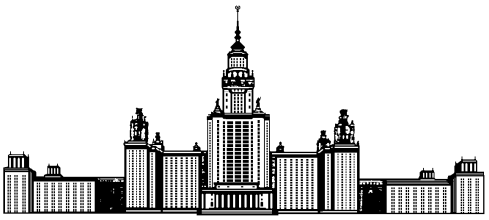


Московский государственный университет имени М.В. Ломоновоа

Факультет вычислительной математики и кибернетики

Кафедра математических методов прогнозирования

Шаталов Николай Алексеевич

Методы обучения без учителя для автоматического выделения составных терминов в текстовых коллекциях

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

Научный руководитель:

д.ф.-м.н., профессор К.В. Рудаков **Научный консультант:**

д.ф.-м.н., доцент К.В. Воронцов

Москва, 2019

Содержание

Аннотация

На сегодняшний день большинство алгоритмов в задачах обработки естественного языка используют алгоритмы с моделью *мешка слов*.В данной модели каждой документ из коллекции текстов выглядит как неупорядоченный набор слов без сведениий о связях между ними. Главным недостатком этой модели является «расщепление» словосочетаний, в следствие чего мы теряем смыслы устойчивых выражений.

В данной работе предложен алгоритм выделения составных терминов, в основу которого положены три подхода: статистический анализ, синтаксический анализ и тематическое моделлирование. Эксперименты показали, что в совокупности три подхода дают максимальную эффективность алгоритма, позволяя не рассматривать все возможные словосочетания в модели «мешка слов».

Итоговый алгоритм может быть применим в различных задачах, например в *вероятностном тематическом моделировании*. Важным показателем качества тематической модели является **интерпретируемость тем**, то есть возможность определить и дать название теме по списку наиболее частотных токенов в данной теме. Добавление в словарь устойчивых словосочетаний, которые являются терминами предметных областей, должно повысить желаемую интерпретируемость тем.

1 Введение

Задачи классификации текстов, ранжирования текстового контента и информационного поиска становятся все более актуальными в растущем мире информации. Для решения этих задач применяется вероятностное тематическое моделирование.

Тематическая модель – это модель, которая принимает на вход коллекцию текстовых документов и определяет, к каким темам относится каждый документ коллекции и какие слова наиболее точно описывают каждую тему. На выходе алгоритм для каждого документа выдает числовой вектор, составленный из оценок степени принадлежности данного документа к каждой из тем, и для каждой темы выдается вектор оценок принадлежности слова этой теме.

Используя эти вектора оценок для тем можно составить список наиболее частотных слов и документов, с помощью которых человек сможет определить семантику этой темы, дать ей адекватное название, определить более общие, более частные или наиболее близкие темы. Это важное свойство называется интерпретируемостью тем. Оно характеризует качество построенной тематической модели и объясняет результаты, выдаваемые пользователю информационной системы.

Для повышения интерпретируемости тем в качестве элементов словаря данной коллекции можно использовать не только отдельные слова, но и словосочетания, являющиеся терминами предметных областей. Поиск таких терминов в тексте является нетривиальной и трудоемкой задачей. Поиск составных терминов также является нетривиальной задачей для человека, так как важно не только понимать синтаксис конкретного языка, но и понимать контекст, в котором был написан текст, а также обладать большим словарным запасом в нужной предметной области. Поэтому встает вопрос об автоматизации извлечения терминов из текстов.

2 Цель работы

Цель данной работы исследовать методы обучения без учителя для выделения составных терминов в текстовых коллекциях, использующие различные подходы, а так же исследование композиции этих методов в роли эффективного алгоритма нахождения словосочетаний, являющихся терминами в данной предметной области.

Для фильтрации и обработки таких словосочетаний требуется определить какими свойствами должен обладать составной термин:

1. **Высокая частотность.** Данное словосочетание много раз встречается в коллекции.

2. **Совстречаемость составляющих слов.** Составной термин должен состоять из слов, неслучайно часто встречающихся вместе.

