Санкт-Петербургский государственный университет

ПЕТРИЦКАЯ Ева Олеговна

Выпускная квалификационная работа

Мультимодальное тематическое моделирование многоязычного корпуса общественно-политических текстов

Уровень образования: бакалавриат

Направление 45.03.02 «Лингвистика»

Основная образовательная программа СВ.5106. «Прикладная, компьютерная и математическая лингвистика (английский язык)»
Профиль «Прикладная, компьютерная и математическая лингвистика (английский язык)»

Научный руководитель:
доцент, кандидат филологических наук, Кафедра математической лингвистики,
Митрофанова Ольга Александровна

Рецензент:
старший преподаватель, Национальный исследовательский университет «Высшая школа экономики»,
Москвина Анна Денисовна

Санкт-Петербург
2022

Аннотация

В работе рассматриваются алгоритмы построения мультимодальных тематических моделей и варианты их обучения на многоязычном корпусе текстов. В процессе создания моделей тестируются разные способы лингвистической обработки корпусов текстов и алгоритмы выделения ключевых выражений, реализуемые с помощью библиотек и инструментов RAKE, scikit-learn, Natasha, NLTK и pymorphy2. Производится интеграция алгоритмов выделения ключевых выражений в тематические модели с целью выбора наиболее подходящего способа построения репрезентативных и качественных моделей коллекции документов. Полученные результаты проходят многоаспектную количественную и качественную оценку. Исследование проводится на параллельном корпусе текстов Организации Объединенных Наций (United Nations Parallel Corpus), находящимся в открытом доступе. Результаты работы могут быть применены в задачах обработки текстов на естественных языках, возникших в ответ на растущую потребность анализа документов, а также в дальнейших исследованиях процессов тематического моделирования.

Ключевые слова: тематическое моделирование, коллекция документов, корпус текстов, многоязычный корпус, ключевые выражения.

Abstract

The paper considers algorithms for constructing multimodal topic models and options for their training on a multilingual corpus of texts. In the process of creating models, various methods of linguistic processing of text corpora and algorithms for highlighting key expressions are tested, using such libraries and toolkits as RAKE, scikit-learn, Natasha, NLTK, and pymorphy2. Key expression selection algorithms are integrated into topic models in order to select the most appropriate way to create representative and qualitative models of a collection of documents. The obtained results undergo a multidimensional quantitative and qualitative assessment. The experiment is conducted on United Nations Parallel Corpus which is available in the public domain. The results of the work can be applied in the tasks of natural language processing that have arisen in response to the growing need for document analysis, as well as in further studies of topic modeling processes.

Keywords: topic modelling, collection of texts, text dataset, multilingual corpus, key phrases.

Оглавление

Введение

Современная компьютерная лингвистика занимается разными аспектами анализа и обработки текстов на естественных языках. В числе актуальных направлений исследований в области компьютерной лингвистики присутствует автоматическое определение тематики документов, исследование лексического состава тем, кластеризации документов по темам, и т.д.

Тематическое моделирование – это способ построения модели коллекции текстовых документов, которая определяет, к каким темам относится каждый из документов. Существует два крупных класса тематических моделей: алгебраические (основанные на счетных векторных моделях текстов) и вероятностные (описываемые вероятностными распределениями) [Blei, Ng, and Jordan 2003; Daud et al. 2010; Lee, Song, Kim 2010; Коршунов, Гомзин 2012].

Вероятностное тематическое моделирование – построение тематической модели с использованием вероятностных методов описания корпусов документов, к числу которых относятся вероятностный латентный семантический анализ PLSA (Probabilistic Latent Semantic Analysis), латентное размещение Дирихле LDA (Latent Dirichlet Allocation) и ряд других. [Hofmann 1999; Jelodar et al. 2019; Воронцов, Потапенко 2012; Potapenko, Vorontsov 2013; Blei 2012].

Алгоритм построения тематической модели получает на входе корпус текстовых документов. На выходе для каждого документа выдаётся числовой вектор, значениями координат которого являются оценки принадлежности данного документа каждой из тем. Размерность этого вектора равняется числу тем и может либо задаваться пользователем в начале процесса моделирования, либо определяться моделью автоматически. Основной

принцип работы таких алгоритмов состоит в том, что тема описывается вероятностным распределением на множестве всех слов текста.

Стандартная тематическая модель учитывает только распределение слов по документам, слов по темам и тем по документам. Добавление дополнительных параметров, характеризующих корпус, позволяет осуществить переход от стандартной тематической модели к мультимодальной. [Andrews et al. 2009; Roller, Im Walde 2013; Nokel, Loukachevitch 2015]. Это происходит, например, в моделях, учитывающих разные типы n-грамм (n-граммная тематическая модель), авторство текстов (автор-тематическая модель, Author-Topic Model, ATM), хронологические рамки корпуса (динамическая тематическая модель, Dynamic Topic Model, DTM), языки корпуса (многоязычные тематические модели) и т.д. [Rosen-Zvi et al. 2012; Sha et al. 2020; Vulić et al. 2013; Zosa et al. 2019].

Большинство современных тематических моделей способны находить в корпусе скрытые темы разной степени значимости, однако связи между словами и документами внутри корпуса представлены не во всей полноте в силу ограниченности базовых тематических моделей. Так, темы обычно представлены униграммами, то есть отдельными, наиболее значимыми для понимания текста словами, но не словосочетаниями [Воронцов 2013; Daud et al. 2009]. Это не всегда помогает точно отразить содержание той или иной темы в документе, особенно если речь идет о специализированных текстах, изобилующих терминами и терминосочетаниями, или о художественных текстах, в которых зачастую используются фразеологизмы, метафоры, имена собственные и устойчивые сочетания, которые нельзя разделять и рассматривать как отдельные слова. При создании тематической модели стоит учитывать, что словосочетания, к какому бы типу они не относились (лексико-грамматические конструкции, коллокации, идиомы и т.д.) играют

большую роль в представлении семантической и синтаксической структуры связного текста.

В противовес униграммным моделям существуют, например, биграммные модели [Wallach 2006; Yan et al. 2013; Huang et al. 2020], но и они, в свою очередь, не могут претендовать на репрезентативность в отношении словарного наполнения корпуса, поскольку генерируемые ими темы состоят исключительно из словосочетаний и не включают отдельные слова.

Для решения данной проблемы были созданы специальные алгоритмы построения мультимодальных тематических моделей, учитывающие биграммы и сочетания нескольких слов как полноценные единицы анализа; результатом работы этих алгоритмов является обобщенная n-граммная тематическая модель [Седова, 2017].

Под n-граммной тематической моделью мы будем понимать модель коллекции текстовых документов, содержащую в себе не только отдельные слова, характерные для данного текста, но и сочетания двух и более слов, представляющих одно понятие или предмет. В алгоритмах такого типа строится общая тематическая модель, объединяющая два метода представления данных – униграммый и n-граммный. В результате в темах присутствуют как отдельные слова, так и словосочетания, которые могут являться, например, ключевыми выражениями, что положительно влияет на репрезентативность модели. Алгоритмы построения n-граммных тематических моделей можно разделить на две группы по признаку последовательности выделения биграммных сочетаний: это делается либо на этапе выделения тем, либо на этапе предобработки текста. [Hu et al. 2008; Wang, McCallum, Wei 2007; Lau, Baldwin, Newman 2013; Нокель, Лукашевич 2015].

В данной работе была предпринята попытка создания мультимодальной n-граммной тематической модели для многоязычного параллельного корпуса текстов.

Тематическое моделирование параллельных многоязычных текстов опирается на алгоритм построения тематических моделей на наборе эквивалентных текстовых данных. Такие тематические модели могут рассматриваться в качестве дополнительного ресурса для систем машинного перевода, и в некоторых случаях могут являться прототипом многоязычного машинного словаря.

Многоязычные тематические модели позволяют эффективно изучать структуру параллельных корпусов текстовых данных, выявлять переводные эквиваленты специфических слов и выражений, а также определять меру расхождения между языками и находить различия в приоритетности тем для анализируемых языков.[Vulic, De Smet, Moens 2011; Mimno et al. 2009].

Тем самым, мультимодальность тематических моделей, созданных в рамках данного исследования, проявляется а) в комбинировании униграмм и n-грамм (коллокаций, ключевых выражений) внутри тем, б) в построении параллельных тем для многоязычного корпуса текстов.

Материалом исследования послужил параллельный многоязычный корпус текстов резолюций ООН[1] от 2000 года, находящийся в открытом доступе и созданный для проведения исследований по машинному обучению и автоматической обработке текстов. Наше внимание сосредоточено на английском и русском корпусах в составе данного многоязычного ресурса. Объем корпуса русскоязычных текстов составляет 2 424 172 словоупотребления, объем англоязычного корпуса – 2 716 043 словоупотребления.

[1] URL: https://opus.nlpl.eu/UNPC.php (дата обращения: 31.05.2022 г.).

Цель настоящего исследования состоит в практической реализации и экспериментальной оценке мультимодального алгоритма вероятностного тематического моделирования, комбинирующего униграммную модель латентного размещения Дирихле (LDA), алгоритмы выделения n-грамм и алгоритмы автоматического выделения ключевых выражений RAKE, и применяемого для анализа параллельного корпуса текстов резолюций ООН на русском и английском языках.

Для достижения данной цели требуется решить следующие задачи:

1) исследовать теоретические основания процедур семантической компрессии текста, прежде всего, вероятностного тематического моделирования и автоматического выделения ключевых выражений;

2) обосновать выбор LDA как базового алгоритма вероятностного тематического моделирования, исследовать реализацию LDA в библиотеке scikit-learn;

3) обосновать выбор линейки алгоритмов автоматического выделения n-грамм – ключевых выражений и коллокаций в тексте, исследовать их реализации на языке Python;

4) сформулировать комбинированную методику расширения стандартной униграммной модели LDA до n-граммной, что предполагает введение в состав униграммных тем биграммных и триграммных лексических конструкций – ключевых выражений, а также n-грамм – коллокаций;

5) подготовить лингвистические данные для проведения экспериментов: произвести предобработку находящихся в свободном доступе корпусов текстов ООН на русском и английском языках.

