+B)} \qquad (18)$$

Мера взаимной информации достигает своего максимума, когда термин содержится только в документах предметной области и не содержится в других документах. Из формулы (18) видно, что данный критерий в большей степени предназначен для отбора редких, специфичных для предметной области терминов. Если у двух терминов количество содержащих их документов, относящихся к предметной области, одинаково (A), то большее значение меры имеет тот термин, у которого при ее расчете меньше знаменатель, т.е. меньше общее количество документов (*A+B)*, в которых он встречается.

Информационная выгода, также как мера взаимной информации, представляет собой количество информации, которое несет

наличие термина t о принадлежности документа, содержащего этот термин, предметной области.

$$IG(t,c) = \frac{A}{|TS|} * \log_2\left(\frac{A*|TS|}{(A+C)*(A+B)}\right) + \frac{C}{|TS|} * \log_2\left(\frac{C*|TS|}{(D+C)*(A+C)}\right) + \\ + \frac{B}{|TS|} * \log_2\left(\frac{B*|TS|}{(A+B)*(D+B)}\right) + \frac{D}{|TS|} * \log_2\left(\frac{D*|TS|}{(D+C)*(D+B)}\right) \quad (19)$$

Так же как мера взаимной информации, информационная выгода тем выше, чем меньше документов других категорий, в которых встречается данный термин. Однако в отличие от меры взаимной информации, информационная выгода симметрична, поскольку выдает одинаково высокие значения как для терминов «сильно» характерных для данной предметной области, так и для терминов «сильно» характерных для прочих областей. Благодаря этому свойству критерий информационной выгоды хорошо справляется с очисткой от «стоп-слов», служебных слов, которые равномерно распределены по всем предметным областям.

Критерий Хи-квадрат – это наиболее часто употребляемый в статистике критерий для проверки гипотезы о согласии модели и данных. Для нашей задачи его формула имеет следующий вид:

$$CHI(t,c) = \frac{|TS|*(A*D-C*B)^2}{(A+C)*(B+D)*(A+B)*(C+D)} \quad (20)$$

Критерий достигает максимума, равного количеству документов выборки, если термин входит только в документы позитивного множества, и минимума, равного 0, если термин и категория независимы.

Каждый из приведенных критериев извлечения ключевых слов имеет свои преимущества и недостатки. Все три критерия позволяют избавиться от слов общеупотребительной лексики, в том числе стоп-слов. При этом мера взаимной информации позволяет отобрать достаточно редкие, специфичные для предметной области термины. В отличие от нее информационная выгода и критерий хи-квадрат позволяют отобрать термины, часто употребляемые в документах предметной области («яркие» концепты). Оба критерия симметричны, они извлекают яркие термины как позитивного, так и негативного множеств, поэтому требуется

дополнительная проверка, принадлежит ли термин к интересующей эксперта предметной области. Очевидный способ проверки заключается в том, чтобы сравнивать величины A и C (если A>C, то термин принадлежит именно позитивному множеству).

Ниже представлен алгоритм извлечения ключевых слов из текстов предметной области на основе критерия хи-квадрат.

Алгоритм 1: Извлечение ключевых слов из текстов коллекции

1 Цикл по словарю коллекции
2 Извлечь очередное *слово* из словаря коллекции
3 A:=0; B:=0; C:=0; D:=0;
4 Цикл по документам коллекции
5 Извлечь очередной *документ* из коллекции
6 Если *документ* относится к *предметной области* и содержит *слово*, то A:=A+1
7 Если *документ* относится к *предметной области* и не содержит *слово*, то B:=B+1
8 Если *документ* относится к *предметной области* и не содержит *слово*, то C:=C+1
9 Если *документ* относится к *предметной области* и не содержит *слово*, то D:=D+1
10 Конец цикла
11 Chi2:=(A+B+C+D)*(A*D-B*C)^2/((A+B)*(A+C)*(B+D)*(C+D))
12 Если (Chi2>6.6) И (A>C), то добавить *слово* в *список ключевых слов*
13 Конец цикла

2.3. Методы автоматического распознавания терминов в одиночных документах

Одно из неудобств контрастного извлечения терминов заключается в том, что обязательно требуется альтернативная коллекция текстов. Ее подбор иногда становится слишком нетривиальным, т.к. не существует единой универсальной коллекции, одинаково «оппозиционной» ко всем предметным областям. По-

этому наряду с контрастными методами современные исследователи активно изучают методы извлечения терминов, основанные на анализе внутренних связей целевой коллекции или, если такая коллекция отсутствует, целевого документа.

Первый метод, который выбран для рассмотрения, был создан японскими учеными Matsuo и Ishizuka для извлечения ключевых слов из одиночных документов [35]. Метод основан на предположении, что если слово является термином, то в характере его совместного употребления со словами из группы высокочастотных терминов будет наблюдаться смещение в пользу определенной подгруппы. Степень смещения оценивается на основе критерия хи-квадрат по следующей формуле:

$$\chi^2(w) = \sum_{g \in G} \frac{\left(freq(w,g) - n_w p_g\right)^2}{n_w p_g}, \quad (21)$$

где w – текущее слово, G – множество высокочастотных терминов документа, $freq(w,g)$ – частота совместного появления слова w и высокочастотного термина $g \in G$ в ограниченных контекстах, p_g – безусловная вероятность появления термина g в документе, n_w – общее количество совместных появлений слова w и терминов из G в документе.

Чем выше разница между наблюдаемой частотой (совместного появления) и ожидаемой частотой (безусловной), тем менее вероятна нуль-гипотеза. То есть высокие значения хи-квадрат свидетельствуют о неслучайном характере появления данного термина в окрестности частотных терминов.

Таким образом, при реализации этого метода сначала извлекаются частые термины. Затем вычисляется совместная встречаемость каждого слова с частыми терминами (внутри предложений). Если появление данного термина в окрестности частых терминов не носит случайный характер, а смещено, то данный термин, скорее всего, является ключевым. Степень смещения измеряется мерой хи-квадрат. Метод демонстрирует качество сравнимое с мерой TF-IDF.

Частые термины извлекаются на основе подсчета частот их вхождений в документ. Частоты берутся относительные, т.е. нормируются так, чтобы сумма всех частот составляла 1. Для подсчета совместной встречаемости слов с частыми терминами, документ делится на предложения. Название документа, заголовки разделов и подписи тоже считаются предложениями. Два термина, встретившиеся в предложении, считаются за один случай совместной встречаемости. Если обозначить через N число различных терминов в документе, то матрица совместной представляет собой симметричную матрицу N × N. Для улучшения работы алгоритма авторы используют 2 приема: нормализация длины предложений и повышение устойчивости критерия хи-квадрат.

Первый прием обусловлен тем, что документ состоит из предложений различной длины. Если термин появляется в длинном предложении, он, вероятно, будет встречаться со многими терминами. Если термин появляется в коротком предложении, он менее вероятен для совместного использования с другими терминами. Поэтому для нормализации предложений вводятся новые значения:

- Pg – длина предложений, где встречается термин g, деленная на общую длину документа.
- nw – длина предложений, где появляется термин w.

Второй прием обусловлен тем, что термин, совпадающий с определенным термином из частотного пула $g \in G$, имеет высокое значение $\chi 2$. Однако эти термины иногда дополняют термин g, а не являются важными сами по себе. Например, термин internal имеет высокую связь с частым термином state, поскольку эти термины используются в устойчивом сочетании internal state. Предположив, что state не является частым термином, можно существенно уменьшить значение $\chi 2$ для термина internal.

$$\chi^{2'}(w) = \chi^2(w) - max_{g \in G} \frac{\left(freq(w,g) - n_w p_g\right)^2}{n_w p_g}, \quad (22)$$

Повышения устойчивости также можно добиться за счет кластеризации терминов. Матрица совместной встречаемости

первоначально представляет собой матрицу $N \times N$, где для вычисления извлекаются столбцы, соответствующие частым терминам. Остальные столбцы игнорируются, т.е. совместная встречаемость термина с низкочастотными терминами не берется в расчет, поскольку трудно оценить точную вероятность появления низкочастотных терминов. Итоговый алгоритм выглядит, как показано на рисунке 16.

Как альтернативу методу Matsuo & Ishizuka мы разработали два собственных способа извлечения терминов на основе анализа внутренних связей предметной коллекции. Для этой цели мы использовали тематическое моделирование документов на базе неотрицательной матричной факторизации (NMF) и латентного размещения Дирихле (LDA), которое подробно описали в работе [36]. Суть тематического моделирования заключается в переходе из пространства слов в пространство тем, т.е. в выделении внутри коллекции или документа набора локальных тем, каждая из которых описывается собственным набором слов [37].

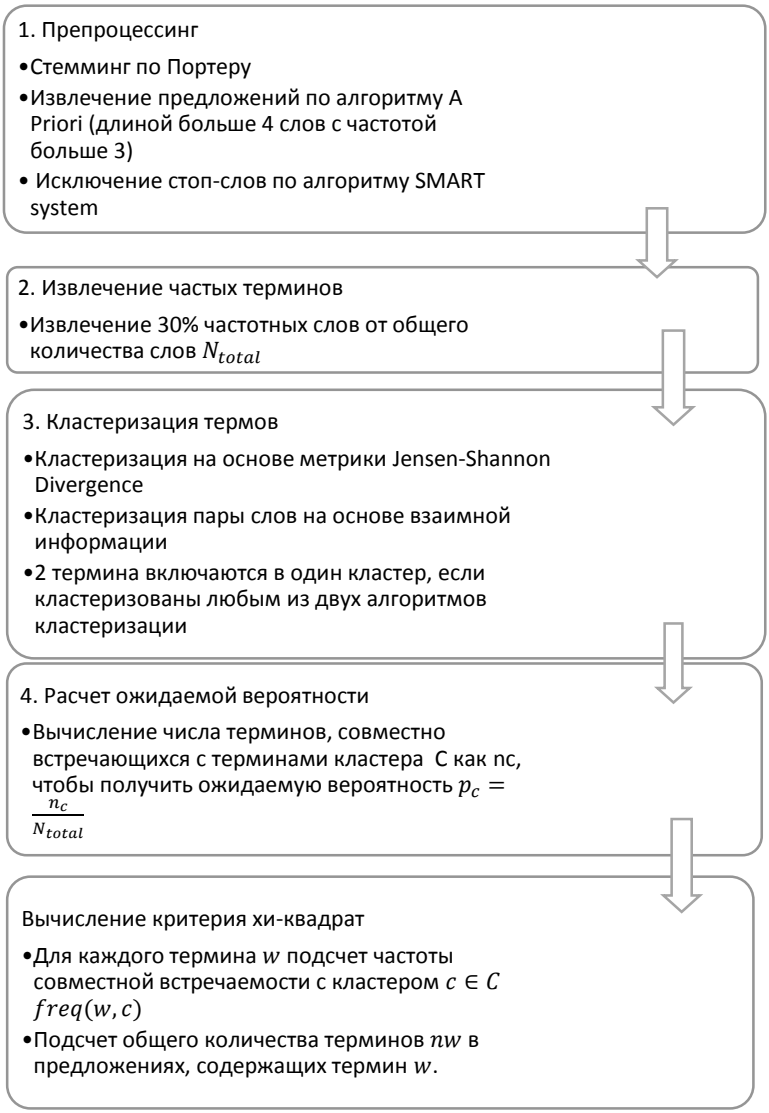


Рисунок 16 – Итоговый алгоритм извлечения терминов из одиночных документов на основе работы [28]

3. КЕЙС-СТАДИ ПО КОНТРАСТНОМУ ИЗВЛЕЧЕНИЮ ТЕРМИНОВ ИНФОРМАЦИОННОГО ПОИСКА

3.1. Предварительные сведения

В данной главе рассматривается практический пример (кейс-стади), направленный на контрастное извлечение терминов из популярного учебника по информационному поиску под авторством К. Маннинга, П. Рагхавана и Х. Шютце. Результаты извлечения терминов оцениваются с помощью эталонного указателя терминов, составленного самими авторами.