3. **Полнота.** Словосочетание-термин является максимальной по включению цепочкой слов («обработка естественного языка», но не «обработка естесвенного»)

4. **Синтаксическая связность.** Слова составного термина образуют грамматически правильное словосочетание.

5. **Тематичность.** Термин должен имеет «пиковую» тему в тематической модели.

В данной работе используются три независимых друг от друга подхода, позволяющие проверить каждое из этих свойств:

1. **Статистический анализ** для пунктов 1-3, использующий информацию о частоте и совстречаемости слов в коллекции. Он позволяет находить высокочастотные токены, выделять слова, неслучайно стоящие рядом и штрафовать неполные последовательности слов, входящие как подмножество в другую высокочастотную последовательность неслучайно стоящих рядом слов;

2. **Синтаксический анализ** для пункта 4, использующий информацию о синтаксических связях слов внутри предложений и их частях речи. Данный анализ позволяет выделять синтаксически связанные словосочетания в предложениях;

3. **Тематическая модель** для пункта 5. Она позволяет сопоставить каждому токену (словам и словосочетаниям) распределение тем. Эту информацию можно использовать для выделения «пиковых» словосочетаний, распределение тем которых имеет пик в одной или двух темах.

Для достижения наилучшего качества поиска словосочетаний требуется сравнить несколько альтернативных стратегий оценки и отбора терминов на каждом этапе.

Каждый из множества таких фильтров не может гарантировать высокую точность, но может дополнять другие фильтры, поэтому ставится задача объединения трех подходов отбора терминов в единую технологию. Она подразумевает решение задачи о том, какие фильтры стоит использовать, а какие нет, как строить комбинацию подходов и оценивать итоговое качество.

3 Описание подходов

В качестве моделей, использующих описанные выше подходы, предлагается использовать:

1. ToPMine – для статистического анализа,
2. UDPipe – для синтаксического анализа,
3. BigARTM – как модуль, содержащий в себе тематическую модель PLSA.

3.1 Статистический анализ: ToPMine

Статистический подход позволяет оценивать частоту словосочетания, неслучайность последовательности слов в словосочетании а также штрафовать словосочетание за то, что оно входит в другие.

Частоту N-граммы w в коллекции, состоящую из последовательности слов $w_1, \ldots, w_N$ обозначим $f(w_1, \ldots, w_n) = f(w)$.

В данной работе для статистического анализа используется алгоритм **ToPMine** (Topical Phrase Mining), представленный в статье [4]. Этот алгоритм итеративно сливает слова и фразы в предложении, рассчитывая для каждого слияния оценку зна-

чимости *SignificanceScore* и останавливается, кода для всех возможных слияний значимость меньше заданного порога. В качестве функции значимости используется:

$$SignificanceScore(W_1, W_2) = \frac{f(W_1 \oplus W_2) - f(W_1)f(W_2)/L}{\sqrt{f(W_1 \oplus W_2)}},$$

где W_1, W_2 – сливаемые N-граммы, L – длина коллекции, операция $\oplus$ означает конкатенацию (слияние) двух N-грамм.