6) произвести планирование и проведение экспериментов:

a) проведение частеречной разметки корпуса текстов на русском языке;

b) автоматическое выделение ключевых выражений из исследовательских корпусов, сравнение списков ключевых выражений, выделенных для разных языков;

c) разметка выделенных ключевых выражений в корпусах;

d) построение комбинированных n-граммных моделей для корпусов;

7) провести анализ результатов экспериментов.

Объектом исследования является тематическое моделирование русскоязычных и англоязычных параллельных текстов общественно-политического характера, **предметом исследования** – алгоритмы построения мультимодальных n-граммных многоязычных тематических моделей. В работе используются разнообразные методы количественного и лингвистического анализа данных.

Новизна исследования заключается в том, что в данной работе впервые реализован эксперимент по обучению многоязычных n-граммных тематических моделей, совмещающих два способа формирования n-грамм (с учетом коллокаций и с учетом ключевых выражений) и предполагающих выравнивание тем, содержащих словосочетания – кандидаты в переводные эквиваленты.

Теоретическая значимость данной работы состоит в исследовании и суммаризации доступных на сегодняшний день инструментов обработки естественных языков, выявлении особенностей работы этих инструментов с разными языками, а также в изучении комбинированных алгоритмов

тематического моделирования применительно к параллельным многоязычным корпусам текстов.

Практическая значимость результатов работы состоит в создании и описании метода построения репрезентативных n-граммных тематических моделей, способных отразить тематическое содержание объемных лингвистических данных, что является актуальной задачей обработки текстов на естественных языках и отвечает современной потребности крупных организаций в структурировании и компрессии данных. Полученный алгоритм может применяться в задачах изучения параллельных корпусов текстов, машинного перевода, а также в задачах, касающихся обработки и интерпретации больших текстовых данных, например, семантической компрессии текстов, извлечения и исследования текстовой информации.

Глава 1. Интеграция процедур семантической компрессии в мультимодальных тематических моделях

1.1. Процедура семантической компрессии

Важной процедурой анализа данных на естественном языке является **семантическая компрессия**. Семантическая компрессия − построение сжатого семантического представления текста, отражающего его содержательно значимые компоненты [Большакова et al. 2011]. В обработке естественного языка семантическая компрессия представляет собой процесс сжатия текста посредством уменьшения языковой гетерогенности с сохранением семантики текста. Таким образом, в конце процедуры получается текст, в котором идеи и темы из оригинального документа представлены с использованием меньшего набора слов.

Семантическая компрессия может быть представлена на локальном и глобальном уровнях. На локальном уровне семантическая компрессия соответствует таким процедурам, как автоматическое выделение ключевых выражений и n-грамм, аннотирование и реферирование (суммаризация). Подробнее о процедурах семантической компрессии см. работу [Добров 2014]. На глобальном уровне семантическая компрессия проявляет себя в построении тематической модели текстов или коллекции документов, в создании рубрикаторов, в процедурах классификации или кластеризации текстов.

Для осуществления семантической компрессии на локальном уровне существуют различные **алгоритмы поиска и выделения ключевых выражений**, например, RAKE[2], YAKE[3], KEA[4], TF-IDF[5], TextRank[6], DegExt[7] и

[2] URL: https://pypi.org/project/rake-nltk/ (дата обращения: 31.05.2022 г.).
[3] URL: https://pypi.org/project/yake/ (дата обращения: 31.05.2022 г.).
[4] URL: https://github.com/EUMSSI/KEA (дата обращения: 31.05.2022 г.).

др. Выделение ключевых выражений – как униграмм, так и сочетаний из нескольких слов – помогает не только в определении общей тематики документов, что необходимо в задачах информационного поиска, но также в выявлении наиболее значимых фрагментов словаря корпуса, в уточнении тематических моделей, насыщении их лексическими единицами, характерными для текстов заданного типа [Браславский, Соколов 2008; Ягунова, Пивоварова 2010]

Задача **аннотирования** включает в себя определение тематики документов, выделение ключевых (по темам) слов и фраз с учетом смысла, поиск предложений, содержащих ключевые слова и фразы, и синтез на этой основе связного текста, состоящего из значимых для содержания документа фраз и предложений, отражающих основные темы текста, подробнее см. [Вознесенская, Леднов 2018]. В таком случае результатом работы являются не ключевые слова, а главные идеи и проблемы, затронутые в тексте.

Реферирование (суммаризация) подразумевает сжатие больших объемов текста до связного краткого содержания, в котором выражены лишь основные идеи. Существует два типа суммаризации: 1) экстрактивный подход, предполагающий выделение из текста наиболее значимых предложений и склеивание их в новый текст (что не гарантирует его связность), 2) абстрактивный подход, основывающийся на выделении основных идей и генерации нового текста “с нуля” [Белякова, Беляков 2020; Нестерова, Герте 2013].

На глобальном уровне семантическая компрессия может осуществляться с помощью **тематического моделирования** – выделения тем

---

[5]URL:https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html (дата обращения: 31.05.2022 г.).

[6] URL: https://analyticsindiamag.com/guide-to-nlps-textrank-algorithm/ (дата обращения: 31.05.2022 г.).

[7] URL: https://link.springer.com/chapter/10.1007/978-3-642-18029-3_13 (дата обращения: 31.05.2022 г.).

и соответствующих этим темам слов, которые представляют содержание какого-либо документа или коллекции текстов. Построение тематической модели можно сравнить с кластеризацией текстов, только в данном случае один документ может относиться сразу к нескольким кластерам – темам, соответственно, одна тема может содержать в себе слова и выражения из разных документов, выявляя, таким образом, связь текстов в коллекции [Воронцов 2020]. Наибольшее применение в современных приложениях находят подходы, основанные на байесовских сетях, вероятностных моделях на ориентированных графах. Вероятностные тематические модели — это относительно молодая область исследований в теории самообучения. Работы по исследованию процедуры вероятностного тематического моделирования представлены в целом ряде работ: хронология разработок и основные алгоритмические решения описаны, например, в обзорах [Воронцов 2013; Daud et al. 2009]. Одним из первых был предложен вероятностный латентно-семантический анализ (PLSA) [Hoffman 1999], основанный на принципе максимума правдоподобия, как альтернатива классическим методам кластеризации, основанным на вычислении функций расстояния. Вслед за PLSA был предложен метод латентного размещения Дирихле (LDA)[Blei et al. 2003] и его многочисленные обобщения. В данной работе рассматривается обогащение униграммной модели LDA ключевыми выражениями и n-граммами. Работы по внедрению n-грамм в тематическую модель проводились рядом исследовательских групп, см. например, [Nokel, Loukachevich 2016; Sedova, Mitrofanova 2017; Moskvina, Sokolova, Mitrofanova 2018].

Кластеризация и классификация текстов – процесс структурирования текстов в корпусе без использования обучающих данных или с использованием таковых[8]. Тексты объединяются в класс на основании

[8] Учебник по машинному обучению ШАД. URL: https://ml-handbook.ru/ (дата обращения: 31.05.2022 г.).

лексического сходства, семантической смежности и ассоциативности. Наиболее эффективны подходы, предполагающие идентификацию семантических классов, например, рубрикация: при таком подходе сравниваются не слова, а эталоны, рубрики, к которым тексты принадлежат см. работу [Добров 2012]. В результате документы на одинаковую тему, в которых используются разные слова, объединяются в один класс (подробнее см. [Воронцов 2015]).

1.2. Вероятностное тематическое моделирование

Построение тематических моделей является одним из актуальных направлений современной компьютерной лингвистики. Тематическое моделирование как вид статистических моделей для нахождения скрытых тем, встреченных в коллекции документов, нашло своё применение в таких областях обработки естественного языка, как извлечение информации и анализ данных. Исследователи используют различные тематические модели для анализа текстов, текстовых архивов документов, для анализа изменения тем в наборах документов.

Вероятностное тематическое моделирование – метод построения тематической модели коллекции текстов, основывающийся на применении математических и статистических данных. В вероятностной тематической модели темы представляются в виде распределений на множестве слов, а документы – в виде распределений на множестве тем [Воронцов, Потапенко 2013]. Так, одна тема может содержать в себе слова и ключевые выражения из нескольких документов, и наоборот, один документ может относиться одновременно к нескольким темам. По определению условной вероятности, формуле полной вероятности и гипотезе условной независимости, распределения слов в документе, p(w|d), описывается вероятностной смесью распределения слов на темах $\phi_{wt} = p(w|t)$ с весами $\theta_{\text{td}} = p(t|d)$. Введем формальное определение тематического моделирования:

$$p(w \mid d) = \sum_{t \in T} p(t \mid d)\, p(w \mid t).$$

Данная формула описывает процесс порождения коллекции документов по известным распределениям $p(w|t)$ и $p(t|d)$. Задача же тематического моделирования состоит в обратном : по некоторой коллекции документов D надо вычислить параметры ϕ_{wt} и θ_{td}.

В основе вероятностного подхода лежит идея о том, что появление слова в документе, принадлежащем какой-либо теме, не зависит от самого документа. Это предположение называется гипотезой условной независимости. Если нам известны распределения тем по документам и слов по темам, то можно описать процесс порождения коллекции документов с помощью вероятностной модели на основе этих распределений. Построение тематической модели является противоположной задачей: по известной нам коллекции документов нужно восстановить породившие её распределения.