Британский корпус академического письменного английского языка, сокращенно именуемый BAWE (The British Academic Written English), создавался как совместный проект трех британских вузов: Уорикского университета, Университета Рединга и Университета Оксфорд Брукс. Цель проекта состояла в том, чтобы собрать в единый корпус лучшие образцы письменных работ студентов-старшекурсников и магистрантов указанных вузов [38]. Таким образом, в корпус вошло около 3000 работ по 35 учебным дисциплинам, представляющим 4 области наук: искусство и гуманитарные науки, науки о жизни, физические науки и социальные науки.

В настоящее время корпус доступен для скачивания из Оксфордского архива текстов как ресурс под номером 2539 [39]. В этом виде он содержит 2761 документ, каждый из которых снабжен подробной аннотацией, включающей в себя такие данные как код работы, ее название, курс, дата написания, жанр работы, учебная дисциплина, полученная оценка, количество слов. Аннотирована и информация об авторе каждой работы, в частности, такая аннотация содержит данные о поле студента, его годе рождения, первом языке, стране, откуда он родом, и т.д.

Изначально корпус создавался для исследования языковых особенностей, присущих письменным работам студентов британских высших учебных заведений [40]. В частности, по собранным в корпусе образцам изучались стиль, лексика, жанровое разнообразие академических письменных работ, зависимость стиля и жанра от области наук и дисциплины. Впоследствии корпус стал широко использоваться не только лингвистами, но и всеми, кто заинтересован в изучении письменного английского языка.

В области обработки естественного языка корпус BAWE стал использоваться в качестве тестовой коллекции документов практически сразу с момента своей публикации в открытом доступе. Так, пилотная версия корпуса, состоящая всего из 500 документов, была использована в работе [41] для проведения экспериментов по автоматическому определению гендерной принадлежности авторов документов. По итогам экспериментов у 81% авторов работ пол был идентифицирован верно.

В работе [42] корпус использовался для проведения экспериментов по тематическому моделированию. Авторы использовали для проверки своего метода тексты корпуса BAWE, относящиеся к области искусства и гуманитарных наук. В работе [43] авторы использовали тексты корпуса для автоматического определения тематики в английских предложениях. Разработанная этими авторами система Theme Analyzer автоматически определяла не только тема-рематическую структуру каждого предложения, но и входящие в него синтаксические узлы, тематические роли и т.д.

Одной из интересных практик применения корпуса BAWE является его использование в качестве альтернативной коллекции документов, необходимой для сопоставления с какой-либо другой коллекцией, интересующей исследователя. В работе [44] авторы используют BAWE вместе с коллекцией текстов, содержащих описания ритуальных действий, для того, чтобы извлечь ключевые слова, связанные с этой предметной областью. Авторы используют хорошо зарекомендовавший себя контрастный подход, выявляющий ключевые слова предметной области с позиции их разной встречаемости внутри предметной области и за ее пределами. Считается, что слова, которые часто употребляются внутри предметной области и крайне редко за ее пределами,

являются ключевыми. В данном цитируемом случае «внутри предметной области» означает в текстах, описывающих ритуалы, а «за ее пределами» означает в текстах альтернативной коллекции, т.е. в текстах корпуса BAWE.

В работе [45] авторы также используют BAWE для сравнения с другим корпусом, что по их словам, является «прямым, практичным и увлекательным способом изучения характеристик корпусов и типов текста». Авторы этой работы анализируют топ-100 ключевых слов каждого из рассматриваемых корпусов и сравнивают эти списки между собой.

Цель данного кейса – показать, что использование контрастного подхода при наличии сбалансированной и представительной альтернативной коллекции, в качестве каковой мы рассматриваем BAWE, позволяет эффективно решать задачу автоматического распознавания терминов, содержащихся в предметной коллекции текстов. В данной работе в качестве предметной коллекции текстов используется учебник по информационному поиску «Introduction to Information Retrieval» [46]. Учебник доступен в электронном виде на сайте Стэнфордского университета [47] и снабжен авторским указателем терминов, который используется в экспериментах как эталон для оценки точности и полноты извлечения терминов.

3.2. Содержательный анализ корпуса BAWE

Основными критериями качества альтернативной коллекции являются ее представительность и сбалансированность. Представительность означает, что альтернативная коллекция по возможности должна охватывать как можно больше текстов из как можно большего числа предметных областей, не смежных с целевой предметной областью. Сбалансированность означает, что различные предметные области в альтернативной коллекции по возможности должны быть представлены либо в равных пропорциях. С точки зрения названных критериев корпус BAWE является достаточно представительным (124516 словоупотреблений) и сбалансированным (4 области наук представлены примерно рав-

ными количествами текстов). В таблице 7 показано распределение текстов BAWE по областям наук, а на рисунке 17 представлена диаграмма, иллюстрирующая сбалансированность BAWE.

Таблица 7

Распределение текстов корпуса BAWE по областям наук

№	Область наук	Английское название	Количество текстов
1	Искусство и гуманитарные науки	Arts and Humanities (AH)	705
2	Науки о жизни	Life Sciences (LS)	683
3	Физические науки	Physical Sciences (PS)	596
4	Социальные науки	Social Sciences (SS)	777
	ИТОГО		2761

В работе [38] дается подробное описание состава корпуса, приводится статистика распределения текстов во всевозможных разрезах: по учебным дисциплинам, по жанрам, по годам, по курсам и т.д. На рисунке 18 мы приводим гистограммы распределения текстов корпуса в каждой из четырех областей наук с детализацией по учебным дисциплинам. Из гистограмм видно, что больше всего текстов в BAWE приходится на долю дисциплины Инженерия (238), затем идет Биология (169) и на третьем месте – Бизнес (146).

Мы проанализировали распределение слов в трех самых представительных дисциплинах и построили их облака (см. рисунки 19-21). Поскольку количество всех слов очень велико для визуализации, то использовались только слова с частотой употребления не меньше 70. Перед построением облаков тексты были подвергнуты предобработке: сначала выполнялась токенизация (разбиение текстов на слова и другие токены), затем лемматизация (приведение слов к нормальным формам), затем были удалены числа, знаки пунктуации и стоп-слова. В таблице 8 приведены попарные пересечения топ-100 ключевых слов для каждой из трех учебных дисциплин.

Таблица 8

Пересечения по топ-словам дисциплин «Инженерия», «Биология» и «Бизнес»

Инженерия и Биология – 26 слов	Инженерия и Бизнес – 47 слов	Биология и Бизнес – 35 слов
activity, change, control, development, factor, figure, form, group, high, important, increase, level, need, number, order, process, product, production, quality, rate, result, role, study, system, table, time, year	analysis, based, business, case, change, company, control, cost, current, customer, development, factor, figure, financial, good, group, high, important, increase, information, level, management, market, model, need, number, order, performance, point, power, price, problem, process, product, profit, project, rate, result, service, strategy, system, table, team, term, time, work, year	area, change, control, data, development, effect, energy, experiment, factor, figure, formula, group, high, higher, important, increase, level, method, need, number, order, picture, process, product, rate, required, result, small, stage, system, table, temperature, time, type, year

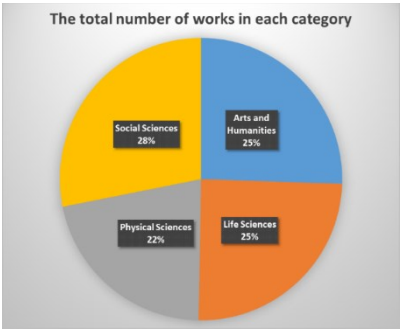


Рисунок 17 – Диаграмма распределения текстов корпуса BAWE по областям наук

Рисунок 18 – Распределение текстов корпуса BAWE по учебным дисциплинам

Попутно из числа извлеченных топ-100 слов были выделены топ-слова, общие для всех трех дисциплин (см. таблицу 8). Это такие распространенные в научных текстах слова как результат (result), система (system), фактор (factor), процесс (process), таблица (table) и т.д. Если распространить данный подход на все дисциплины корпуса BAWE, то можно говорить о перспективе автоматического построения словаря общенаучной и межотраслевой лексики на основе корпуса BAWE. Нам известны такие работы по автоматическому или полуавтоматическому построению словарей общенаучной лексики на основе корпусов научных текстов, например, для английского языка можно указать работу [48].

Таблица 9

Общеакадемические топ-слова, выделенные из пересечения топ-слов дисциплин «Инженерия», «Биология» и «Бизнес»

№	Слово	№	Слово	№	Слово
1	change	8	important	15	product
2	control	9	increase	16	rate
3	develop-ment	10	level	17	result
4	factor	11	need	18	system
5	figure	12	number	19	table
6	group	13	order	20	time
7	high	14	process	21	year

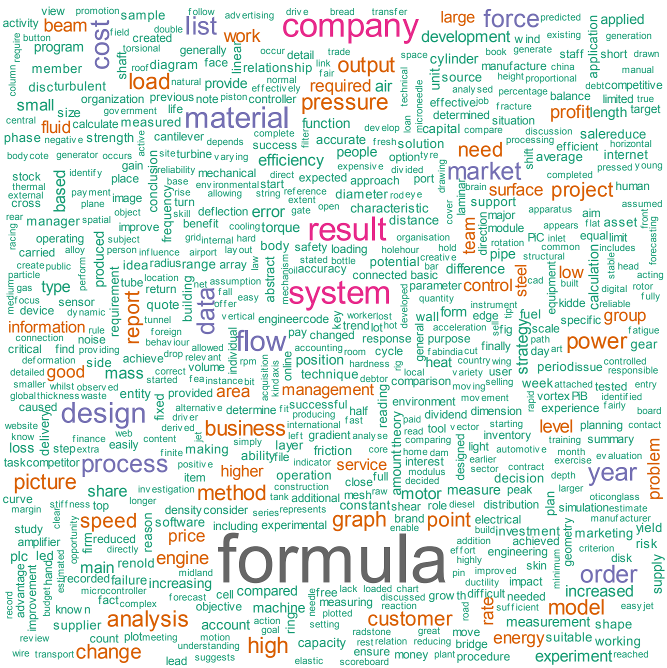


Рисунок 19 – Облако слов на основе текстов дисциплины «Инженерия»

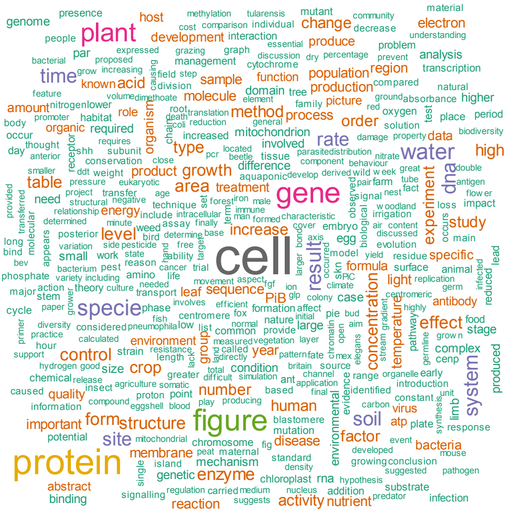


Рисунок 20 – Облако слов на основе текстов дисциплины «Биология»

Рисунок 21 – Облако слов на основе текстов дисциплины «Бизнес»

3.3. Извлечение терминов

В данной работе мы будем извлекать одно-, двух- и трехсловные термины и затем сравнивать список извлеченных терминов с эталонным авторским указателем. Эталонный указатель насчитывает 603 термина, в т. ч. 174 однословных, 335 двухсловных, 78 трехсловных, 14 четырехсловных и 1 шестисловный термин. Примеры даны в таблице 9.