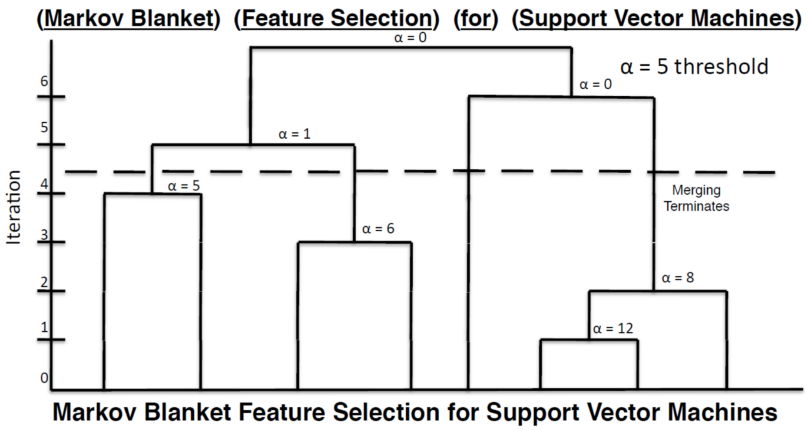


Рис. 1: Пример работы алгоритма ToPMine

Вход: Частоты N-грамм $f(\dots)$, порог α

Выход: Разбиение на N-граммы и значения SignificanceScore для каждой

H ← Куча по возрастанию

Поместить все рядом стоящие пары слов вместе с их SignificanceScore в H

пока размер H > 1

- Best ← максимум по SignificanceScore из H
- **если** Best $\geq \alpha$ **то**
  - New ← Слить(Best)
  - Удалить Best из H
  - Обновить SignificanceScore для New с помощью фраз, находящихся слева и справа
- **иначе**
  - выйти из цикла

Алгоритм 1: ToPMine

3.2 Синтаксический анализ: UDPipe

Синтаксический подход позволяет определить синтаксическую связанность слов в словосочетании. В данной работе для синтаксического анализа используется **UDPipe** [5] – предобученная модель, одной из возможностей которой является распознавание синтаксических связей в предложениях и разметка частей речи слов в предложениях. Данная модель поддерживает 64 различных языка, включая английский, что делает эту модель *универсальной* к задачам на разных языках.

На вход алгоритму подается список прдложений. Для каждого слова в предложениях вычисляется:

- Часть речи слова (NOUN, VERB, ADJ, ...);
- Член предложения (nsubj, conj, cc, ...);
- ID родительского слова (для построения синтаксического дерева).

На рисунке 2 приведен пример работы инструмента UDPipe.

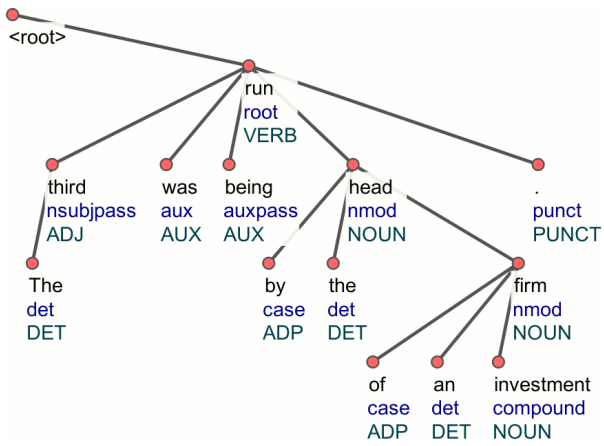


Рис. 2: Пример работы UDPipe

Выделенные части речи слов используется для улучшения качества лемматизации текстов на английском языке алгоритмом, основанном на онтологии **WordNet** [6].

Для других языков лемматизаторы также могут принимать дополнительно на вход часть речи слова для улучшения лемматизации.

Лемматизированная коллекция нужна на вход ToPMine и тематической модели, чтобы улучшить качество соответсвующих алгоритмов.

Используя синтаксическое дерево, для каждого словосочетания-кандидата подсчитывается оценка его синтаксической связанности: максимальное среди возможных расстояний в синтаксическом дереве между словами, составляющее это словосочетание:

$$\mathit{SyntaxScore}(W) = \max_{\substack{1 \le i \le N \\ 1 \le j \le N \\ i \ne j}} \mathit{SyntaxDistance}(w_i, w_j),$$

где $W = (w_1, \ldots, w_N)$ – рассматриваемая N-грамма, $\mathit{SyntaxDistance}(w_i, w_j)$ – расстояние между словами w_i, w_j в синтаксическом дереве.

Таким образом, если одно из слов в словосочетании-кандидате синтаксически не связанно с остальным, *SyntaxScore* это выявит.

Данная технология является свободной и высоко расширяемой на другие языки и синтаксисы. В ней не используются вручную созданные правила и зависимости, а происходит стандартный метод обучения на размеченных данных. С другой стороны это влечет за собой снижение качества распознавания по сравнению с аналогами, где часть правил обработки задана с помощью специалистов, некоторые специфичные случаи (имена собственные, редкие склонения слов) UDPipe может обработать неправильно. Но в случае определения связей в простой структуре термина детальное распознавание не нужно, достаточно только базовых синтаксических зависимостей между словами.