Классической версией вероятностного тематического моделирования является алгоритм PLSA (Probabilistic Latent Semantic Analysis), опубликованный в 1999 году Т.Хоффманом [Hoffmann 1999]. Вероятностный латентно-семантический анализ основан на смешанном разложении, в свою очередь берущем своё начало из модели скрытых классов. Модель скрытых классов соотносит множество наблюдаемых случайных переменных с множеством скрытых переменных. Скрытыми называются переменные, которые содержатся в наборе данных, но могут быть выведены только из наблюдаемых переменных. В рамках модели скрытых классов существует процедура анализа скрытых классов, которая подразумевает поиск скрытых групп или объединений в наборе данных. Работая с каким-либо набором данных, модель скрытых классов находит неявные группы и объединения, в которые можно собрать эти данные, при этом элементы внутри класса или группы статистически независимы друг от друга. В тематическом

моделировании такой анализ означает соотнесение набора слов – наблюдаемых переменных – с какой-либо темой − скрытой переменной.

Расширением модели PLSA является алгоритм LDA (Latent Dirichlet Allocation) [Blei et al. 2003]. Это порождающая статистическая модель, в которой за основу распределения тем по документам берется распределение Дирихле.

Данная модель объясняет результаты наблюдений с помощью неявных групп, благодаря чему возможно выявление причин сходства некоторых частей данных. Например, если наблюдениями являются слова, собранные в документы, утверждается, что каждый документ представляет собой смесь небольшого количества тем и что появление каждого слова связано с одной из тем документа. Алгоритм LDA доступен в многих библиотеках, например, scikit-learn[9], gensim[10], Mallet[11], Bigartm[12] и tomotopy[13]. Существуют библиотеки, интегрирующие алгоритмы тематического моделирования с контекстуализированными векторными моделями, например, CTM[14], Bertopic[15]. В данной работе используется реализация алгоритма LDA в библиотеке scikit-learn. Мы выбрали данную реализацию из соображений практического удобства и возможности обращаться к другим компонентам библиотеки для решения более частных задач и задействования дополнительных функций при построении тематической модели. Так, реализация LDA scikit-learn позволяет провести дополнительную процедуру выделения биграмм и триграмм на этапе построения самой модели, что

[9] URL:https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html (дата обращения: 31.05.2022 г.).
[10] URL: https://pypi.org/project/gensim/ (дата обращения: 31.05.2022 г.).
[11] URL: https://mimno.github.io/Mallet/ (дата обращения: 31.05.2022 г.).
[12] URL: https://bigartm.readthedocs.io/en/stable/api_references/python_interface/lda_model.html (дата обращения: 31.05.2022 г.).
[13] URL: https://bab2min.github.io/tomotopy/v0.12.2/en/ (дата обращения: 31.05.2022 г.).
[14] URL: https://github.com/MilaNLProc/contextualized-topic-models (дата обращения: 31.05.2022 г.).
[15] URL: https://arxiv.org/abs/2203.05794 (дата обращения: 31.05.2022 г.).

может значительно улучшить результат моделирования тем с учетом n-грамм и коллокаций.

В данной работе используется один из самых распространенных методов построения тематических моделей – **метод латентного размещения Дирихле (LDA).** LDA – это порождающая вероятностная модель, которая рассматривает документ как коллекцию латентных тем, определяющихся вероятностным распределением на множестве слов текста. Слова, в свою очередь, определяются наличием в документе той или иной латентной темы. Этот алгоритм основан на априорном распределении Дирихле и использует в своей работе модель “bag-of-words”, которая рассматривает слова из документов вне их линейной последовательности: для этой модели важна лишь частотность тех или иных слов, но не их окружение. Таким образом, обнаруживаются неявные связи между словами и документами с учетом полисемии. На рисунке 1 представлена схема работы алгоритма LDA. Наблюдаемыми переменными являются слова в документах, на схеме эти вершины выделены серым цветом. Остальные переменные являются скрытыми.

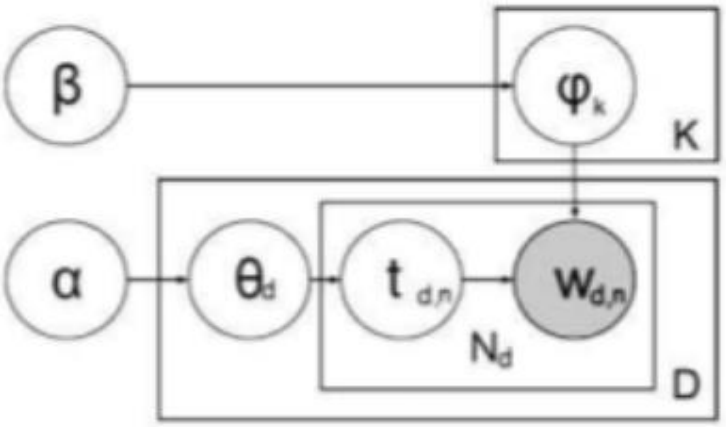

Рисунок 1. Схема работы алгоритма LDA

В базовом алгоритме LDA есть возможность задания нескольких параметров: количество выделяемых тем, количество слов в каждой теме,

количество итераций алгоритма. На выходе алгоритм выдает список тем, представленных самыми частотными словами – униграммами. LDA в своей базовой конфигурации не рассматривает устойчивые сочетания и n-граммы как единые элементы и не включает их в итоговую выборку.

Посмотрим на принцип работы алгоритма LDA. Каждый документ из коллекции генерируются независимо в два этапа:

1) для документа выбирается его распределение по темам $\theta_{td} = p(t|d)$;
2) для каждого слова в документе выбирается тема θ_{td} и слово из распределения $\phi_{wt} = p(w|t)$.

Алгоритм LDA подразумевает что параметры θ_{td} и ϕ_{wt} распределены следующим образом: $\theta \sim Dir(\alpha)$, $\phi \sim Dir(\beta)$, где α и β являются гипер параметрами, или векторами-параметрами распределения Дирихле. Обычно параметр α принимается равным 50/T, где T — число тем, а β принимается равным 0.1 [Daud 2010].

Одним из преимуществ модели LDA является ее интерпретация распределения тем. В отличие от PLSA, ещё одной популярной модели тематического моделирования, LDA опирается на представление о том, что в документах присутствуют лишь несколько мелких тем, которые представлены не всеми словами в документе, а лишь небольшой их частью. Таким образом, предполагается, что в каждом документе содержится несколько тем, выраженных несколькими ключевыми словами, и разные темы могут одновременно присутствовать в разных документах. Это предположение позволяет избавиться от многих частотных, но малозначимых для тематики текстов слов и найти скрытые темы в разных документах. На практике этот подход оказывается более эффективным в отношении распознавания тематически важных слов и соотнесения слов с темами.

1.3. Автоматическое выделение ключевых выражений

Проведение процедуры семантической компрессии на локальном уровне подразумевает нахождение ключевых выражений – сочетаний слов, имеющих наибольшую значимость для понимания текста или имеющих наибольшую семантическую ценность. Ключевые выражения не включают в себя знаки пунктуации, служебные слова, местоимения и специальные символы, поэтому процедура выделения ключевых слов требует также обработки исходного корпуса с помощью стоп-словаря. Существует три основных подхода к выделению ключевых выражений :

1) статистический (TFxIDF, логарифмическая функция правдоподобия LogLikelihood, критерий Хи-квадрат, C-Value и др., графовые алгоритмы типа PageRank, TextRank, DegExt, алгоритмы, основанные на моделях распределенных векторов и т.д.),
2) лингвистический (использование морфологической аннотации, лексико-грамматических шаблонов, данных о синтаксических группах; словари типа WordNet и т.д.),
3) гибридный (KEA, RAKE и т.д.).

Также при автоматическом выделении ключевых выражений и словосочетаний могут использоваться дополнительные источники информации, например, тематические и фоновые корпусы, обучающие выборки с уже размеченными в них ключевыми выражениями и стоп-словари.

Все эти алгоритмы, несмотря на своеобразие, включают следующий набор операций:

1) выделение фраз-кандидатов;
2) подсчет весов выделенных фраз;
3) ранжирование весов и выделение наиболее значимых фраз.

В данной работе для выделения ключевых выражений мы использовали возможности алгоритма RAKE, который реализован на основе программной библиотеки NLTK. Изначально алгоритм создавался для анализа текстов на английском языке, для работы с русским языком RAKE был адаптирован в работах [Moskvina et al. 2017; Sedova, Mitrofanova 2017; Sokolova, Mitrofanova 2017; Sokolova, Moskvina, Mitrofanova 2017]. В рамках данного исследования мы оценивали его эффективность в отношении возможности расширения униграммных моделей n-граммами и качество построенных с выделенными им выражениями тематических моделей корпусов текстов на русском и английском языках.

RAKE — это гибридный алгоритм выделения ключевых слов и выражений в тексте, использующий как лингвистический (синтаксические алгоритмы и словарные данные), так и статистический теоретический аппарат. Работа алгоритма основывается на предположении о том, что ключевые выражения представляют собой не только отдельные слова, но и фразы. Такие фразы могут состоять из нескольких слов, не включая знаки пунктуации, служебные слова и слова, не несущие ярко выраженного лексического значения.

Алгоритм RAKE не требует предварительного обучения на размеченных данных и не зависит от размера текстов. Он также не требует подключения других специализированных корпусов, тезаурусов или словарей и может использоваться на динамических коллекциях, подстраиваясь под их стилистику. Алгоритм выделяет как униграммы, имеющие наибольшую семантическую и смысловую значимость, так и последовательности слов вплоть до пяти – шести элементов, не требуя при этом настройки специальных параметров. Таким образом, за один рабочий сеанс выделяются все значимые слова и выражения. Данный алгоритм использовался нами в предыдущей работе и показал свою эффективность при сравнении с другими алгоритмами выделения ключевых слов (NLTK

Collocations и YAKE). Более того, эффективность RAKE в задаче поиска ключевых слов и выражений была доказана при сравнении его с другим алгоритмом TextRank [Sokolova et al 2018].