Таблица 10

Примеры эталонных терминов из учебника «Introduction to Information Retrieval», содержащихся в авторском указателе (взяты в случайном порядке)

№	Однословные термины	Двухсловные термины	Трехсловные термины
1	accumulator	authority score	ad hoc retrieval
2	break-even	auxiliary index	binary independence model
3	BSBI	average-link clustering	blind relevance feedback
4	lemmatization	Bayes risk	clickthrough log analysis
5	likelihood	cumulative gain	maximum likelihood estimation
6	LSA	data-centric xml	multivariate Bernoulli model
7	NLP	support vector	natural language processing
8	regression	term frequency	principal left eigenvector
9	regularization	term-document matrix	unigram language model
10	Reuters-21578	word segmentation	vector space model

Благодаря наличию эталонного списка можно оценить точность $Precision$ и полноту $Recall$ рассматриваемых контрастных методов извлечения терминов. Для этого необходимо рассчитать значения, как показано в таблице 10. Полученные оценки точности и полноты извлечения терминов можно объединить в единый показатель, называемый F-мерой, с помощью среднего гармонического.

Таблица 11

Опорные значения для вычисления точности и полноты извлечения терминов

Обозначение	Название	Как определяется
TP	True Positive (число истинных обнаружений)	число извлеченных терминов, которые входят в эталонный список
FP	False Positive (число ложных обнаружений)	число извлеченных терминов, которые не входят в эталонный список
FN	False Negative (число ложных пропусков)	число терминов эталонного списка, которые не входят в число извлеченных терминов

3.4. Экспериментальная работа

Эксперименты по извлечению терминов проводились в R с использованием библиотек tm и quanteda [49]. Обе коллекции (главы книги и тексты корпуса BAWE) были загружены в R и затем преобразованы в вид удобный для обработки (представлены в виде разреженных матриц документы-на-термины).

Коллекция, составленная на основе глав книги «Introduction to Information Retrieval», была спарсена с сайта Стэнфордского университета, где она была выложена в открытом доступе в виде html-страниц. При парсинге была удалена html-разметка, и содержимое книги было экспортировано в 245 текстовых файлов по числу глав и параграфов книги. Коллекция, составленная на основе текстов корпуса BAWE, была скачана с сайта Оксфордского архива текстов, где она также была выложена в открытом доступе в виде архива текстовых файлов.

Текст каждого файла был лемматизирован с помощью инструмента Wordnet Lemmatizer, входящего в состав пакета NLTK – открытой библиотеки программ для символьной и статистической обработки естественного языка. Пос-таггинг при извлечении терминов не использовался, соответственно, поиск по лексическим шаблонам, характерным для двух- и трехсловных терминов, не применялся. Безусловно, это существенно понизило точность

извлечения терминов, т.к., например, для двухсловных терминов в основном характерны лексические шаблоны вида A+N (прилагательное + существительное), в то время как мы отбирали все двухсловные комбинации (биграммы). То же самое касается и трехсловных комбинаций (триграмм). В таблице 11 представлены топ-30 одно- и двухсловных терминов, извлеченных с помощью меры TF-DCF при использовании в качестве альтернативных коллекций всех четырех разделов BAWE.

Не все из извлеченных единиц являются терминами по авторской версии. Например, в авторском эталонном списке отсутствуют такие биграммы, как machine learning (хотя, по нашему мнению, это явный термин) или set document (это, безусловно, не термин). Очевидно, что эти примеры как раз и свидетельствуют о высокой сложности задачи извлечения терминов.

Таблица 12

Топ-30 одно- и двухсловных терминов, извлеченных с помощью TF-DCF при использовании в качестве альтернативных коллекций всех 4-х категорий BAWE

Ранг	Термин	Ранг	Термин	Ранг	Термин
1	postings list	11	relevant document	21	machine learning
2	query term	12	term document	22	term frequency
3	information retrieval	13	language model	23	IDF
4	text classification	14	crawler	24	number document
5	web search	15	nonrelevant	25	Rocchio
6	document collection	16	multinomial	26	document query
7	relevance feedback	17	single-link	27	complete-link
8	training set	18	Reuters-RCV1	28	set document
9	KNN	19	SVM	29	IR system
10	inverted index	20	linear classifier	30	centroid

В таблице 12 мы приводим топ-30 одно- и двухсловных терминов, извлеченных с помощью критерия хи-квадрат. Хотя на первый взгляд, между двумя списками терминов из таблиц 6 и 7 не слишком

большая разница, на самом деле, при внимательном сравнении можно определить, что критерий хи-квадрат сработал менее эффективно. Например, в топ-30 были включены такие слова как algorithm, compute, vector, Boolean, которые не являются терминами, принадлежащими только области информационного поиска, эти термины достаточно широко представлены и в таких областях как математика, компьютерные науки, инженерия.

Таблица 13

Топ-30 одно- и двухсловных терминов, извлеченных с помощью критерия хи-квадрат при использовании в качестве альтернативных коллекций всех 4-х категорий BAWE

Ранг	Термин	Ранг	Термин	Ранг	Термин
1	query	11	inverted index	21	IR system
2	retrieval	12	postings list	22	term occur
3	query term	13	vector	23	centroid
4	information retrieval	14	Boolean	24	naive
5	algorithm	15	document collection	25	IDF
6	posting	16	classifier	26	relevance feedback
7	compute	17	text classification	27	relevant document
8	vector space	18	retrieval system	28	naive Bayes
9	search engine	19	term frequency	29	machine learning
10	web search	20	IR	30	nonrelevant

Таким образом, как и ожидалось, показатели точности, полноты и F-меры при использовании метода TF-DCF оказались выше, чем при использовании критерия хи-квадрат. (см. таблицы 13-14). Поэтому при формировании онтологии в области информационного поиска, именно термины, отобранные с помощью меры TF-DCF, использовались как базовые конструкции для построения концептов. На рисунке 22 показано облако концептов

онтологии, извлеченных в соответствии с мерой TF-DCF. Отметим, что описание того, как из извлеченных терминов были сформированы концепты предметной области, выходит за рамки настоящей работы.

Таблица 14

Максимальные показатели качества извлечения терминов при использовании четырех альтернативных коллекций (AH+ LS+SS+PS) за вычетом дисциплины Computer Science

Показатель	Однословные термины		Двухсловные термины		Трехсловные термины	
	Хи-квадрат	TF-DCF	Хи-квадрат	TF-DCF	Хи-квадрат	TF-DCF
Точность	0,1818	0,24	0,1654	0,2018	0,1443	0,1271
Полнота	0,3086	0,24	0,1284	0,2627	0,1489	0,2447
F-мера	**0,2288**	**0,24**	**0,1446**	**0,2283**	**0,1466**	**0,1673**

Таблица 15

Показатели качества извлечения одно-, двух- и трехсловных терминов при использовании четырех альтернативных коллекций (AH+LS+SS+PS) за вычетом дисциплины Computer Science

Показатель	Хи-квадрат (с порогом 24)	TF-DCF (с порогом 5,5)
Точность	0,1045	0,2196
Полнота	0,2913	0,2470
F-мера	**0,1539**	**0,2325**

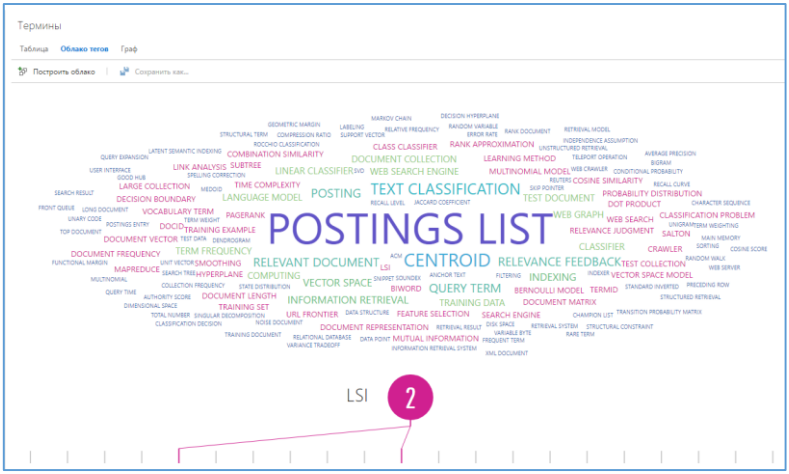


Рисунок 22 – Облако концептов предметной области «Информационный поиск»

Примечательным оказался результат сравнения показателей качества извлечения терминов при использовании меры TF-DCF в зависимости от числа альтернативных коллекций (см. таблицу 16). Максимальные показатели качества были достигнуты при использовании максимального количества альтернативных предметных областей – 4. При этом оказалось важным, чтобы альтернативные предметные области не пересекались с целевой предметной областью. Таким образом, при исключении из альтернативной коллекции части текстов, относящихся к компьютерным наукам (т.е. предметной области смежной с предметной областью информационного поиска) и точность, и полнота извлечения терминов повысились.

Еще один примечательный факт состоит в том, что при увеличении числа альтернативных текстов, полнота извлечения терминов падает. Например, при использовании 3 коллекций в число терминов, отсутствующих в альтернативных коллекциях, попали такие слова как index, stemming, entropy, в то время, как при использовании 4 коллекций эти термины уже оказались в числе рядовых слов, одинаково распространенных и в целевой коллекции,

и в альтернативных. Соответственно, полнота охвата терминологии снижалась за счет исключения этих слов из списка терминов. В целом, качество извлечения трех- и двухсловных терминов ниже, чем однословных, т.к., как мы отмечали выше, не использовались лексические шаблоны.

Таблица 16

Зависимость точности, полноты и F-меры от числа альтернативных коллекций при извлечении терминов с помощью меры TF-DCF (порог 5.5)

Показатель	2 альтернативные коллекции (LS+SS)	3 альтернативные коллекции (AH + LS + SS)	4 альтернативные коллекции (AH + LS + SS + PS)	4 альтернативные коллекции (AH + LS + SS + PS) без Computer sciences
Точность	0,2110	0,2156	0,2218	0,2196
Полнота	0,2675	0,2623	0,2419	0,2470
F-мера	**0,2359**	**0,2367**	**0,2314**	**0,2325**

Большой интерес вызывает проблема выбора порогового значения, как для критерия хи-квадрат, так и для меры TF-DCF. На рисунке 14 показано, как меняются значения F-меры при изменении порогового значения меры TF-DCF при извлечении одно-, двух- и трехсловных терминов. Согласно этим результатам, оптимальное пороговое значение, дающее максимум F-меры, лежит в диапазоне от 3 до 6 (4.4 – для однословных терминов; 5.6 – для двухсловных терминов; 3.2 – для трехсловных терминов).

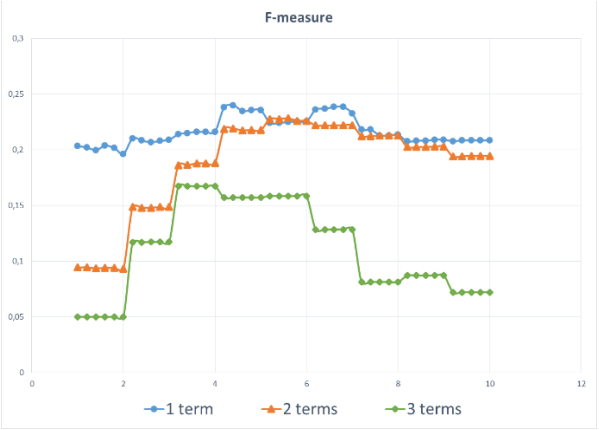


Рисунок 23 – Зависимость F-меры от порогового значения меры TF-DCF

Проведенные эксперименты позволили сделать следующие 3 важных вывода:

1) При увеличении количества альтернативных коллекций точность и полнота извлечения терминов увеличиваются, при этом важно использовать критерии, учитывающие не совокупное распределение терминов в альтернативных коллекциях, а их отдельные частоты в каждой отдельной коллекции.