3.3 Тематический анализ: BigARTM

3.3.1 Тематическая модель

Для тематического анализа используется открытая библиотека **BigARTM** [1][2][3]. В качестве тематической модели в данной работе применяется модель PLSA.

Метка части речи	Описание и примеры
ADJ	Прилагательные
ADP	Предлоги
ADV	Наречия
NOUN	Существительные
NUM	Числительные
PRON	Местоимение
PROPN	Имена собственные
VERB	Глаголы
CONJ	Только 7 слов, которые могут быть помечены так: and, but, for, nor, or, so, yet. Обозначает связь между другими словами
AUX	Разные формы глаголов be, have, do, get, используемые для построения грамматических конструкций (например разные времена глаголов)
SCONJ	Слова для связи частей предложения: that, whether, if, when, since, before и т.д.
PUNCT	Пунктуация
DET	Слова the, my, this, some, twenty, each, any и т.д, используемые перед существительными
PART	Части слов: 's, ', not, to (как обозначение инфинитива)
INTJ	Междометия
SYM	Различные символы, отличные от пунктуации

Таблица 1: Части речи, получаемые в процессе работы UDPipe. В верхней части обозначены части речи, которые встречаются в терминах, снизу те, которые встречаться не должны

Пусть D – коллекция текстовых документов, W – словарь токенов, из которых состоят документы. Каждый документ $d \in D$ представляет собой последовательность входящих в него n_d токенов из словаря W. Заметим, что токенами являются не только одиночные слова, но и словосочетания-кандидаты. Приняв гипотезу «мешка слов» о том, что порядок токенов в документе не имеет значения для определения тематики документа можно перейти к определению документа как подмножество $d \subset W$, в котором каждому токену $w \in d$ поставлено в соответствие число n_{dw} ее вхождений в документ d.

Тематическая модель описывает вероятности появления токенов w в документах d при предположении условной независимости:

$$p(w|d,\, t) = p(w|t)p(w|d),$$
$$p(w|d) = \sum_{t\in T} p(w|t)p(t|d) = \sum_{t\in T} \phi_{wt}\theta_{td},$$

где условные вероятности $\phi_{wt} = p(w|t)$, $\theta_{td} = p(t|d)$ являются обучаемыми параметрами модели. В модели они хранятся в виде матриц $\Phi = (\phi_{wt})_{W\times T}$ и $\Theta = (\theta_{td})_{T\times D}$, показывающие условные распределения зависимостей токенов от тем и тем от документов соответственно. Для обучения параметров модели Φ и Θ по коллекции документов D максимизируется логарифм прадоподобия

$$L(\Phi,\, \Theta) = \sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} \to \max_{\Phi,\,\Theta}$$

при ограничениях неотрицательности и нормировки

$$\sum_{w\in W} \phi_{wt} = 1,\ \phi_{wt} \geq 0,\ \ \sum_{t\in T} \theta_{td} = 1,\ \theta_{td} \geq 0.$$

3.3.2 Тематический отбор

На основе полученных параметров Φ и Θ для каждого словосочетания-кандидата считается распределение тем $p(t|w) = \phi_{wt}\frac{p(t)}{p(w)}$. Суть тематического отбора кандидатов заключается в поиске пиковых словосочетаний, то есть таких N-грам w, для которых распределения тематик $p(t|w)$ имеет высокую вероятность только для некоторых тем t из малого подмножества всех тем T.