На этапе генерирования фраз-кандидатов алгоритм RAKE [Moskvina et al. 2017] выделяет в фрагменты текста по знакам пунктуации и стоп-словарю. Внутри синтагм мы не ожидаем увидеть знаки препинания, но для верности можно поставить фильтр готовых фраз. Затем для каждого слова рассчитывается вес на основе его длины и частотности. Вес фразы-кандидата определяется как сумма весов составляющих ее слов. Для улучшения результата выделения ключевых выражений алгоритм допускает введение дополнительных параметров – задание грамматических характеристик фразы, то есть, введение шаблонов согласования типа ADJ+N, N+Adj+N и правил из грамматики. На выходе алгоритм выдает ранжированный список фраз – от самых длинных (пентаграмм и квадриграмм) до одиночных слов – униграмм.

1.4. Описание комбинированной методики построения n-граммной модели LDA

Комбинированная методика расширения стандартной модели LDA заключается в добавлении биграмм и триграмм в корпус текстов, на котором реализуется работа алгоритма. В базовых алгоритмах ТМ темы представлены униграммами, но это не всегда удобно и верно, так как в текстах зачастую встречаются устойчивые сочетания, несущие большую семантическую нагрузку и представляющие потенциальную значимость для описания темы. Включение таких сочетаний в темы позволило бы улучшить точность тематического моделирования какой-либо коллекции текстов.

На сегодняшний день существует два основных подхода расширения тематической модели биграммами и коллокациями: первый подразумевает выделение сочетаний слов одновременно с темами, второй − выделение коллокаций на этапе обработки исходного текста. Модели первого типа

имеют логичное теоретическое обоснование, но могут снизить результативность алгоритма ТМ или же заметно усложнить его. Первый тип моделей представлен биграммной тематической моделью и скрытой тематической марковской моделью с применением LDA:

1) биграммная тематическая модель (Bigram Topic Model): данная модель является иерархической порождающей моделью, и при её работе в качестве основополагающего используется предположение о том, что появление определенного слова зависит исключительно от его соседей;

2) скрытая тематическая марковская модель с применением LDA (HMM-LDA) тоже выделяет биграммы на этапе выделения тем; предложение разбивается на функциональные слова, порожденные с помощью скрытой марковской модели, и на термины, генерируемые моделью LDA; состоит из последовательности слов, тематических переменных и бинарных классификаций, которые показывают, создают ли данное и предыдущее слово словосочетание.

Ко второму типу можно отнести модели с использованием меры T-Score, PLSA-SIM, и PLSA-ITER:

1) модель с использованием меры T-Score [Lau et al, 2013]: алгоритмы выявляются на этапе предварительной обработки текста, полученные словосочетания добавляются в корпус в виде единого токена, и в процессе построения тематической модели они рассматриваются наравне с другими униграммами как токены;

2) PLSA-SIM, PLSA-ITER – усовершенствованные модели вероятностного латентного семантического анализа, предложенные в работах [Nokel 2016; Nokel, Loukachevitch 2000]; PLSA-SIM основана на введении слоя скрытых переменных; в модели PLSA-ITER биграммы составляются из самых частотных слов в теме.

Второй подход, подразумевающий выделение биграмм на этапе обработки текстов, является более простым, так как не требует настройки

большого количества параметров. Алгоритм, рассмотренный в данной работе, принадлежит ко второму типу.

Вопросу расширения стандартной процедуры тематического моделирования с выделением биграмм на этапе обработки текста посвящены работы [Sedova, Mitrofanova 2017; Nokel, Loukachevitch 2000; Loukachevitch, Nokel 2017]. Алгоритм построения тематической модели для LDA с учетом биграмм и коллокаций был предложен в работе [Sedova, Mitrofanova 2017]. Исследование проводилось на модели латентного размещения Дирихле, которая используется и в данной работе. Биграммы выделялись на этапе обработки текста, а затем добавлялись в корпус, с которым работал алгоритм LDA. В результате работы были выделены темы, содержащие в себе как униграммы, так и сочетания слов.

В нашем исследовании процесс расширения модели n-граммами проходил в соответствии с алгоритмом, представленным в работе [Sedova, Mitrofanova 2017] с незначительными изменениями. Этот же алгоритм применялся нами в предыдущих работах [Петрицкая 2020; 2021].

На первом этапе осуществляется морфологическая разметка текстов: с помощью морфологического анализатора pymorphy2[16] в текстах проводится частеречная разметка и выделяются морфологически связанные сочетания слов, например, прилагательное и существительное, существительное и существительное в родительном падеже, глагол и существительное в винительном падеже, и т.д. В отдельные группы выделяются предлоги, союзы, частицы, наречия, глаголы без дополнений и инфинитивные формы.

После проведения частеречной разметки тексты подаются алгоритму выделения ключевых слов и выражений RAKE, который находит наиболее значимые униграммы и коллокации для каждого текста с учетом ранее выделенных морфологических групп. Результаты работы заносятся в отдельный файл для каждого текста из коллекции.

[16] URL: https://pymorphy2.readthedocs.io/en/latest/ (дата обращения: 31.05.2022 г.).

На третьем шаге проводится фильтрация выделенных выражений и добавление их в исходный корпус. Из всех выделенных на предыдущем шаге слов и коллокаций выбираются только биграммы и триграммы; сочетания из четырех и более слов в корпус не добавляются, чтобы не создавать избыточности коллокаций в текстах. Выбранные биграммы и триграммы проходят процесс лемматизации с помощью парсера Natasha[17] и помечаются специальным символом '_'. В таком виде выражения добавляются в исходный корпус, где уже учитываются как единый токен.

Далее проводится подготовка текстов для работы с алгоритмом LDA: с помощью стоп-словаря удаляются местоимения, междометия, частицы, знаки пунктуации, цифры, специальные символы и слова на нерелевантном языке, все тексты приводятся к нижнему регистру и очищаются от ненужных отступов и пробелов.

После подготовки коллекция текстов подается алгоритму латентного размещения Дирихле. В рамках данной работы мы опробовали несколько реализаций LDA: с дополнительным выделением биграмм и триграмм, с дополнительным выделением биграмм, а также алгоритм LDA, расширенный только коллокациями RAKE. В процессе выделения тем и подсчета самых репрезентативных слов модель учитывает не только униграммы, но и выделенные в текстах n-граммы, включая наиболее значимые выражения в итоговую выборку; также первые два варианта подразумевают выделение дополнительных коллокаций на этапе поиска тем. В итоге алгоритм выделяет встретившиеся в коллекции темы и относящиеся к ним слова, в которые входят биграммы и триграммы. То, что строится n-граммная модель, является проявлением ее мультимодального характера. Особенностью n-граммных моделей, строящихся в рамках нашего исследования, является то, что на разных этапах анализа сборка n-грамм производится и по формально-количественному критерию (коллокации с заданной структурой), и по

[17] URL: https://github.com/natasha/natasha (дата обращения: 31.05.2022 г.)

содержательному (наиболее значимые, важные для текста ключевые выражения определенной длины).

1.5. Многоязычные тематические модели

Изначально алгоритм тематического моделирования подразумевал анализ и создание модели тем для коллекции одноязычных текстов, однако с появлением и возрастанием важности параллельных корпусов в прикладных лингвистических задачах исследователи заинтересовались возможностью создания тематических моделей для коллекции текстов на двух и более языках. Такие модели получили название многоязычных и в настоящее время строятся в основном на двух типах корпусов: параллельных и сопоставимых. В данной работе рассматривается тематическая модель для параллельного корпуса текстов.

Тематическое моделирование параллельных корпусов текстов – достаточно новое направление, которое может значительно улучшить работу как с параллельными корпусами, так и способствовать в решении других лингвистических задач. Например, в рамках тематического моделирования параллельных корпусов можно облегчить подготовку текстовых данных для инструментального исследования различных языков. Более того, многоязычные тематические модели могут использоваться в качестве дополнительного ресурса для систем машинного перевода, а в некоторых случаях могут служить прототипом многоязычного машинного словаря [Митрофанова 2019].

Параллельный корпус – это коллекция текстов на одном языке вместе с их выровненным переводом на другой язык (или языки) [Добровольский, Кретов, Шаров 2005]. В таких корпусах каждому документу и предложению для одного языка соответствует документ и предложение второго языка. Ценность параллельного корпуса растет вместе с его размером и количеством языков, для которых существуют переводы. В то время как параллельные

корпусы для некоторых языков существуют в изобилии, для большинства других языковых пар параллельных корпусов мало или вообще нет. Поэтому особой ценностью обладают коллекции текстов, включающие в себя не только самые распространенные языки мира, но и более редкие. В рамках данной работы мы сосредоточились на анализе русско-английской пары коллекций текстов.

К более всеобъемлющим коллекциям можно отнести, например, корпус текстов слушаний Европарламента[18], содержащий параллельные тексты на языках некоторых стран-участниц Европейского союза. Всего корпус охватывает одиннадцать языков, или десять параллельных корпусов для каждого языка. В работе [Koehn 2005] этот корпус был задействован в качестве тренировочных данных для алгоритма статистического машинного перевода и показал хороший результат в рамках задачи по улучшению качества автоматизированного перевода.

В исследовании [Mimno et al. 2009] данные из корпуса параллельных текстов слушаний Европарламента и корпуса сопоставимых текстов Википедии на двенадцати языках были использованы для экспериментов по построению многоязычных тематических моделей. В результате эксперимента были получены темы для каждого языкового корпуса, а проведенный после этого сопоставительный анализ состава тем выявил, что в них входят преимущественно переводные эквиваленты. Это показывает потенциал использования тематических моделей параллельных корпусов для решения задач машинного перевода и составления специализированных словарей. Важно, что по входящим в состав тем словам и выражениям можно также судить о специфике употребления самостоятельных и служебных слов в разных языках [Mimno et al. 2009].