2) Альтернативные коллекции не должны быть смежными рассматриваемой целевой коллекции.

3) Британский корпус академического письменного английского языка полностью удовлетворяет вышеперечисленным условиям и несмотря на свой сравнительно небольшой размер может с успехом использоваться в качестве набора альтернативных коллекций.

Следует также подчеркнуть, что при проведении экспериментов не использовались лексико-синтаксические шаблоны, описанные в разделе 1.3. Этим можно объяснить не слишком высокие показатели точности и полноты извлечения терминов.

В заключении покажем как ключевые слова и термины могут использоваться в тематическом моделировании. В качестве входных учебных материалов были взяты 3 источника на английском

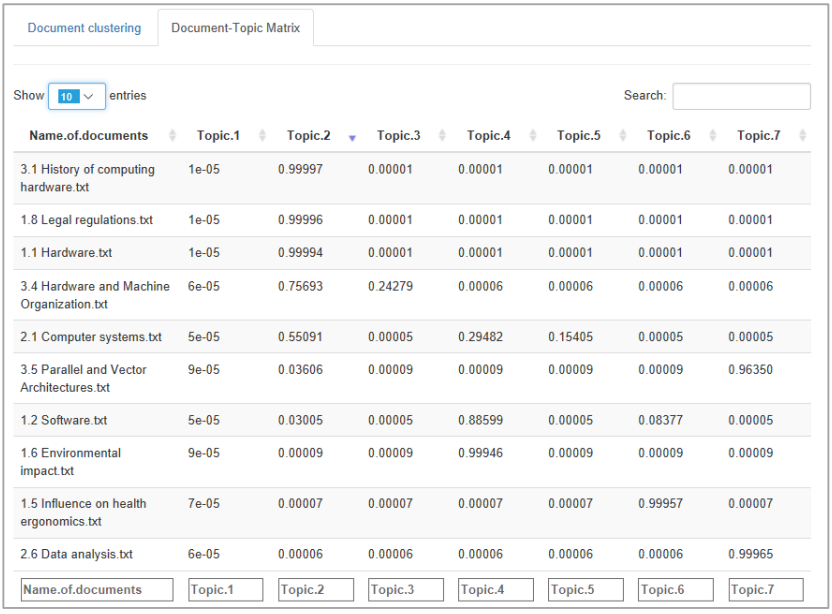
языке. Первый источник – электронный учебник «Основные понятия информационно-коммуникационных технологий», под авторством Гораны Целебич и Дарио Рендулича. Содержание учебника включает в себя 8 разделов, которые полностью соответствуют типовому содержанию учебной программы. Второй источник – сборник тезисов лекций по дисциплине «Информационные и коммуникационные технологии», разработанный преподавателями кафедры компьютерного моделирования и информационных технологий ВКГУ имени С. Аманжолова и утвержденный методическим советом вуза. Третий источник содержит материалы открытого и бесплатного онлайн-курса по основам компьютерных наук «CS301: Computer Architecture» на интернет-площадке Saylor Academy.

Для выполнения тематического моделирования в R предназначена функция LDA() из пакета topicmodels. Функция принимает в качестве параметра матрицу документы-на-термины, число выделяемых скрытых тем k (определяется эмпирически) и метод оценки правдоподобия:

```
res <- LDA (dfm, k, method = "VEM")
```

Функция возвращает два основных слота, первый слот показывает распределение слов по темам, второй – распределение документов по темам. На рисунке 24 показан фрагмент табличного представления второго слота, т.е. показаны веса каждой из 7 выявленных тем в каждом из документов корпуса. Произведена сортировка по весам второй темы, и явно видно, что это тема связана с аппаратным обеспечением. Например, документ «1.8. Legal regulations» («Правовые нормы»), который представляет собой 8-ю главу 1-го источника тесно связан с темой аппаратного обеспечения, из чего следует, что он раскрывает суть правового регулирования не в области ИКТ вообще, не в области программного обеспечения, а именно в области аппаратного обеспечения. То же самое справедливо в отношении документа «2.1. Computer Systems».

Разработанный интерактивный интерфейс для представления результатов тематического моделирования позволяет пользователю (эксперту) глубже понять тематическое представление источников и их разделов, а также выявить глубинные предметные связи между разными источниками и темами.



Document clustering | Document-Topic Matrix

Show 10 entries — Search:

Name.of.documents	Topic.1	Topic.2	Topic.3	Topic.4	Topic.5	Topic.6	Topic.7
3.1 History of computing hardware.txt	1e-05	0.99997	0.00001	0.00001	0.00001	0.00001	0.00001
1.8 Legal regulations.txt	1e-05	0.99996	0.00001	0.00001	0.00001	0.00001	0.00001
1.1 Hardware.txt	1e-05	0.99994	0.00001	0.00001	0.00001	0.00001	0.00001
3.4 Hardware and Machine Organization.txt	6e-05	0.75693	0.24279	0.00006	0.00006	0.00006	0.00006
2.1 Computer systems.txt	5e-05	0.55091	0.00005	0.29482	0.15405	0.00005	0.00005
3.5 Parallel and Vector Architectures.txt	9e-05	0.03606	0.00009	0.00009	0.00009	0.00009	0.96350
1.2 Software.txt	5e-05	0.03005	0.00005	0.88599	0.00005	0.08377	0.00005
1.6 Environmental impact.txt	9e-05	0.00009	0.00009	0.99946	0.00009	0.00009	0.00009
1.5 Influence on health ergonomics.txt	7e-05	0.00007	0.00007	0.00007	0.00007	0.99957	0.00007
2.6 Data analysis.txt	6e-05	0.00006	0.00006	0.00006	0.00006	0.00006	0.99965
Name.of.documents	Topic.1	Topic.2	Topic.3	Topic.4	Topic.5	Topic.6	Topic.7

Рисунок 24 – Табличное представление LDA

Более удобным средством визуализации тематической модели LDA является интерактивный инструмент LDAvis., при выборе круга справа отображается список ключевых слов, соответствующих теме (см. рисунок 25). Длина списка регулируется. Как показано на рисунке 25, визуализация состоит из двух основных частей. Левая панель визуализации отвечает за отображение тем и их отношений. Каждая тема на этой визуализации представляет собой пронумерованный круг, размер круга определяется весом темы в коллекции. Более близкие темы показаны ближе друг к другу, некоторые даже пересекаются. Правая панель визуализа-

ции представлена в виде горизонтальной шкалы, которая отображает наиболее подходящие термины для объяснения выбранной темы. Эти слова раскрывают суть каждой темы.

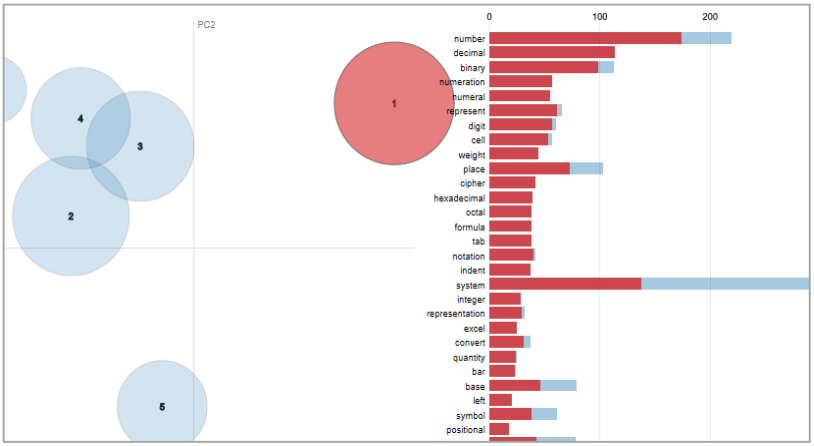


Рисунок 25 – Инструмент LDAvis для визуализации темы

Ползунок над списком ключевых слов позволяет пользователям изменять значение λ, параметра, который может изменить ранжирование ключевых слов темы. По умолчанию для параметра λ установлено значение 0.6. Если параметр равен 1, то вверх поднимаются отличительные термины темы, которые практически отсутствуют в других терминах. Чем ниже параметр, тем выше ранги общеупотребительных терминов темы. Красная и синяя полосы рядом с термином показывают соответственно его долю в текстах данной темы и текстах других тем. Так, например, на указанном рисунке в панели ключевых слов высокий ранг имеют слова «число», «система», «десятичная», «двоичная», «восьмеричная», «нумерация», «цифра», «представлять», «цифра», «шестнадцатеричная», Сочетание этих терминов позволяет определить, что речь идет о теме «Системы счисления», а доля красного и синего напротив каждого термина показывает, насколько этот термин привязан только к данной теме. Например,

высокая доля синего напротив слова «система» говорит об универсальности этого термина, а высокая доля красного напротив слова «десятичный» об узкой направленности последнего.

Таким образом, разработанные модели извлечения терминов и тематического моделирования позволяет пользователю не только описать предметную область, но и визуализировать ее, показать взаимосвязь между порзаделами предметной области. Это экономит время, затрачиваемое на изучение, анализ, подбор соответствующей литературы, а также определяет наиболее подходящий контент в учебном курсе.

4. ИЗВЛЕЧЕНИЕ ИМЕНОВАННЫХ СУЩНОСТЕЙ

4.1. Предварительные сведения

В настоящее время существуют различные подходы для извлечения информации. Они разнообразны и сложно сказать, что один лучше другого, так как тот или иной показывает хорошие результаты в разных ситуациях. Подходы для извлечения информации могут быть классифицированы на следующие категории:

– подходы, основанные на правилах. Эксперты составляют вручную наборы правил, необходимые для извлечения определённых данных.

– подходы, основанные на знаниях. Сюда относятся модели, основанные на онтологиях [50], модели, основанные на тезаурусах [51].

– статистические подходы. Включают скрытые марковские модели [52-54], условные марковские модели [55], условные случайные поля [56].

– подходы, основанные на машинном обучении.

Одной из подзадач извлечения информации является распознавание именованных сущностей. Из текста выделяются именованные объекты, такие как имена людей, названия организаций, локации, геополитические объекты, временные метки, расширенный вариант может включать специфичные для определённой предметной области термины, такие как медицинские, биологические, др. Существующие методы распознавания именованных сущностей можно разделить на категории:

– Основанные на правилах. Являются одними из ранних систем для распознавания именованных сущностей. Правила основываются на лексико-синтаксических шаблонах специфичных

для определённого языка. Поэтому системы, основанные на данном методе, считаются наиболее эффективными [57]. Однако данные системы ограничены для области, для которой они определены, и непереносимы. [58, 59].

– Обучение с учителем. Данный метод требует большого количества тренировочных данных, размеченных вручную экспертами. Затем система на основе предоставленных данных извлекает правила для распознавания именованных сущностей. К данной категории относятся такие методы, как условные случайные поля (Conditional Random Fields) [60], максимальная энтропия (Maximum Entropy) [61], деревья решений (Decision trees) [62], др.