Такую «пиковость» можно оценивать через отдаленность $p = p(t|w)$ от равномерного $p_{unif}(t) = 1/T$, $t = 1, \ldots, T$. Расстояние между двумя распределениями считается несколькими способами:

- **Дивергенция Кульбака-Лейблера**

  $\mathrm{KL}(pp_{unif}) = \sum_{t \in T} \frac{1}{|T|} \ln \frac{\frac{1}{|T|}}{p(t|w)}$;

- **Дивергенция Йенсена-Шеннона**

  $\mathrm{JS}(pp_{unif}) = \frac{1}{2}\mathrm{KL}(p_{unif}\hat{p}) + \frac{1}{2}\mathrm{KL}(p\hat{p})$,

  где $\hat{p}(t|w) = \frac{1}{2}(p(t|w) + \frac{1}{|T|}$;

- **Сумма степенных функций**, $\gamma > 1$

  $\mathrm{Deg}(p) = \sum_{t \in T} p(t|w)^{\gamma}$.

Чем больше значение данных метрик для N-граммы w, тем больше расстояние от распределения тем данного словосочетания $p(t|w)$ до равномерного, следовательно тем тематичнее данная N-грамма.

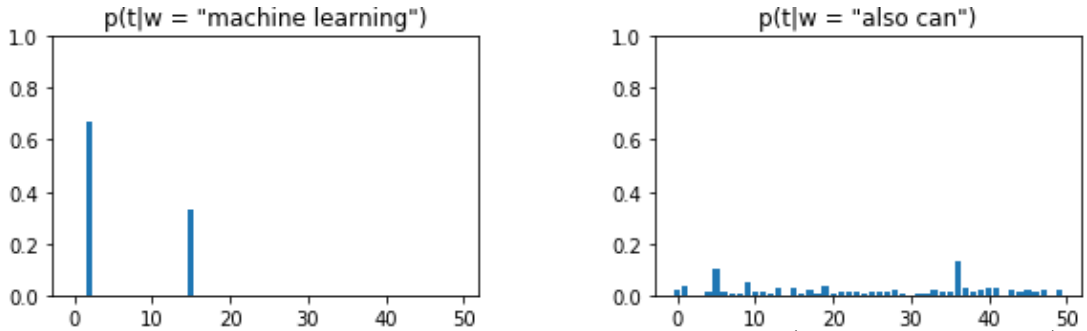


Рис.2: Вероятностное распределение тем для термина (“machine learning”) и «нетермина» (“also can”)

4 Описание признаков

Три описанных подхода по отдельности не могут гарантировать высокую точность, поэтому требуется скомбинировать их для получения лучшего результата. В каждом из них есть несколько стратегий, по которым можно отбирать и оценивать термины.

Данные стратегии генерируют для каждой N-граммы оценку, и чем вы- ше эта оценка, тем уверенее можно сказать, что рассматриваемое словосочетание действительно термин. Таким образом, из этих оценок можно составить признаковое описание N-грамм, по которому предлагается настраивать финальный алгоритм отбора.

В качестве признаков отобраны описанные ранее показатели:

1. Статистический анализ
    - $Sigificance Score$, показывающий, насколько неслучайно встречаются вместе слова, составляющие составной термин,
    - Частота – дополнительный признак, показывающий насколько часто термин встречается в коллекции,
2. Синтаксический анализ
    - $Syntax Score$, показывающий синтаксическую связность составного термина,
3. Тематическое моделлирование
    - значение KL-дивергенции между распределением тем и равномерным распределением,
    - степенная функции с параметром $\gamma = 2$.

5 Вычислительные эксперименты

5.1 Данные

В данной работе исследование проводится на двух датасетах.

- Коллекция статей Википедии
    - $\sim$1.2K документов, $\sim$1.4K слов на документ,
    - 4.8K составных терминов,
    - статьи взяты из шести различных разделов,

  - разметка производилась в полуавтоматическом режиме;

- Коллекция ACL-RD-TEC 2.0
  - 310 документов, ∼1.1K слов на документ,
  - 2K составных терминов,
  - разметка производилась при помощи двух ассесоров;

5.1.1 Wikipedia

Коллекция документов с размеченными словосочетаниями-терминами была собрана на основе статей Wikipedia на английском языке, где за словосочетания-термины были взяты названия гиперссылок, ссылавшиеся на другие статьи в Wikipedia. Было предположено, что словосочетания с гиперссылками ссылаются на статьи, где можно более подробно изучить о предметной области, в которой используется данное словосочетание, а значит оно является термином предметной области. Коллекция состояла из 1200 статей различных тематик, в которых содержится около 640 тысячи слов. Из гиперссылок было выделено ∼4500 уникальных словосочетаний-терминов.