[18] URL: https://www.statmt.org/europarl/ (дата обращения: 31.05.2022 г.).

Еще один пример – корпус документов Евросоюза Acquis Communautaire[19], находящийся в открытом доступе и содержащий параллельные тексты на двадцати двух языках стран-участниц Европейского союза. Создатели корпуса также отмечают большую ценность таких материалов в области изучения и обработки естественных языков и своими усилиями показывают готовность к содействию языковым исследователям в данной сфере. Этот корпус является самым большим параллельным корпусом из существующих, как по размеру, так и по количеству включенных в него языков, а также по количеству редких сочетаний языковых пар (например, мальтийско-эстонский, словенско-финский и т.д.).

Помимо разнообразия источников параллельных и сопоставимых текстов существуют ещё и разные способы создания многоязычных тематических моделей. Многоязычные тематические модели могут строиться как для одной коллекции всех многоязычных текстов, так и отдельно для каждого языка в корпусе. Результаты в таком случае будут отличаться и отражать разные аспекты и особенности коллекции. Например, в работе [Lind et al. 2019] был реализован алгоритм построения одной тематической модели на семь тем для общего корпуса текстов об эмиграции на семи языках. В такой модели каждая тема отображала не конкретную тематическую область, а представляла язык документов в корпусе. Поскольку алгоритм LDA, задействованный в данном эксперименте, учитывает статистику совместной встречаемости слов, в результате были получены группы токенов, принадлежащих к одному языку, ведь слова одного языка в документах встречаются только вместе со словами этого же языка. Таким образом, полученная модель не отражала присутствующие в коллекции темы, а позволяла отождествлять представленные в ней языки.

[19] URL: https://joint-research-centre.ec.europa.eu/language-technology-resources/jrc-acquis_en) (дата обращения: 31.05.2022 г.).

Другой вариант построения стандартной тематической модели на многоязычном корпусе подразумевает построение отдельной тематической модели для каждого языка. Такой алгоритм реализован в исследовании [Zheng et al. 2014], где авторы сравнивают тематики постов популярных японских и китайский блогеров. Для каждой коллекции строится тематическая модель LDA, затем каждому документу присваивается одна тема, являющаяся доминирующей в этом документе. Полученные результаты сравниваются, и на основании выделенных тем выводится гипотеза о тематической близости двух текстовых коллекций. Таким образом, анализ многоязычного корпуса проводится при помощи построения двух независимых тематических моделей для каждого языка в коллекции. Этот способ, в силу своей простоты и большей практической ценности в поставленной задаче, применяется и автором в данной работе.

Помимо этого существует множество других подходов к моделированию тем корпусов параллельных текстов, подробнее см. работу [Тюрева 2021]. В частности, подход при котором алгоритм тематического моделирования работает сразу на всей коллекции текстов, генерируя при этом параллельные темы для двух языков. Такие модели могут требовать дополнительных решений, например, выравнивания на уровне слов, перевода текстов на язык-посредник, опоры на словари или модели на тематически сопоставимых корпусах.

В данной работе был выбран способ построения отдельной тематической модели для каждого корпуса текстов с последующим сопоставлением тем. Такой подход поможет избежать смешения языков в темах и более точно проследить процесс выделения и использования биграмм и триграмм в процессе создания модели и на этапе предобработки текстовых данных. Более того, такой подход помогает опробовать разные варианты реализации алгоритма LDA с использованием дополнительных параметров выделения коллокаций. Таким образом, мультимодальность

алгоритма в данной работе обеспечивается многоязычием лингвистических данных, комбинированием разных способов выделения n-грамм и выделением ключевых выражений и коллокаций в параллельных текстах.

Выводы к Главе 1

В данной главе мы описали основные процедуры семантической компрессии текстов и один из ее видов - алгоритм тематического моделирования. Нами был описан вариант вероятностного тематического моделирования, реализованный в алгоритме LDA. Был дан обзор процедуры расширения стандартной тематической модели с униграммами, би- и триграммами, выделяемыми с помощью алгоритма RAKE. Был описан принцип действия данного алгоритма и дан краткий обзор иных инструментов поиска и выделения ключевых выражений и коллокаций.

Также было дано определение многоязычной тематической модели и описаны разные виды и способы создания таких моделей на материале параллельных многоязычных корпусов текстов; описан процесс интеграции процедур семантической компрессии в мультимодальные тематические модели. Была описана архитектура тематической модели, реализованная в данной работе и обоснована ее мультимодальность.

Глава 2. Эксперимент по построению мультимодальной тематической модели корпуса параллельных текстов резолюций ООН

В нашем исследовании осуществляется попытка создать мультимодальную тематическую модель специализированного корпуса общественно-политических текстов, а именно, корпуса параллельных текстов резолюций Организации Объединенных Наций (ООН) на английском и русском языках.

Мультимодальность тематической модели обеспечивается тем, что

а) строится тематическая модель для каждого корпуса текстов,

б) в модели учитываются лексические единицы разной структуры – униграммы (леммы), биграммы, триграммы, при этом среди словосочетаний, входящих в тематическую модель, присутствуют как устойчивые сочетания (коллокации), так и ключевые выражения, значимым образом характеризующие содержание текстов корпуса.

2.1. Подготовка лингвистических данных

Для проведения эксперимента был выбран корпус параллельных текстов резолюций ООН [Ziemski et al. 2016], находящийся в свободном доступе на сайте ООН.[20] Данный корпус был собран в рамках инициативы ООН по поддержке языкового разнообразия и развития современных лингвистических технологий. Корпус текстовых данных ООН призван помогать в задачах исследования процедур обработки естественного языка, машинного перевода и других лингвистических задач, отвечающих современным потребностям.

Выбранный для экспериментов корпусной ресурс представляет собой коллекцию текстов резолюций, принятых ООН в 2000 году. Эта коллекция

[20] URL: https://conferences.unite.un.org/UNCORPUS/en/DownloadOverview (дата обращения: 31.05.2022 г.).

была подготовлена специально для проведения исследований по автоматической обработке текстов и машинному переводу. В рамках нашего проекта было принято решение сосредоточиться на той части коллекции, которая содержит параллельные тексты на русском и английском языках. Для каждого из выбранных языков в корпусе содержится 2055 текстов резолюций (или 65 тысяч предложений). Каждому предложению из коллекции текстов на одном языке соответствует переводной эквивалент в корпусе второго языка. Таким образом, обе коллекции представляют собой гармонизированный набор связанных данных на разных языках. Корпус текстов на русском языке насчитывает 2 424 172 словоупотребления, англоязычный корпус – 2 716 043. Для удобства анализа и обработки каждый корпус был разбит на 2055 отдельных текстовых файлов, содержащих целый связный текст одной резолюции включая номер резолюции, знаки препинания и специальные символы.

Изначально оригинальные файлы резолюций содержали также символы табуляций и дополнительные данные, которые было решено сразу исключить из корпуса: это введение к самой резолюции, результаты голосования стран-участниц ООН за конкретную резолюцию, а также длинные предложения, содержащие только перечисления различных стран, связанных с обсуждаемым в резолюции вопросом и не представляющих большой ценности для определения темы документа.

Пример исходных языковых данных (русский язык):

[*РЕЗОЛЮЦИЯ 55/100*

Принята на 81-м пленарном заседании 4 декабря 2000 года регистрируемым голосованием 106 голосами против 1 при 67 воздержавшихся по рекомендации Комитета (A/55/602/Add.2, пункт 94)Проект резолюции, рекомендованный в докладе, на рассмотрение в Комитете внесли: Боливия, Гана, Гондурас, Куба и Сальвадор.; голоса распределились следующим образом:

Голосовали за: Алжир, Ангола, Антигуа и Барбуда, Аргентина, Армения, Афганистан, Багамские Острова, Бангладеш, Барбадос, Беларусь, Белиз, Бенин, Боливия, Ботсвана, Бразилия, Буркина-Фасо, Бурунди, Бутан, Вануату, Венесуэла, Вьетнам, Габон, Гаити, Гайана, Гамбия, Гана, Гватемала, Гвинея, Гондурас,

Гренада, Демократическая Республика Конго, Доминика, Доминиканская Республика, Египет, Замбия, Зимбабве, Индия, Индонезия, Иордания, Иран (Исламская Республика), Йемен, Кабо-Верде, Камбоджа, Камерун, Катар, Китай, Коморские Острова, Конго, Корейская Народно-Демократическая Республика, Кот-д'Ивуар, Куба, Кувейт, Лаосская Народно-Демократическая Республика, Лесото, Ливан, Ливийская Арабская Джамахирия, Маврикий, Мавритания, Мадагаскар, Малави, Мали, Марокко, Мексика, Мозамбик, Мьянма, Намибия, Непал, Нигерия, Никарагуа, Объединенная Республика Танзания, Объединенные Арабские Эмираты, Оман, Пакистан, Панама, Папуа-Новая Гвинея, Парагвай, Перу, Российская Федерация, Руанда, Сальвадор, Сан-Томе и Принсипи, Саудовская Аравия, Свазиленд, Сенегал, Сент-Винсент и Гренадины, Сент-Китс и Невис, Сент-Люсия, Сирийская Арабская Республика, Соломоновы Острова, Судан, Суринам, Сьерра-Леоне, Того, Тринидад и Тобаго, Тунис, Турция, Уганда, Уругвай, Филиппины, Чад, Чили, Шри-Ланка, Эквадор, Эритрея, Эфиопия, Ямайка

Голосовали против: Соединенные Штаты Америки

Воздержались: Австралия, Австрия, Азербайджан, Албания, Андорра, Бахрейн, Бельгия, Болгария, Бруней-Даруссалам, бывшая югославская Республика Македония, Венгрия, Германия, Греция, Грузия, Дания, Джибути, Израиль, Ирландия, Исландия, Испания, Италия, Казахстан, Канада, Кения, Кипр, Колумбия, Коста-Рика, Кыргызстан, Латвия, Литва, Лихтенштейн, Люксембург, Малайзия, Мальдивские Острова, Мальта, Маршалловы Острова, Микронезия (Федеративные Штаты), Монако, Монголия, Науру, Нидерланды, Новая Зеландия, Норвегия, Палау, Польша, Португалия, Республика Корея, Республика Молдова, Румыния, Самоа, Сан-Марино, Сингапур, Словакия, Словения, Соединенное Королевство Великобритании и Северной Ирландии, Таиланд, Узбекистан, Украина, Фиджи, Финляндия, Франция, Хорватия, Чешская Республика, Швеция, Эстония, Южная Африка, Япония

55/100. Уважение права на всеобщую свободу передвижения и чрезвычайная важность воссоединения семей.