– Обучение без учителя. Система использует небольшой набор экземпляров именованных сущностей, например, страны {«Канада», «Южная Корея», «Япония», …}. Данный набор изучается системой и на основе предложений, в которых встречаются объекты из данного набора, вырабатываются некоторые правила извлечения именованных сущностей. Эти правила применяются для определения новых сущностей. Затем изучается новый набор правил. Таким образом обучение продолжается до тех пор, пока не будут обнаружены новые правила. Например, в работе [63] обсуждается неконтролируемая модель классификации именованных сущностей с использованием немаркированных примеров данных. В работе [64] предлагается неконтролируемая классификация именованных сущностей и техника ансамбля, в которой использовался небольшой словарь именованных сущностей и немаркированный корпус для именованных сущностей.

– Гибридные системы. Включают в себя две или более техники обучения на основе машинного обучения или на основе правил. [65, 66].

4.2. Предыдущие работы

Классические системы извлечения именованных сущностей используют выделенные вручную свойства [67]. В некоторых ранних системах использовались правила, разработанные вручную [68, 69], однако подавляющее большинство современных систем опираются на модели машинного обучения [70], такие как условное случайное поле (CRF) [71], Скрытая Марковская модель (HMM) [72], метод опорных векторов (SVM) [73]. Хотя традиционные модели машинного обучения не основаны на ручных правилах, они требуют ручного процесса разработки функций, что является довольно дорогостоящим и зависит от домена и языка. В последнее время множество работ с применением нейронных сетей превзошли классические системы. В последние годы модели с рекуррентной нейронной сетью (RNN), такие как Long-Short-Term-Memory (LSTM) [74], Gated Recurrent Unit (GRU) [75] были очень успешными в задачах моделирования последовательностей, например, Language Modeling [76, 77], машинный перевод [78], Dialog Act classification [79, 80]. Одной из сильных сторон моделей RNN является их способность обучаться на основных компонентах текста (то есть словах и символах). Эта возможность обобщения облегчает построение независимых от языка моделей NER [81, 82], которые основаны на неконтролируемом изучении свойств и небольшом аннотированном корпусе.

Впервые применение нейронных моделей в задаче маркировки последовательности было предложено Collobert et. al. [83]. Однако для данной модели существуют некоторые ограничения. Во-первых, здесь используется простая нейронная сеть с прямой связью, что ограничивает диапазон рассматриваемого контекста вокруг слов. Модель забывает полезные отношения между словами на большом расстоянии. Во-вторых, из-за зависимости исключительно от векторизации слов, невозможно определение и использование свойств, представленных на уровне символов, таких как суффиксы и префиксы.

Позже были предложены модифицированные модели с использованием двунаправленной LSTM или Stacked LSTM [84, 85]. Например, в работе [84] используется архитектура на основе bi-LSTM и CRF. Авторы работы [86] используют архитектуру bi-LSTM-CNNs (рис. 26). Для векторизации символов они предлагают применение свёрточных нейронных сетей. Были найдены

новые подходы, использующие CNN или LSTM для извлечения информации подслов из ввода символов, результаты которых превзошли другие модели [87]. Rei et. al. [88] предложили модель, в качестве входных данных на которую подаются слова и символы.

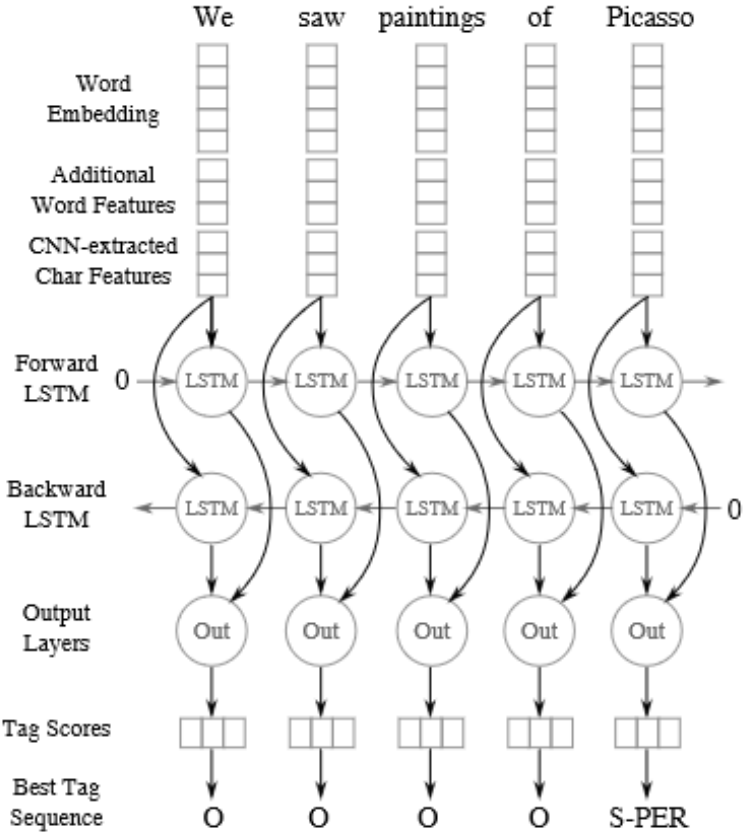


Рисунок 26 – Архитектура модели, предложенная Jason P.C. Chiu и E. Nichols [86]

Kuru et. al. [89] предложили нейронную модель на основе символов. Данная модель, принимающая на вход только символы, демонстрирует хорошие показатели, с условием что никакие внешние данные не используются. Данная модель предсказывает тег для каждого символа и следит, чтобы у всех символов в слове предсказанные теги были одинаковыми.

В данной работе рассматривается нейронная модель, где и слова, и символы представляются в векторной форме и подаются на вход блоку двунаправленной LSTM для моделирования контекстной информации. Для кодировки информации на уровне символов используется также двунаправленная LSTM.

4.3. Модель

Для решения задачи извлечения именованных сущностей разработана модель, основанная на bi-LSTM блоке с векторизацией символов и слов (рис. 27). При построении модели был использован подход, представленный в [90]. Авторы работы предложили новую архитектуру нейронной сети, которая автоматически обнаруживает функции на уровне слов и символов, используя гибридную двунаправленную архитектуру bi-LSTM и CNN.

4.3.1. LSTM

LSTM (Long Short-Term Memory) – разновидность рекуррентных нейронных сетей. Рекуррентные нейронные сети обладают способностью запоминания результатов прошлых итераций, однако они не способны запоминать их долгосрочно. Появляется проблема исчезновения градиента [91]. LSTM сети [t] разработаны с целью бороться с данной проблемой. Они содержат три основных блока, которые контролируют какая информация будет забыта, а какая будет передана на последующие итерации. Схематично LSTM блок может быть изображен как на рис. 28.

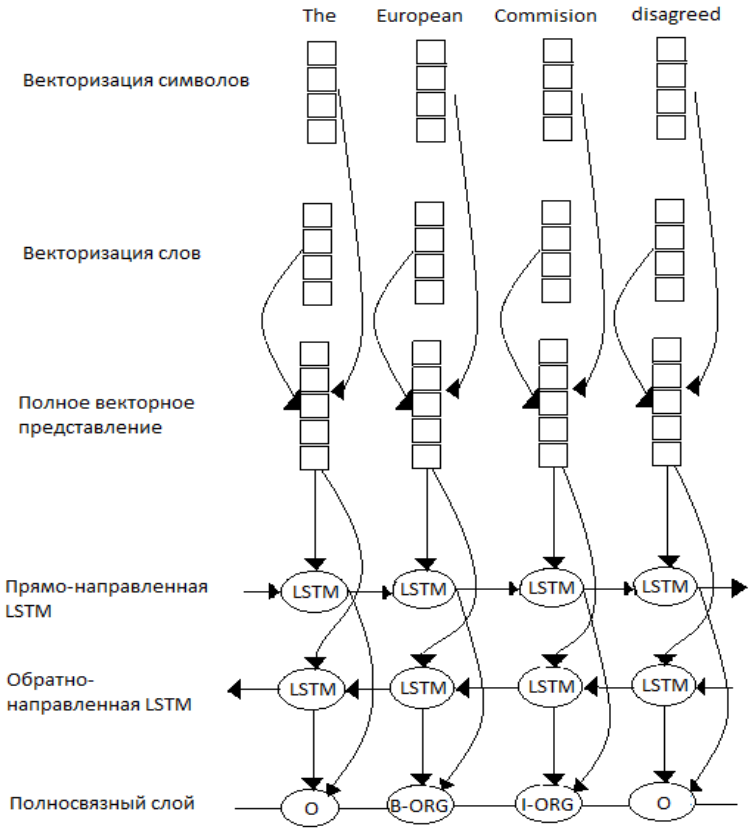


Рисунок 27 - Основная архитектура сети

Формально формулы для обновления LSTM блока в момент времени t можно представить как:

$$i_t = \sigma(W_i h_{t-1} + U_i x_t + b_i)$$

$$f_t = \sigma(W_f h_{t-1} + U_f x_t + b_f)$$

$$\tilde{c}_t = tanh(W_c h_{t-1} + U_c x_t + b_c)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t$$

$$o_t = \sigma(W_o h_{t-1} + U_o x_t + b_o)$$

$$h_t = o_t \odot \tanh(c_t)$$

где

x_t – входной вектор в момент времени t,

h_t – вектор скрытого состояния (вывода) в момент времени t,

σ- сигмоидная функция,

Ui, Uf, Uc, Uo – матрицы весов различных фильтров для входного вектора x_t,

Wi,Wf,Wc,Wo – матрицы весов для вектора скрытого состояния h_t,

$\odot$ - поэлементное умножение,

b_i, b_f, b_c, b_o – векторы смещения.

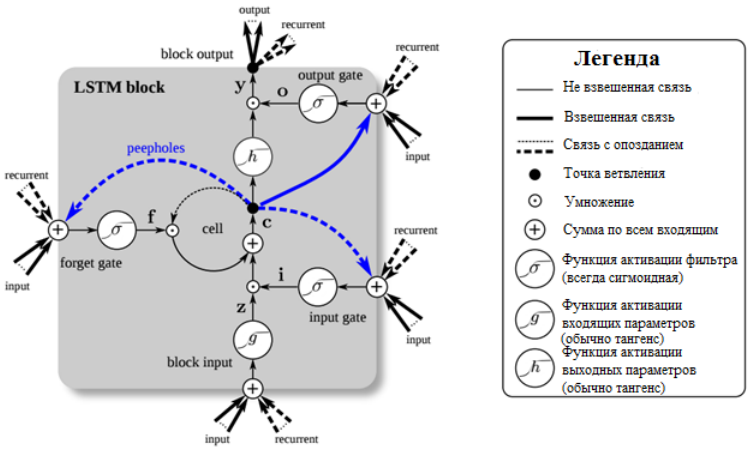


Рисунок 28 - LSTM блок [92]

4.3.2. Bi-LSTM

Главной идеей bi-LSTM является учитывание последовательности символов не только до текущей, но и после [93]. Таким образом происходит рассмотрение полного окружения. Bi-LSTM состоит из прямо направленной LSTM, которая рассматривает последовательность до и обратно направленной, рассматривающей последовательность после. Затем полученные последовательности конкатенируются.

4.3.3. Векторизация

Целью векторного представления или векторизации является извлечение информации из текстового корпуса и сопоставление каждому его элементу (слову/символу) уникального числового вектора. Векторизация является одним из подходов к моделированию языка и обучению представлений в обработке естественного языка, направленных на сопоставление словам из некоторого словаря векторов значительно меньшего количества слов в словаре. Теоретической базой для векторных представлений является дистрибутивный подход к обработке естественного языка, который представляет собой группу методов, предназначенных для изучения семантической близости между языковыми единицами (словами, понятиями, документами) на основании оценки распределения (дистрибуции) слов в текстах. Основными инструментами дистрибутивного анализа являются контекстные векторы и матрицы совместной встречаемости [94].