5.1.2 ACL-RD-TEC 2.0

ACL RD-TEC 2.0 включает в себя 1200 уникальных рефератов из сборника ACL Anthology, которые вручную аннотируются для содержащихся в них терминов.

Эти термины помечены несколькими категориями компьютерной лингвистики: технологии, системы, языковые ресурсы, а также метка класса для остальных.

5.2 Эксперимент

Цель эксперимента состояла в проверки предположения о том, что комбинация описанных выше подходов к отбору терминов может выдавать достаточно хорошее качество отсеивания в сравнении с отдельными техниками. Также проверялось то, что каждый из трех подходов существенно влияет на итоговый результат.

5.2.1 Объединение результатов работ компонент

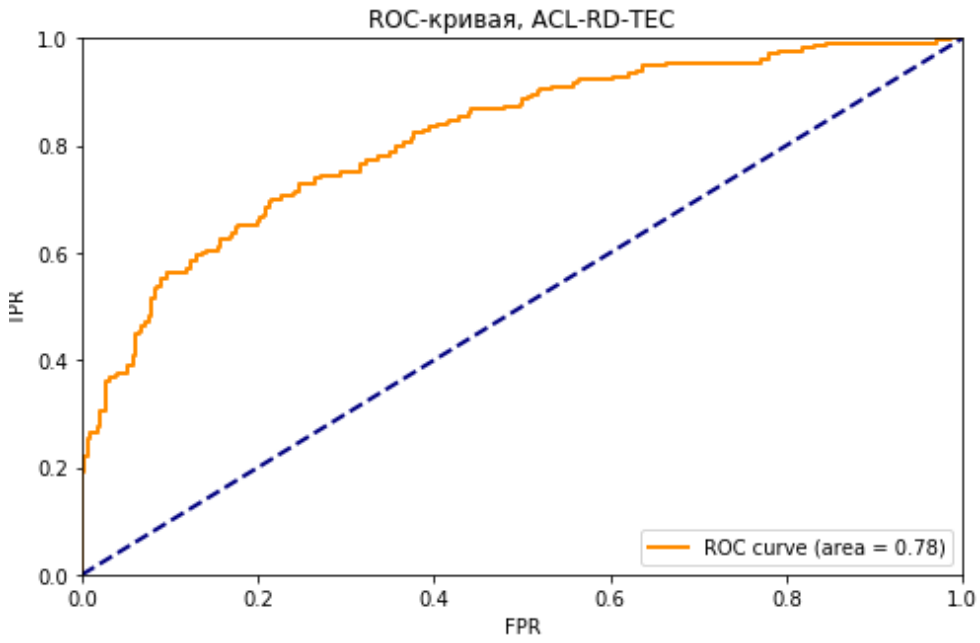
На основе описанного ранее признакового пространства была построена логистическая регрессия, как алгоритм, позволяющий объединить все три подхода.

Model	ACL-RD-TEC 2.0	Wikipedia
[1]	0.7431	0.7031
[2]	0.5148	0.5159
[3]	0.5302	0.5439
[1] + [2]	0.7387	0.7172
[2] + [3]	0.5386	0.5539
[1] + [3]	0.7729	0.7440
[1] + [2] + [3]	0.7801	0.7472