Генеральная Ассамблея,

подтверждая, что все права человека и основные свободы универсальны, неделимы, взаимозависимы и взаимосвязаны,

ссылаясь на положения Всеобщей декларации прав человекаРезолюция 217 A (III)., а также на статью 12 Международного пакта о гражданских и политических правахСм. резолюцию 2200 A (XXI), приложение.,

подчеркивая, что, согласно Программе действий Международной конференции по народонаселению и развитиюДоклад Международной конференции по народонаселению и развитию, Каир, 5-13 сентября 1994 года (издание Организации Объединенных Наций, в продаже под № R.95.XIII.18), глава I, резолюция 1, приложение., воссоединение семей зарегистрированных мигрантов является важным фактором международной миграции, а денежные переводы зарегистрированных мигрантов в их страны происхождения зачастую составляют очень важный источник валютных поступлений и способствуют повышению благосостояния оставшихся в странах происхождения родственников,

ссылаясь на свою резолюцию 54/169 от 17 декабря 1999 года,

1. вновь призывает все государства гарантировать общепризнанное право на свободу передвижения всем иностранным гражданам, проживающим на их территории на законных основаниях;

2. подтверждает, что все правительства, в частности правительства принимающих стран, должны признавать чрезвычайную важность воссоединения семей и содействовать включению этого положения в национальные законы в целях обеспечения защиты единства семей зарегистрированных мигрантов;

3. призывает все государства разрешать в соответствии с международно-правовыми документами свободный перевод денежных средств проживающими на их территории иностранными гражданами родственникам в странах происхождения;

4. призывает также все государства воздерживаться от принятия и обеспечить отмену уже существующих законов, призванных служить в качестве принудительной меры, дискриминационной по отношению к отдельным лицам или группам лиц из числа законных мигрантов в силу отрицательного воздействия на воссоединение семей и право направлять денежные переводы родственникам в стране происхождения;

5. постановляет продолжить рассмотрение этого вопроса на своей пятьдесят седьмой сессии по пункту, озаглавленному «Вопросы прав человека».]

Пример исходных языковых данных (английский язык):

[*55/100 55/100. Respect for the right to universal freedom of travel and the vital importance of family reunification The General Assembly, Reaffirming that all human rights and fundamental freedoms are universal, indivisible, interdependent and interrelated, Recalling the provisions of the Universal Declaration of Human Rights,Resolution 217 A (III). as well as article 12 of the International Covenant on Civil and Political Rights,See resolution 2200 A (XXI), annex. Stressing that, as stated in the Programme of Action of the International Conference on Population and Development,Report of the International Conference on Population and Development, Cairo, 5-13 September 1994 (United Nations publication, Sales No. E.95.XIII.18), chap. I, resolution 1, annex. family reunification of documented migrants is an important factor in international migration and that remittances by documented migrants to their countries of origin often constitute a very important source of foreign exchange and are instrumental in improving the well-being of relatives left behind, Recalling its resolution 54/169 of 17 December 1999, 1. Once again calls upon all States to guarantee the universally recognized freedom of travel to all foreign nationals legally residing in their territory; 2. Reaffirms that all Governments, in particular those of receiving countries, must recognize the vital importance of family reunification and promote its incorporation into national legislation in order to ensure protection of the unity of families of documented migrants; 3. Calls upon all*

States to allow, in conformity with international legislation, the free flow of financial remittances by foreign nationals residing in their territory to their relatives in the country of origin; 4. Also calls upon all States to refrain from enacting, and to repeal if it already exists, legislation intended as a coercive measure that discriminates against individuals or groups of legal migrants by adversely affecting family reunification and the right to send financial remittance to relatives in the country of origin; 5. Decides to continue its consideration of this question at its fifty-seventh session under the item entitled "Human rights questions".]

При подготовке корпуса для проведения исследования из всех текстов резолюций были также удалены знаки табуляции и деление на абзацы, однако, ради корректного проведения процедуры морфологического анализа, оставлены знаки пунктуации и числа, слова, записанные в верхнем регистре, термины на иностранных языках и специальные символы.

Пример итоговых языковых данных (русский язык):

[*55/100 55/100. Уважение права на всеобщую свободу передвижения и чрезвычайная важность воссоединения семей. Генеральная Ассамблея, подтверждая, что все права человека и основные свободы универсальны, неделимы, взаимозависимы и взаимосвязаны, ссылаясь на положения Всеобщей декларации прав человека Резолюция 217 A (III)., а также на статью 12 Международного пакта о гражданских и политических правах См. резолюцию 2200 A (XXI), приложение., подчеркивая, что, согласно Программе действий Международной конференции по народонаселению и развитию Доклад Международной конференции по народонаселению и развитию, Каир, 5-13 сентября 1994 года (издание Организации Объединенных Наций, в продаже под № R.95.XIII.18), глава I, резолюция 1, приложение., воссоединение семей зарегистрированных мигрантов является важным фактором международной миграции, а денежные переводы зарегистрированных мигрантов в их страны происхождения зачастую составляют очень важный источник валютных поступлений и способствуют повышению благосостояния оставшихся в странах происхождения родственников, ссылаясь на свою резолюцию 54/169 от 17 декабря 1999 года, 1. вновь призывает все государства гарантировать общепризнанное право на свободу передвижения всем иностранным гражданам, проживающим на их территории на законных основаниях; 2. подтверждает, что все правительства, в частности правительства принимающих стран, должны признавать чрезвычайную важность воссоединения семей и содействовать включению этого положения в национальные законы в целях обеспечения защиты единства семей зарегистрированных мигрантов; 3. призывает все государства разрешать в соответствии с международно-правовыми документами свободный перевод денежных средств проживающими на их территории иностранными*

гражданами родственникам в странах происхождения; 4. призывает также все государства воздерживаться от принятия и обеспечить отмену уже существующих законов, призванных служить в качестве принудительной меры, дискриминационной по отношению к отдельным лицам или группам лиц из числа законных мигрантов в силу отрицательного воздействия на воссоединение семей и право направлять денежные переводы родственникам в стране происхождения; 5. постановляет продолжить рассмотрение этого вопроса на своей пятьдесят седьмой сессии по пункту, озаглавленному «Вопросы прав человека».]

Пример итоговых языковых данных (английский язык):

[55/100 55/100. Respect for the right to universal freedom of travel and the vital importance of family reunification. The General Assembly, Reaffirming that all human rights and fundamental freedoms are universal, indivisible, interdependent and interrelated, Recalling the provisions of the Universal Declaration of Human Rights, Resolution 217 A (III). as well as article 12 of the International Covenant on Civil and Political Rights, See resolution 2200 A (XXI), annex. Stressing that, as stated in the Programme of Action of the International Conference on Population and Development, Report of the International Conference on Population and Development, Cairo, 5-13 September 1994 (United Nations publication, Sales No. E.95.XIII.18), chap. I, resolution 1, annex. family reunification of documented migrants is an important factor in international migration and that remittances by documented migrants to their countries of origin often constitute a very important source of foreign exchange and are instrumental in improving the well-being of relatives left behind, Recalling its resolution 54/169 of 17 December 1999, 1. Once again calls upon all States to guarantee the universally recognized freedom of travel to all foreign nationals legally residing in their territory; 2. Reaffirms that all Governments, in particular those of receiving countries, must recognize the vital importance of family reunification and promote its incorporation into national legislation in order to ensure protection of the unity of families of documented migrants; 3. Calls upon all States to allow, in conformity with international legislation, the free flow of financial remittances by foreign nationals residing in their territory to their relatives in the country of origin; 4. Also calls upon all States to refrain from enacting, and to repeal if it already exists, legislation intended as a coercive measure that discriminates against individuals or groups of legal migrants by adversely affecting family reunification and the right to send financial remittance to relatives in the country of origin; 5. Decides to continue its consideration of this question at its fifty-seventh session under the item entitled "Human rights questions".]

Таким образом, в итоговом корпусе оказались представлены тексты резолюций, принятых ООН в 2000 году, с исключенными из них избыточными языковыми данными. Готовые данные представляют собой не набор уникальных предложений, не зависящих друг от друга, а тематически связные тексты, в этом состоит принципиальное отличие данного

исследования от нашей предыдущей работы, которая выполнялась на материале корпуса разрозненных публикаций ООН [Петрицкая 2021].

Преимуществом рассматриваемого корпуса официальных общественно-политических текстов можно назвать его узкую специализированность и определенную тематическую направленность, что позволяет предугадать возможные термины и темы; более точным анализ таких текстов делает и характерное для публицистического стиля обилие терминов и устойчивых сочетаний, отсутствие метафор и сложных синтаксических оборотов, синонимии и экспрессивности.

2.2. Предобработка корпусных данных

Обработка текстов корпуса проходила в несколько этапов, ее детали различались в зависимости от языка.