Под контекстным вектором слова понимается вектор, указывающий на слова, с которыми данное слово встречается в одном контексте [95]. Под контекстным вектором документа, понимается вектор, указывающий на слова, которые встречаются в данном документе. Тогда семантическое расстояние между двумя словами или документами определяется как евклидово расстояние или косинусная близость между соответствующими им контекстными векторами.

Под матрицей совместной встречаемости понимается матрица, строками и столбцами которой являются языковые единицы, а на пересечении строк и столбцов записаны показатели совместного употребления или соотнесенности языковых единиц в общем контексте. Например, матрица термины-на-термины может быть сформирована на основе совместной встречаемости терминов в одном документе или в одном документе. Она может быть бинарная, т.е. состоять из нулей и единиц, если два термина входят в один документ или одно предложение, указывается единица, иначе 0. Также такая матрица может быть частотной, например, указывать на количество документов или предложений, в которых два термина встречаются вместе. Классическая матрица совместной встречаемости представляет собой матрицу «термины-на-документы», элементами которой являются частоты (относительные или абсолютные) вхождения терминов в документы предметной коллекции.

Покажем, например, как можно использовать дистрибутивную матрицу «документы-на-термины» для оценки силы семантических связей между терминами. Каждый термин в матрице представляет собой вектор-столбец, таким образом, семантическую связь между любыми двумя терминами можно рассматривать как близость или расстояние между соответствующими векторами, используя при этом любые известные метрики векторного пространства. Например, как отмечалось выше, косинусную меру:

$$r_{ij} = cos\left(\bar{T}_i, \bar{T}_j\right) = \frac{\bar{T}_i \cdot \bar{T}_j}{|\bar{T}_i| \cdot |\bar{T}_j|},$$

где $\bar{T}_i, \bar{T}_j$ – это вектор-столбцы матрицы «документы-на-термины», соответствующие *i*-му и *j*-му терминам соответственно (i и j пробегают весь список терминов), r_{ij} – это значение близости, элемент матрицы семантических связей.

Роль дистрибутивных матриц и контекстных векторов в задаче семантического анализа естественно-языковых текстов трудно переоценить ([96, 97, 98]). К числу наиболее известных

моделей и методов обработки естественного языка и информационного поиска, основанных на дистрибутивных подходах, относятся:

- Модель Bag-of-words;
- Модель Bag-of-related words;
- Латентный семантический анализ (LSA);
- Латентное размещение Дирихле (LDA);
- Неотрицательная матричная факторизация;
- Машина опорных векторов;
- Модель Word2Vec;
- Модель Word embedding (погружение слов в линейное векторное пространство).

В ходе исследований авторами были использованы все перечисленные модели и методы. В частности, модель Bag-of-words и ее усложненный вариант Bag-of-related words использовались для классификации документов, для автоматического сегментирования текстов, для извлечения семантических связей между терминами и документами [99]. Латентный семантический анализ использовался для поиска ассоциативных связей, для снижения признакового пространства при индексировании документов ключевыми словами, для очистки дистрибутивных матриц от шума и разреженности [100]. Как отмечается в работе [101], все частотные матрицы являются разреженными и зашумленными. Поэтому классический способ улучшения характеристик таких матриц – это подвергнуть их сингулярному разложению и уменьшить признаковое пространство, оставив только главные признаковые компоненты. Латентное размещение Дирихле и неотрицательная матричная факторизация использовались для автоматической сегментации текстов [99] и формирования визуального образа документа в виде набора тем и опорных ключевых слов. Метод и программный инструмент Word2Vec использовался при разработке системы извлечения фактов.

Одним из перспективных применений матрицы совместной встречаемости терминов является ее использование в качестве

основы для построения таксономии. В этом случае матрица строится не на всем множестве слов, а на множестве ключевых терминов и их атрибутов, т.е. слов, употребляемых с ключевыми словами в одном паттерне. Примерами таких паттернов являются лексико-семантические шаблоны вида: Подлежащее + Определение, Подлежащее + Сказуемое, Подлежащее + Дополнение. Сформированная таким образом матрица может использоваться в качестве формального контекста, на основе которого строится решетка понятий, т.е. выделяются понятия предметной области и их иерархии [102]. Такой подход применялся при парсинге медицинских текстов с целью выявления ключевых групп понятий в области медицины [103]. Вкупе с методами кластеризации и машинного обучения анализ формальных понятий на основе дистрибутивных матриц представляет собой мощный инструмент автоматического построения онтологий.

Дистрибутивные матрицы можно рассматривать также как компактный и удобный способ описания связей между словами, которые затем можно визуализировать с помощью семантического графа или концепт-карты [104, 105, 106]. Вершинами графа являются строки и столбцы матрицы, а ребрами – связи, веса которых определяются из значений матрицы.

В данной работе использовался способ векторизации – one-hot embedding [107].

4.4. Экспериментальная часть

Набор данных

Обучение построенной модели проводилось на данных на английском языке, был использован набор данных CoNLL-2003 [108]. Объём набора составил 1629 предложений. Эксперименты проводились в Python.

Таблица 17

Количество именованных сущностей по категориям в наборе данных

	LOC	MISC	ORG	PER
Тренировочный набор	7140	3438	6321	6600
Валидационный набор	1837	922	1341	1842
Тестовый набор	1668	702	1661	1617

Данные были размечены с применением девяти тегов:

1. B-PER
2. I-PER
3. B-LOC
4. I-LOC
5. B-ORG
6. I-ORG
7. B-MISC
8. I-MISC
9. O

где PER – PERSON (персона, имена людей), LOC – LOCATION (локация), ORG – ORGANIZATION (организация), MISC – MISCELLANEOUS (названия смешанных именованных сущностей, не относящихся к предыдущим трём категориям), B – BEGINNING (начало, первый токен именованной сущности), I – INSIDE (последующие (внутренние) токены именованной сущности), O - OTHER, токен, не являющийся именованной сущностью.

4.5. Результаты

Обучение модели проходило в течение 10 эпох. На рисунке 29 показано изменение точности маркировки именованных сущностей с течением эпох.

Пример распознавания именованных сущностей в предложении, где во второй колонке показаны исходные теги, а в третьей – теги, предсказанные моделью, представлены на рис. 30.

Анализ построенной модели осуществлялся с помощью F1-меры. Значение F1 было вычислено с макро-усреднённой и микро-усреднённой мерой (рис. 31).

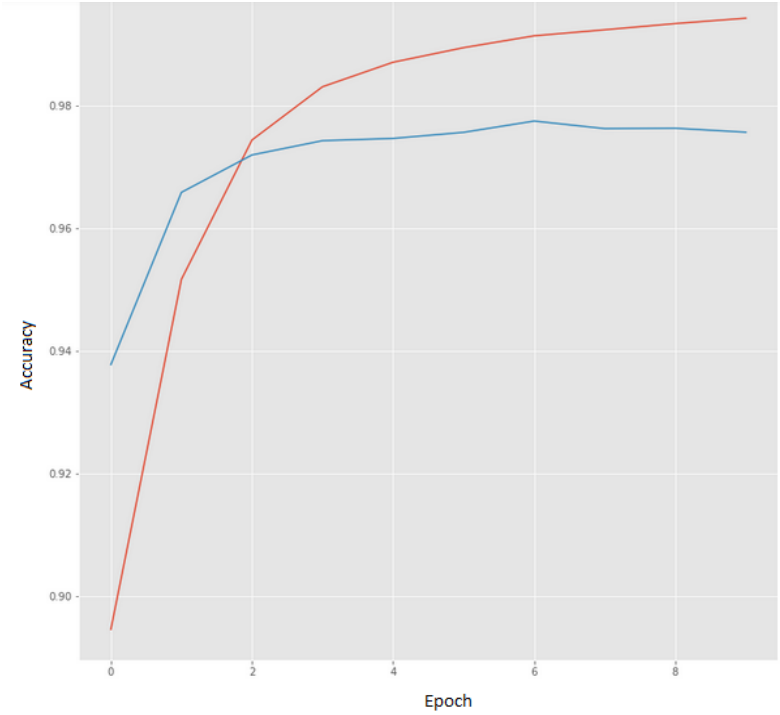
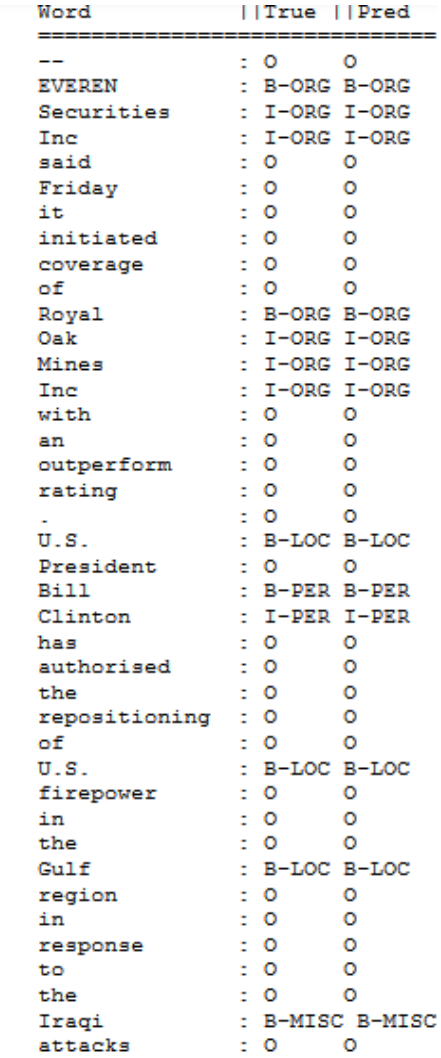


Рисунок 29 - Изменение точности (accuracy) в зависимости от эпохи, где красной линией показана точность при тренировке, синией линией – точность при валидации



```
Word            ||True ||Pred
==============================
--              : O     O
EVEREN          : B-ORG B-ORG
Securities      : I-ORG I-ORG
Inc             : I-ORG I-ORG
said            : O     O
Friday          : O     O
it              : O     O
initiated       : O     O
coverage        : O     O
of              : O     O
Royal           : B-ORG B-ORG
Oak             : I-ORG I-ORG
Mines           : I-ORG I-ORG
Inc             : I-ORG I-ORG
with            : O     O
an              : O     O
outperform      : O     O
rating          : O     O
.               : O     O
U.S.            : B-LOC B-LOC
President       : O     O
Bill            : B-PER B-PER
Clinton         : I-PER I-PER
has             : O     O
authorised      : O     O
the             : O     O
repositioning   : O     O
of              : O     O
U.S.            : B-LOC B-LOC
firepower       : O     O
in              : O     O
the             : O     O
Gulf            : B-LOC B-LOC
region          : O     O
in              : O     O
response        : O     O
to              : O     O
the             : O     O
Iraqi           : B-MISC B-MISC
attacks         : O     O
```

Рисунок 30 - Пример распознавания именованных сущностей в предложении, где во второй колонке показаны исходные теги, а в третьей – теги, предсказанные моделью

```
f1-score with average macro: 0.8982352937221123
f1-score with average micro: 0.9799387757074335
```

Рисунок 31 - Значение F1-меры

Построенная модель нейронной сети, которая включает в себя bi-LSTM с векторизацией символов и слов, достигает хороших результатов в распознавании именованных сущностей. Областью дальнейших исследований будет более эффективное построение и применение гибридных подходов нейронных сетей с погружением слов в линейное векторное пространство (word embedding).