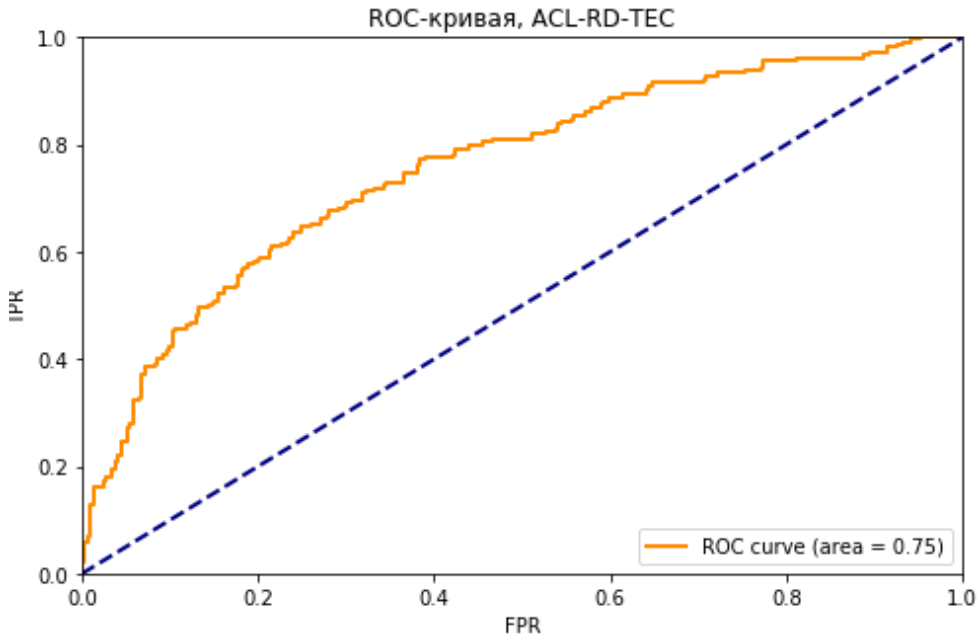
Таблица 2: Результаты экспериментов, точность, [1] – ToPMine, [2] – UDPipe, [3] – BigARTM.

	ACL-RD-TEC		Wikipedia	
	Точность	Полнота	Точность	Полнота
[1]	0.57	0.70	0.53	0.81
[2]	0.36	0.69	0.38	0.80
[3]	0.40	0.70	0.39	0.82
[1] + [2]	0.59	0.70	0.55	0.81
[1] + [3]	0.61	0.71	0.60	0.82
[2] + [3]	0.40	0.70	0.42	0.82
[1] + [2] + [3]	0.66	0.71	0.64	0.82

Таблица 3: Результаты экспериментов, точность и полнота, [1] – ToPMine, [2] – UDPipe, [3] – BigARTM.



ROC-кривая алгоритма с всеми тремя компонентами на датасете ACL-RD-TEC



ROC-кривая алгоритма с всеми тремя компонентами на датасете Wikipedia

5.3 Выводы

Эксперимент на двух коллекциях показал, что можно построить достаточно эффективный алгоритм предсказания терминов на основе трех основных групп признаков: синтаксической, статистическо1 и тематической. Каждая из групп вносит существенный вклад в итоговый ответ.

6 Заключение

В рамках выпускной квалификационной работы было проведенно исследование работы различных методов обучения без учителей для выделения составных терминов в текстовых коллекциях. Также был предложен алгоритм автоматического выделения составных терминов, объединяющий в себе три рассмотренных подхода к выделению терминов. Результат эксперимента показал, что данный алгоритм эффективен.

Список литературы

[1] Vorontsov K. V., Potapenko A. A. Additive Regularization of Topic Models // Machine Learning Journal, Special Issue "Data Analysis and Intelligent Optimization Springer, 2014.

[2] Vorontsov K., Frei O., Apishev M., Romov P., Dudarenko M. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections Analysis of Images, Social Networks and Texts. 2015.

[3] Vorontsov, Konstantin and Frei, Oleksandr and Apishev, Murat and Romov, Peter and Suvorova, Marina and Yanina, Anastasia; Non-Bayesian Additive Regularization for Multimodal Topic Modeling of Large Collections // Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications

[4] Scalable Topical Phrase Mining from Text Corpora / Ahmed El-Kishky, Yanglei Song, Chi Wang et al. // *PVLDB*. – 2014, – Vol. *, no. 3. – Pp. 305-316.

[5] Milan Straka and Jana Strakov a. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Vancouver, Canada, August 2017

[6] Fellbaum C (1998). WordNet: An Electronic Lexical Database. Bradford Books.