На первом шаге для русскоязычного корпуса проводилась частеречная разметка каждого документа морфологическим теггером pymorphy2 [21], что требовало сохранения всех знаков пунктуации и стоп-слов. Данный инструмент относится к морфоанализаторам словарного типа со встроенным предсказателем, в pymorphy2 используется словарь ресурса OpenCorpora[22], отражающий современное состояние словарного состава языка. Выбор данного инструмента из всех доступных (mystem[23], Spacy[24], Stanza[25], UDPipe[26], RNNMorph[27] и т.д.) обуславливается тем, что долгое время pymorphy2 предоставляет стандарт морфологической разметки русскоязычных текстов. Pymorphy2 позволяет приводить слово к его нормальной форме, ставить

[21] URL: https://pymorphy2.readthedocs.io/en/latest/ (дата обращения: 31.05.2022 г.)
[22] URL: http://opencorpora.org/ (дата обращения: 31.05.2022 г.)
[23] URL: https://yandex.ru/dev/mystem/ (дата обращения: 31.05.2022 г.)
[24] URL: https://spacy.io/api/tokenizer (дата обращения: 31.05.2022 г.)
[25] URL: https://stanfordnlp.github.io/stanza/tokenize.html (дата обращения: 31.05.2022 г.)
[26] URL: https://lindat.mff.cuni.cz/services/udpipe/ (дата обращения: 31.05.2022 г.)
[27] URL: https://github.com/IlyaGusev/rnnmorph/blob/master/README.md (дата обращения: 31.05.2022 г.)

слово в заданную форму и получать полную грамматическую характеристику слова. Более подробно функционал pymorphy2 описан в работе [Korobov 2015]. Частеречная разметка помогает определять границы возможных синтаксических групп и, тем самым, избегать формирования n-грамм, внутри которых нарушены грамматические связи.

Пример текста после частеречной разметки:

[| *55/100* | | *55/100.* | | *уважение права* | *на* | | *всеобщую свободу передвижения* | | *и* | | *чрезвычайная важность* | | *воссоединения семей* | *генеральная ассамблея, подтверждая,* | *что* | *все права* | *человека* | | *и* | *основные свободы* | *универсальны, неделимы,* | *взаимозависимы* | | *и* | *взаимосвязаны,* | *ссылаясь* | *на* | *положения* | *всеобщей декларации* | *прав* | *человекарезолюция* | *217* | | *a* | | *(iii).,* | | *a* | *также на статью* | *12* | *международного пакта* | *о гражданских* | *и* | *политических правахсм. резолюцию* | *2200* | | *a* | | *(xxi),* | | *приложение.,* | *подчеркивая, что, согласно программе* | *действий* | *международной конференции* | *по* | *народонаселению* | | *и* | *развитию* | *доклад международной конференции* | *по* | *народонаселению* | | *и* | *развитию, каир,* | *5-13* | *сентября* | | *1994* | *года* | *(издание организации* | *объединенных наций,* | *в продаже под* | *№* | | *r.95.xiii.18),* | *глава* | *i,* | *резолюция* | *1,* | | *приложение.,* | | *воссоединение* | *семей* | *зарегистрированных мигрантов* | | *является* | *важным фактором международной миграции,* | *a* | *денежные переводы зарегистрированных мигрантов* | *в их страны* | | *происхождения* | | *зачастую* | | *составляют* | | *очень* | *важный* | *источник* | *валютных* | *поступлений* | | *и* | | *способствуют* | *повышению благосостояния* | *оставшихся в странах* | *происхождения* | *родственников,* | *ссылаясь* | *на свою резолюцию* | *54/169* | *от* | *17* | *декабря* | | *1999* | *года,* | *1.* | | *вновь* | | *призывает* | *все государства* | | *гарантировать* | *общепризнанное право на свободу* | *передвижения* | *всем иностранным гражданам, проживающим на их территории на законных основаниях;* | *2.* | | *подтверждает,* | | *что* | *все правительства, в частности правительства* | *принимающих стран,* | | *должны* | | *признавать* | *чрезвычайную важность* | | *воссоединения* | *семей* | | *и* | | *содействовать* | *включению этого* | *положения* | *в национальные законы в целях обеспечения* | *защиты* | *единства* | *семей* | *зарегистрированных мигрантов;* | | *3.* | | *призывает* | *все государства* | | *разрешать* | *в* | *соответствии* | *с международно-правовыми документами свободный перевод денежных средств* | *проживающими на их территории иностранными гражданами* | *родственникам в странах происхождения;* | *4.* | | *призывает* | *также все государства* | | *воздерживаться* | *от принятия* | | *и* | | *обеспечить* | *отмену* | *уже* | *существующих законов, призванных* | *служить* | *в качестве принудительной меры, дискриминационной по отношению к отдельным лицам* | *или* | *группам лиц* | *из числа* | *законных мигрантов* | *в силу отрицательного* | *воздействия* | *на* | *воссоединение* | *семей* | | *и* | *право* | *направлять* | *денежные переводы* | *родственникам в стране происхождения;* | *5.* | | *постановляет* | | *продолжить* | *рассмотрение этого вопроса* | *на своей пятьдесят седьмой сессии по пункту, озаглавленному «вопросы прав* | | *человека».* |]

Далее на размеченных данных осуществлялась работа по выделению n-грамм (ключевых выражений с помощью алгоритма RAKE [Rose et al. 2010] и коллокаций с помощью внутреннего алгоритма ngram_range в библиотеке scikit-learn[28]). После выделения n-граммы приводились в нормальную форму с помощью парсера Natasha[29] и встраивались в исходный необработанный корпус. На последнем шаге тексты с выделенными биграммами и триграммами проходили процесс очищения от знаков пунктуации, специальных символов, местоимений и других малозначимых элементов с помощью стоп-словаря. Так как используемый стоп-словарь не включает в себя узкоспециализированную лексику, некоторые слова из корпуса, имеющие низкую значимость для построения тематической модели, приходилось удалять вручную, проводя таким образом дополнительную фильтрацию текстов. Для обоих языков помимо предлогов, местоимений, сокращений, цифр и знаков пунктуации в стоп-словари вошли названия месяцев и дней недели, а также высокочастотные общие глаголы, например, “делать”, “осуществлять”, “быть”, “являться”; также было решено удалить высокочастотную лексику, характерную именно для данного корпуса текстов - это название должностей и слова, относящиеся к формальному аспекту работы организации, например “сессия”, “резолюция”, “генеральный секретарь”, “генеральная ассамблея”. При подготовке к работе с алгоритмом тематического моделирования LDA все слова в документах приводились к начальной форме, то есть, проходили процесс лемматизации, при этом они приводились к нижнему регистру, все лишние пробелы и отступы удалялись. Так как работа осуществлялась с корпусом текстов по тематике международных отношений, содержащим слова на латинице и специальные символы, эти элементы также были удалены. В результате был получен

[28] URL:https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html (дата обращения: 31.05.2022 г.)
[29] URL: https://github.com/natasha/natasha (дата обращения: 31.05.2022 г.)

сбалансированный корпус текстов, не содержащий лишних символов и знаков, которые могли бы затруднить построение тематической модели.

Пример подготовленного текста на русском языке:

[*уважение право свобода передвижение _чрезвычайный_важность_ _воссоединение_семья_ право человек _основный_свобода_ неделимый взаимозависимый взаимосвязанный положение _всеобщий_декларация_ право человек _международный_пакт_ гражданский _политический_правах_ _международный_конференция_ народонаселение развитие _доклад_международный_конференция_ народонаселение развитие каир _издание_организация_ продажа _воссоединение_семья_ _зарегистрировать_мигрант_ международный миграция _денежный_перевод_ _зарегистрировать_мигрант_ страна происхождение зачастую источник валютный _повышение_благосостояние_ остаться _страна_происхождение_ родственник гарантировать _общепризнанный_право_ свобода передвижение _иностранный_гражданин_ проживать территория _законный_основание_ правительство правительство _принимать_страна_ _чрезвычайный_важность_ _воссоединение_семья_ содействовать включение положение _национальный_закон_ _цель_обеспечение_ защита единство семья _зарегистрировать_мигрант_ разрешать международно-правовой свободный перевод денежный проживать _территория_иностранный_гражданин_ родственник _страна_происхождение_ _существовать_закон_ призвать служить _принудительный_мера_ дискриминационный _отдельный_лицо_ _группа_лицо_ _законный_мигрант_ _сила_отрицательный_ _воссоединение_семья_ право направлять _денежный_перевод_ родственник _страна_происхождение_ _вопрос_право_человек_*]

Для англоязычного корпуса был применен иной вариант предобработки текстов. В отличие от русского корпуса, для которого необходимо проводить отдельную процедуру морфологической аннотации, предваряющую процесс выделения ключевых выражений, английский корпус претерпевает морфологическую разметку в рамках самой процедуры выделения ключевых выражений с помощью алгоритма RAKE. В том варианте, который реализован для английского языка, алгоритм RAKE выделяет необходимые выражения в необработанном тексте, и поэтому на вход ему подавался корпус документов с неразмеченными данными. После выделения ключевые выражения записывались в отдельный файл для каждого документа. На следующем шаге выделенные конструкции

проходили лемматизацию – процедуру приведения к нормальной форме – и вставлялись в исходный текст с пометой '_', позволяющей учитывать словосочетание как единый элемент. Далее тексты проходили необходимую предобработку для обучения алгоритма тематического моделирования: приводились к нижнему регистру, с помощью стоп-словаря для английского языка nltk.stopwords в составе библиотеки NLTK[30] [Bird et al. 2009] удалялись предлоги, артикли, наречия и другие стоп-слова. Слова, характерные для данного корпуса и не несущие семантической нагрузки, удалялись с помощью специально созданного фильтра вручную, подобно тому, что применялся для русского языка. Все оставшиеся слова и конструкции лемматизировались с помощью встроенного в пакет библиотеки NLTK лемматизатора WordNetLemmatizer[31].