ЗАКЛЮЧЕНИЕ

Автоматическое извлечение терминов из текстов предметной области представляет собой задачу, которая имеет множество приложений. Термины, извлекаемые автоматическим способом, могут использоваться как классификационные признаки для рубрикации документов, как семантические концепты для генерации тезаурусов и онтологий, как опорные понятия для контент-анализа СМИ. Практически во всех задачах, связанных с автоматической обработкой текстов, как то аннотирование, индексирование, классификация, машинный перевод, извлечение знаний и т.д., требуется извлечение терминологии.

Для решения указанной задачи разработано большое количество эффективных методов и подходов, среди которых самыми простыми и устойчивыми являются методы, основанные на статистике употребления слов: мера информационной выгоды, критерий хи-квадрат, мера взаимной информации. Большой класс методов образуют контрастные подходы, такие как TF-DCF, Weirdness и т.д. Достаточно перспективными зарекомендовали себя и неконтрастные методы извлечения терминов, которые определяют важность терминов, исходя из внутренних связей документа.

В данной монографии мы рассмотрели широкий спектр контрастных методов извлечения терминологии, а также показали связь задачи извлечения терминов с задачей тематического моделирования. Помимо этого, была представлена модель по извлечению именованных сущностей, в которой слова и символы представляются в векторной форме и подаются на вход блоку двунаправленной LSTM для моделирования контекстной информации. Для кодировки информации на уровне символов используется также двунаправленная LSTM. Особое внимание уделено вопросам погружения слов в линейное векторное пространство.

Примеры из разделов 1-3 были выполнены в экосистеме R, которая в настоящее время обладает развитым аппаратом обработки естественного языка для большого количества языков, в том числе русского и казахского. Примеры из раздела 4

выполнены на языке Python. В Python имеется большое количество библиотек для обработки естественного языка: сверхбыстрой токенизации, анализа, лемматизации текстов и распознавания сущностей. С помощью инструментов Python можно решать задачи создания векторов семантических слов с помощью подхода Word2Vec и реализовывать алгоритмы глубокого обучения.

СПИСОК ЛИТЕРАТУРЫ

1. Frantzi K. T., Ananiadou S., Tsujii J. The c-value/nc-value method of automatic recognition for multi-word terms //International Conference on Theory and Practice of Digital Libraries. – Springer, Berlin, Heidelberg, 1998. – С. 585-604.
2. Heylen K., De Hertog D. Automatic term extraction //Handbook of Terminology. – 2015. – Vol. 1. – 27 pp.
3. Simpson M. S., Demner-Fushman D. Biomedical text mining: a survey of recent progress //Mining text data. – Springer, Boston, MA, 2012. – С. 465-517.
4. Гусева О.И. Критерии терминологичности и корреляция “термин–слово общего языка” //Вісник Маріупольського державного гуманітарного університету. Сер.: Філологія. – 2008. – №. 1.
5. Коршунов А. В. Извлечение ключевых терминов из сообщений микроблогов с помощью Википедии //Труды Института системного программирования РАН. – 2011. – Т. 20.
6. Большакова Е. И., Васильева Н. Э. Терминологическая вариантность и ее учет при автоматической обработке текстов //Одиннадцатая национальная конференция по искусственному интеллекту с международным участием. – 2008. – Т. 2. – С. 174-182.
7. Захаров В. П., Хохлова М. В. Автоматическое извлечение терминов из специальных текстов с использованием дистрибутивно-статистического метода как инструмент создания тезаурусов //Структурная и прикладная лингвистика. – 2012. – №. 9. – С. 222-233.
8. Ефремова Н. Э. и др. Терминологический анализ текста на основе лексико-синтаксических шаблонов //Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог». – 2010. – С. 124-130.
9. Браславский П.И., Соколов Е. А. Сравнение четырех методов автоматического извлечения двухсловных терминов из текста //Компьютерная лингвистика и интеллектуальные технологии: Труды Междунар. конф. Диалог. – 2006. – С. 88-94.
10. Новикова Д. С. Автоматическое выделение терминов из текстов предметных областей и установление связей между ними //Информационно-телекоммуникационные технологии и математическое моделирование высокотехнологичных систем. – 2012.
11. Бессмертный И.А., Нугуманова А.Б., Платонов А.В. "Интеллектуальные системы: учебник и практикум для академического бакалавриата." М.: Издательство Юрайт (2017).
12. Ванюшкин А.С., Гращенко Л.А. Методы и алгоритмы извлечения ключевых слов // Новые информационные технологии в автоматизированных системах. – 2016. – №. 19.

13. Conrado M., Pardo T., Rezende S. A machine learning approach to automatic term extraction using a rich feature set //Proceedings of the 2013 NAACL HLT Student Research Workshop. – 2013. – С. 16-23.
14. Астраханцев, Н. А. "Автоматическое извлечение терминов из коллекции текстов предметной области с помощью Википедии." Астраханце Н. А. Автоматическое извлечение терминов из коллекции текстов предметной области с помощью Википедии //Труды Института системного программирования РАН. – 2014. – Т. 26. – №. 4.
15. Mihalcea R., Tarau P. Textrank: Bringing order into text //Proceedings of the 2004 conference on empirical methods in natural language processing. – 2004.
16. Turing A. M. Computing machinery and intelligence. 1950 //The Essential Turing: The Ideas that Gave Birth to the Computer Age. Ed. B. Jack Copeland. Oxford: Oxford UP. – 2004. – С. 433-64.
17. Straka M., Hajic J., Straková J. UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing //LREC. – 2016.
18. Бессмертный И.А. и др. Метод контрастного извлечения редких терминов из текстов на естественном языке //Научно-технический вестник информационных технологий, механики и оптики. – 2017. – Т. 17. – №. 1.
19. Nugumanova A. et al. A New Operationalization of Contrastive Term Extraction Approach Based on Recognition of Both Representative and Specific Terms //International Conference on Knowledge Engineering and the Semantic Web. – Springer, Cham, 2016. – С. 103-118.
20. Nugumanova A. et al. A contrastive approach to term extraction: Case-study for the information retrieval domain using BAWE corpus as an alternative collection. //Eurasian Journal of Mathematical and Computer Applications. – 2017. – Vol. 5. – P. 73-86.
21. da Silva Conrado M., Pardo T. A. S., Rezende S. O. A Machine Learning Approach to Automatic Term Extraction using a Rich Feature Set //HLT-NAACL. – 2013. – С. 16-23.
22. Ahmad K. et al. University of Surrey Participation in TREC8: Weirdness Indexing for Logical Document Extrapolation and Retrieval (WILDER) //TREC. – 1999.
23. Gillam L., Tariq M., Ahmad K. Terminology and the construction of ontology //Terminology. – 2005. – Т. 11. – №. 1. – С. 55-81.
24. Peñas A. et al. Corpus-based terminology extraction applied to information access //Proceedings of Corpus Linguistics. – 2001. – Т. 2001.
25. S.N. Kim, T. Baldwin, and M-Y. Kan. 2009. An unsupervised approach to domain-specific term extraction. In Australasian Language Technology Association Workshop 2009, page 94-98
26. Basili R. A contrastive approach to term extraction. In Proceedings of the 4th Terminological and Artificial Intelligence Conference (TIA2001).
27. Wong W., Liu W., Bennamoun M. Determining termhood for learning domain ontologies using domain prevalence and tendency //Proceedings of the

sixth Australasian conference on Data mining and analytics-Volume 70. – Australian Computer Society, Inc., 2007. – C. 47-54.
28. Sclano F., Velardi P. Termextractor: a web application to learn the shared terminology of emergent web communities //Enterprise Interoperability II. – Springer London, 2007. – C. 287-290.
29. Astrakhantsev N.A., Fedorenko D.G., Turdakov D.Y. Methods for automatic term recognition in domain-specific text collections: A survey //Programming and Computer Software. – 2015. – Т. 41. – №. 6. – С. 336-349.
30. Kit C., Liu X. Measuring mono-word termhood by rank difference via corpus comparison //Terminology. – 2008. – Т. 14. – №. 2. – С. 204-229.
31. Lopes L., Fernandes P., Vieira R. Estimating term domain relevance through term frequency, disjoint corpora frequency-tf-dcf //Knowledge-Based Systems. – 2016.
32. Z. Zheng, X. Wu, and R. Srihari. Feature Selection for Text Categorization on Imbalanced Data. // ACM SIGKDD Explorations Newsletter, 2004. – Vol. 6. – P. 80-89.
33. G. Forman. An extensive empirical study of feature selection metrics for text classiffication. // J. Mach. Learn. Res., 2003. – P. 1289-1305,
34. Lancaster, Henry Oliver, and E. Seneta. Chi-Square Distribution. John Wiley & Sons, Ltd, 1969.
35. Matsuo Y., Ishizuka M. Keyword extraction from a single document using word co-occurrence statistical information // International Journal on Artificial Intelligence Tools. – 2004. – Vol. 13(01). – Pp. 157-169.
36. Nugumanova A. et al. Using Non-Negative Matrix Factorization for Text Segmentation. // Int. Conference on Mathematical and Informational Technologies. – Beograd, 2016. – Pp. 233-242.
37. Blei D. M. Probabilistic topic models //Communications of the ACM. – 2012. – Vol. 55(4). – Pp. 77-84.
38. Heuboeck A., Holmes J., Nesi H. The BAWE corpus manual. – Technical report, Universities of Warwick, Coventry and Reading, 2007.
39. http://ota.ahds.ac.uk/headers/2539.xml [Электронный ресурс]
40. Ebeling S. O., Heuboeck A. Encoding document information in a corpus of student writing: the British Academic Written English corpus //Corpora. – 2007. – Т. 2. – №. 2. – С. 241-256.
41. Doyle J., Kešelj V. Automatic categorization of author gender via n-gram analysis //The 6th Symposium on Natural Language Processing, SNLP. – 2005. – С. 1-5.
42. Allahyari M., Kochut K. Automatic topic labeling using ontology-based topic models //Machine Learning and Applications (ICMLA), 2015 IEEE 14th International Conference. – IEEE, 2015. – С. 259-264.
43. Park K., Lu X. Automatic analysis of thematic structure in written English //International Journal of Corpus Linguistics. – 2015. – Т. 20. – №. 1. – С. 81-101.
44. Reiter N. et al. Adapting standard NLP tools and resources to the processing of ritual descriptions //ECAI 2010. – 2010. – С. 39.

45. Kilgarriff A. Getting to know your corpus //International Conference on Text, Speech and Dialogue. – Springer Berlin Heidelberg, 2012. – С. 3-15.
46. Manning C. D. et al. Introduction to information retrieval. – Cambridge : Cambridge university press, 2008. – 496 с.
47. https://nlp.stanford.edu/IR-book/ [Электронный ресурс]
48. Da Sylva L. Corpus-based derivation of a “basic scientific vocabulary” for indexing purposes //Journal of Linguistics. – 2009. – Т. 45. – №. 1. – С. 167-201.
49. Feinerer I. Introduction to the tm Package Text Mining in R //2013-12-01]. http://www, dainf, ct. utfpr, edu. br/-kaestner/Min-eracao/RDataMining/tm, pdf. – 2017.
50. D. W. Embley, D. M. Campbell, R. D. Smith, and S. W. Liddl. Ontology-based extraction and structuring of information from data-rich unstructured documents. In Conference on Information and Knowledge Management (CIKM), pages 52–59, 1998.
51. C. Cardie. A case-based approach to knowledge acquisition for domain-specific sentence analysis. In Proceedings of the Eleventh National Conference on Artificial Intelligence, pages 798–803. AAAI Press, 1993.
52. T. Scheffer, C. Decomain, and S. Wrobel. Active hidden Markov models for information extraction. In Proceedings of the International Symposium on Intelligent Data Analysis, 2001.
53. T. Scheffer, S. Wrobel, B. Popov, D. Ognianov, C. Decomain, and S. Hoche. Learning hidden Markov models for information extraction actively from partially labeled text. K¨unstliche Intelligenz, (2), 2002.
54. M. Skounakis, M. Craven, and S. Ray. Hierarchical hidden Markov models for information extraction. In IJCAI, 2003.
55. A. K. McCallum, D. Freitag, and F. Pereira. Maximum entropy Markov models for information extraction and segmentation. In ICML, 2000.
56. A. K. McCallum and D. Jensen. A note on the unification of information extraction and data mining using conditional-probability, relational models. In IJCAI’03 Workshop on Learning Statistical Models from Relational Data, 2003.
57. K.Shaalan. Rule-based approach in Arabic natural language processing. Int.J. Inf.Commun.Technol.3(3). 2010.11–19.
58. Farmakiotou, D., Karkaletsis, V., Koutsias, J., Sigletos, G., Spyropoulos, C. D., & Stamatopoulos, P. (2000, September). Rule-based named entity recognition for Greek financial texts. In Proceedings of the Workshop on Computational lexicography and Multimedia Dictionaries (COMLEX 2000) (pp. 75-78). Patras, Greece
59. Chiticariu, L., Krishnamurthy, R., Li, Y., Reiss, F., & Vaithyanathan, S. (2010, October). Domain adaptation of rule-based annotators for named-entity recognition tasks. In Proceedings of the 2010 conference on empirical methods in natural language processing (pp. 1002-1012). Cambridge, Massachusetts.

60. Ekbal, A., & Bandyopadhyay, S. (2009). A conditional random field approach for named entity recognition in Bengali and Hindi. Linguistic Issues in Language Technology, 2(1), 1-44.
61. Curran, J. R., & Clark, S. (2003, May). Language independent NER using a maximum entropy tagger. In Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 (Vol. 4, pp. 164-167). Edmonton, Canada.
62. Szarvas, G., Farkas, R., & Kocsor, A. (2006, October). A multilingual named entity recognition system using boosting and c4. 5 decision tree learning algorithms. In Proceedings of the International Conference on Discovery Science (pp. 267-278). Berlin, Heidelberg: Springer.
63. Collins, M., & Singer, Y. (1999, June). Unsupervised models for named entity classification. In 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, EMNLP (pp. 100-110). PG County, USA.
64. Kim, J. H., Kang, I. H., & Choi, K. S. (2002, August). Unsupervised named entity classification models and their ensembles. In Proceedings of the 19th International Conference on Computational Linguistics-Volume 1 (pp. 1-7). Association for Computational Linguistics, Taipei, Taiwan.
65. Saha, S. K., Sarkar, S., & Mitra, P. (2008, January). A Hybrid Feature Set based Maximum Entropy Hindi Named Entity Recognition. In Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I (pp. 343-349). Hyderabad, India.
66. Rocktäschel, T., Weidlich, M., & Leser, U. (2012). ChemSpot: a hybrid system for chemical named entity recognition. Bioinformatics, 28(12), 1633-1640.
67. Gang Luo, Xiaojiang Huang, Chin-Yew Lin, and Za-iqing Nie. Joint Entity Recognition and Dis-ambiguation. InProceedings of the 2015 Confer-ence on Empirical Methods in Natural LanguageProcessing, pages 879–888, Lisbon, Portugal. As-sociation for Computational Linguistics. 2015.
68. Lisa F Rau. 1991. Extracting company names from text. In Artificial Intelligence Applications, 1991.Proceedings., Seventh IEEE Conference on. IEEE, volume 1, pages 29–32.
69. Satoshi Sekine and Chikashi Nobata. 2004. Definition, dictionaries and tagger for extended named entity hi-erarchy. In LREC. pages 1977–1980.
70. David Nadeau and Satoshi Sekine. 2007.A sur-vey of named entity recognition and classification. Lingvisticae Investigationes 30(1):3–26.
71. Andrew McCallum and Wei Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume4. Association for Computational Linguistics, pages188–191.
72. Daniel M Bikel, Scott Miller, Richard Schwartz, and Ralph Weischedel. 1997. Nymble: a high-performance learning name-finder. In Proceedings of the fifth conference on Applied natural language processing. Association for Computational Linguistics, pages 194–201.

73. Masayuki Asahara and Yuji Matsumoto. 2003. Japanese named entity extraction with redundant morphological analysis. In Proceedings of the 2003Conference of the North American Chapter of the Association for Computational Linguistics on Hu-man Language Technology-Volume 1. Association for Computational Linguistics, pages 8–15.
74. Sepp Hochreiter and J¨urgen Schmidhuber. Long short-term memory. Neural computation, 9(8):1735–1780. 1997.
75. Junyoung Chung, Caglar Gulcehre, KyungHyun Cho,and Yoshua Bengio. 2014. Empirical evaluation ofgated recurrent neural networks on sequence model-ing.arXiv preprint arXiv:1412.3555.
76. Tomas Mikolov, Martin Karafi´at, Lukas Burget, Jan Cernock`y, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In Interspeech. volume 2, page 3.
77. Martin Sundermeyer, Ralf Schl¨uter, and Hermann Ney.2012. Lstm neural networks for language modeling. In Interspeech. pages 194–197.
78. Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. arXiv preprintarXiv:1409.0473.
79. Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent convolutional neural networks for discourse compositionality. arXiv preprint arXiv:1306.3584.
80. Quan Hung Tran, Ingrid Zukerman, and Gholamreza Haffari. 2017. A hierarchical neural model for learning sequences of dialogue acts. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume1, Long Papers. Association for Computational Linguistics, Valencia, Spain, pages 428–437.
81. Xuezhe Ma and Eduard Hovy. 2016.End-to-end sequence labeling via bidirectional lstm-cnns-crf. arXiv preprint arXiv:1603.01354.
82. Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer.2016. Neural architectures for named entity recognition. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, San Diego, California, pages 260–270. http://www.aclweb.org/anthology/N16-1030.
83. Ronan Collobert, Jason Weston, L´eon Bottou, MichaelKarlen, Koray Kavukcuoglu, and Pavel Kuksa.2011. Natural language processing (almost) fromscratch.Journal of Machine Learning Research,12(Aug):2493–2537.
84. Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural Architectures for Named Entity Recognition. CoRR, abs/1603.01360. 2016.
85. Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. arXivpreprint arXiv:1508.01991.
86. Jason P.C. Chiu, Eric Nichols. Named Entity Recognition with Bidirectional LSTM-CNNs. 2016
87. Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural Architectures for Named Entity Recognition. CoRR, abs/1603.01360. 2016.

88. Marek Rei, Gamal Crichton, and Sampo Pyysalo. 2016.Attending to characters in neural sequence label-ing models. In Proceedings of the 26th International Conference on Computational Linguistics, pages 309–318.
89. Onur Kuru, Arkan Ozan Can, and Deniz Yuret. 2016.Charner: Character-level named entity recognition. In Proceedings of the 26th International Conferenceon Computational Linguistics, pages 911–921.
90. Jason P.C. Chiu, E. Nichols. Named Entity Recognition with Bidirectional LSTM-CNNs. arXiv:1511.08308v5 [cs.CL] 2016.
91. Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty oftraining recurrent neural networks. 2012.
92. K. Greff, R. K. Srivastava, J. Koutník, Bas R. Steunebrink and J. Schmidhuber. LSTM: A Search Space Odyssey. 2017.
93. C. Dyer, M. Ballesteros, W. Ling, A. Matthews, and N. A. Smith.Transition-based dependency parsing withstack long short-term memory. In Proceedingsof ACL-2015 (Volume 1: Long Papers), pages334–343, Beijing, China, July. 2015.
94. Нугуманова А. Б. и др. Обогащение модели Bag-of-words семантическими связями для повышения качества классификации текстов предметной области //Программные продукты и системы. – 2016. – №. 2 (114).
95. Wei Y., Wei J., Xu H. Context vector model for document representation: a computational study //National CCF Conference on Natural Language Processing and Chinese Computing. – Springer International Publishing, 2015. – С. 194-206.
96. Mikolov T., I. Sutskever, K. Chen, G. Corrado, J. Dean. Distributed representations of words and phrases and their compositionality. NAACL HLT. 2013.
97. Deerwester S., Dumais S. T., Furnas G. W., Landauer T. K., Harshman R. Indexing by Latent Semantic Analysis. The American Society for Information Science.1990;41:391-407
98. J. Pennington, R. Socher, C. D. Manning. GloVe: Global vectors for word representation. 2014.
99. Nugumanova, A.,Mansurova, M.,Baiburin, Y.,Alimzhanov, Y. Using non-negative matrix factorization for text segmentation // 2017. CEUR Workshop Proceedings – P.233-242.
100. Nugumanova, A.,Mansurova, M.,Alimzhanov, E.,Zyryanov, D.,Apayev, K. An automatic construction of concept maps based on statistical text mining // Communications in Computer and Information ScienceData Management Technologies and Applications. 4th International Conference, DATA Colmar, France, July 20-22, 2015, Revised Selected Papers. 2016. P. 29-38.
101. 3. Slonim, N., Tishby, N. The power of word clusters for text classification. Proceedings of the 23rd European Colloquium on Information Retrieval Research, Vol. 1. –2001.
102. 5. Самойлов Д. Е., Семенова В. А., Смирнов С. В. Анализ неполных данных в задачах построения формальных онтологий //Онтология проектирования. – 2016. – Т. 6. – №. 3 (21).

103. Досанов Б.Б., Мансурова М.Е., Нугуманова А.Б. Статистикалық әдістер негізінде пәндік аймақ онтологиясын құру // Вестник ВКГТУ №3 Том 1, Часть 3, 2018 г. – С. 112-120.
104. Nugumanova, A., Mansurova, M.,Alimzhanov, E.,Zyryanov, D.,Apayev, K. Automatic generation of concept maps based on collection of teaching materials // DATA 2015 - 4th International Conference on Data Management Technologies and Applications, Proceedings, 2015. - P.248-254.
105. Nugumanova, A.,Mansurova, M.,Alimzhanov, E.,Zyryanov, D.,Apayev, K. An automatic construction of concept maps based on statistical text mining // Communications in Computer and Information ScienceData Management Technologies and Applications. 4th International Conference, DATA Colmar, France, July 20-22, 2015, Revised Selected Papers. 2016. P. 29-38.
106. Nugumanova A., Baiburin Y., Mansurova M., Alimzhanov Ye. An Investigation of the Educational Curriculum with Use of Formal Concept Analysis // Proc. of 10th IADIS International Conference on Information Systems, Budapest, 10-12 of March 2017. – P. 567-574.
107. https://www.tensorflow.org/beta/tutorials/text/word_embeddings [электронный ресурс]
108. https://github.com/davidsbatista/NER-datasets/tree/master/CONLL2003 [электронный ресурс]

СОДЕРЖАНИЕ

Научное издание

Алия Нугуманова
Мадина Мансурова

АВТОМАТИЧЕСКОЕ РАСПОЗНАВАНИЕ ТЕРМИНОВ В ТЕКСТАХ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

Монография

Выпускающий редактор *Г.С. Бекбердиева*
Компьютерная верстка *Г.К. Шаккозовой*
Дизайн обложки: *А. Калиева*

ИБ №
Подписано в печать 15.11.2018. Формат 60х84/16.
Бумага офсетная. Печать цифровая. Объем п.л.
Тираж 500 экз. Заказ № . Цена договорная.
Издательский дом «Қазақ университеті»
Казахского национального университета имени аль-Фараби.
050040, г. Алматы, пр. аль-Фараби, 71, КазНУ.
Отпечатано в типографии издательского дома «Қазақ университеті».