Пример подготовленного текста на английском языке:

[*_universal_freedom_ travel _vital_importance_ _family_reunification_ _human_right_ _fundamental_freedom_ universal indivisible interdependent interrelate provision _universal_declaration_ _human_right_ _international_covenant_ civil _political_right_ _international_conference_ population _international_conference_ population cairo _united_nation_publication_ _family_reunification_ _documented_migrant_ _important_factor_ _international_migration_ remittance _documented_migrant_ _origin_often_constitute_ _important_source_ _foreign_exchange_ instrumental well-being _relative_left_behind_ _call_upon_ guarantee _universally_recognized_freedom_ travel foreign legally reside territory _receiving_country_ _must_recognize_ _vital_importance_ _family_reunification_ incorporation _national_legislation_ _ensure_protection_ unity family _documented_migrant_ _call_upon_ conformity inter_national_legislation_ _free_flow_ _financial_remittance_ _foreign_national_residing_ territory relative origin _call_upon_ refrain repeal _already_exists_ legislation _coercive_measure_ discriminates _legal_migrant_ adversely _family_reunification_ send _financial_remittance_ relative origin _human_right_questions_*]

После обработки тексты подавались на вход алгоритмам RAKE и LDA для построения мультимодальной тематической модели.

[30] URL: https://www.nltk.org/ (дата обращения: 31.05.2022 г.)
[31] URL: https://www.nltk.org/_modules/nltk/stem/wordnet.html (дата обращения: 31.05.2022 г.)

2.3. Выделение ключевых выражений в корпусе текстов

Для выделения ключевых выражений в данной работе был применен алгоритм **RAKE (Rapid Automatic Keyword Extractor)**, который определяет ключевые слова как n-граммы, ранжированные по определенным признакам, характеризующиеся с точки зрения лексико-грамматических шаблонов, частоты употребления в корпусе, совместной встречаемости слов и словосочетаний в тексте, а также ряда других количественных характеристик текстов.

В данной работе использовалась реализация алгоритма RAKE для языка программирования Python в рамках инструментария NLTK[32]. Для русского языка на вход подавался текст с частеречной разметкой, на выходе алгоритм выдавал ранжированный список найденных выражений (униграмм, биграмм, триграмм и т.д.). В процессе работы алгоритм RAKE не учитывал знаки пунктуации и специальные символы в качестве кандидатов или элементов ключевых выражений, поэтому не требовал настройки дополнительных параметров для фильтрации результатов; алгоритм также ориентировался на границы словосочетаний, выделенных на предыдущем шаге. Количество выделенных фраз не ограничивалось пользовательскими установками, поэтому для каждого документа выдача различалась по объему, при этом даже в самом маленьком тексте было выделено достаточно ключевых выражений. Сами тексты резолюций различаются по размеру, поэтому и количество выделенных в них вражений также отличается. Количество элементов в выделенных выражениях ограничивалось шестью (дефолтное значение параметра в RAKE). Результаты работы RAKE записывались в отдельный файл для каждого текста.

[32] URL: https://www.nltk.org/ (дата обращения: 31.05.2022 г.)

Пример выделенных ключевых выражений для русскоязычного текста с морфологической разметкой приведен в таблице 1.

Таблица 1. Ключевые выражения, выделенные с помощью RAKE на морфологически размеченном корпусе (русский язык)

Коллокация	**Вес**
международно-правовыми документами свободный перевод денежных средств	*36.0*
важным фактором международной миграции	*15.0*
денежные переводы зарегистрированных мигрантов	*11.166666666666666*
доклад международной конференции	*8.5*
территории иностранными гражданами	*8.0*
вопросы прав человека	*7.0*
денежные переводы	*6.0*
зарегистрированных мигрантов	*5.166666666666666*
законных мигрантов	*4.5*
законных основаниях	*4.0*
политических правах	*4.0*
основные свободы	*4.0*
повышению благосостояния	*4.0*
иностранным гражданам	*4.0*
принимающих стран	*4.0*

национальные законы	*4.0*
группам лиц	*4.0*
стране происхождения	*3.5*
воссоединение семей	*3.1*
человека	*2.0*
семей	*1.6*
право	*1.5*

В таблице 1 сразу выделяются первые два результата – они имеют наибольший вес, так как содержат в себе наибольшее количество словоформ. Несмотря на то, что эти сочетания являются грамматически правильнооформленными и семантически связанными, они не включаются в корпус при подготовке текстов к обработке алгоритмом тематического моделирования, так как данные выражения являются узклспециализированными и при этом избыточными. Таким образом, на этапе обработки результатов отсекаются высокочастотные выражения вверху ранжированного списка и низкочастотные униграммы внизу списка. К анализу привлекаются только биграммы и триграммы, размечаемые в документах корпуса как ключевые выражения. В целом, работа алгоритма RAKE показала хорошие результаты в отношении количества и качества выделенных выражений.

Для сравнения результатов работы RAKE с разными типами данных при выделении ключевых выражений в корпусе текстов на английском языке не проводилась процедура морфологической разметки. На вход алгоритму подавался необработанный текст с сохранением стоп-слов, знаков препинания и табуляции. При работе с текстом алгоритм RAKE сам очищал данные от малозначимых слов и символов с помощью встроенного пакета

nltk.stopwords для английского языка. Было установлено ограничение по размеру выделяемых коллокаций – от двух до трех словоупотреблений. После этого алгоритм RAKE начинал работу с документами.

Пример выделенных ключевых выражений для текста на английском языке без морфологической разметки приведен в таблице 2.

Таблица 2. Коллокации, выделенные с помощью RAKE на неразмеченном корпусе текстов (английский язык)

Коллокация	***Вес***
united nations publication	*9.0*
relatives left behind	*9.0*
universally recognized freedom	*9.0*
origin often constitute	*9.0*
foreign nationals residing	*8.5*
universal freedom	*4.5*
foreign exchange	*4.5*
universal declaration	*4.0*
receiving countries	*4.0*
political rights	*4.0*
national legislation	*4.0*
international migration	*4.0*

legal migrants	*4.0*
human rights	*4.0*
fundamental freedoms	*4.0*
family reunification	*4.0*
documented migrants	*4.0*
coercive measure	*4.0*
financial remittance	*4.0*

Сравнивая полученные ключевые выражения, можно заключить, что RAKE одинаково успешно справляется с выделением устойчивых выражений как в размеченных, так и в необработанных текстах. Веса выражений одной структуры (униграммы, биграммы, триграммы) принимают значения из одного и того же диапазона для рассматриваемых языков, но сами сочетания незначительно различаются. Таким образом, для идентичных текстов на русском и английском языках мы получили разные наборы выделенных алгоритмом RAKE ключевых выражений.

Однако с разными языками алгоритм работает по-разному. Так, тексты на русском языке без разметки обрабатываются значительно хуже. Это объясняется тем, что RAKE изначально был ориентирован на работу с английским языком ср. примеры коллокаций из неразмеченного английского и русского текстов в таблице 3. Как видно, в русских текстах без предварительной обработки выделяется очень мало коллокаций, не имеющих смысловой цельнооформленности и плохо отражающих тематику документа.

Таблица 3. Коллокации, выделенные с помощью RAKE на неразмеченном корпусе текстов (русский и английский языки)

Русский	**Английский**
что все правительства	*united nations publication*
резолюцию 2200 a	*origin often constitute*
взаимозависимы и взаимосвязаны	*foreign nationals residing*
резолюция 1	*foreign exchange*
100 55	*vital importance*
	seventh session
	receiving countries
	national legislation
	legal migrants
	international migration

2.4. Лемматизация текстов и интеграция ключевых выражений в корпус

На этом этапе выбранные n-граммы проходили процесс приведения к нормальной форме и интегрировались в исходный необработанный текст. Для лемматизации словосочетаний на русском языке использовался встроенный лемматизатор библиотеки Natasha[33]. Этот лемматизатор был опробован нами ранее как альтернатива лемматизатору Mystem[34] [Сегалович, Титов 2005] . Хотя последний и правильно приводит все n-граммы к нормальной форме, работа анализатора занимает большое количество времени, что не всегда может быть удобно. Лемматизатор Natasha, в свою

[33] URL: https://pypi.org/project/natasha/ (дата обращения: 31.05.2022 г.)

[34] URL: https://yandex.ru/dev/mystem/ (дата обращения: 31.05.2022 г.)

очередь, работает намного быстрее и так же хорошо справляется с поставленной задачей.

В процессе лемматизации алгоритм помечал обработанные коллокации знаком «_» на месте пробелов для того, чтобы эти коллокации были хорошо видны в тексте и считались как один токен. После приведения к нормальной форме и добавления метки коллокации вставлялись в исходный текст.

Для лемматизации слов на английском языке существует большое количество инструментов, однако не все они просты и понятны в использовании, более того, не все могут лемматизировать целые выражения без применения дополнительных библиотек или частеречной разметки. В данной работе для приведения к нормальной форме коллокаций на английском языке использовался представленный в пакете NLTK лемматизатор WordNetLemmatizer.

Данный лемматизатор в своей первоначальной реализации хорошо работает с униграммами, то есть с отдельными словами, но не может лемматизировать более сложные выражения. Для того чтобы лемматизатор смог обработать целое выражение, каждому слову вначале приписывался морфологический тег. Морфологическая разметка производилась при помощи метода nltk.pos_tag, доступного в самом пакете NLTK[35]. Далее размеченные коллокации подавались на вход алгоритму WordNetLemmatizer, который с опорой на лексическую базу данных WordNet для английского языка[36] приводил каждое слово в его нормальную форму в зависимости от назначенного ему морфологического тега. На выходе алгоритма мы получили полностью лемматизированные выражения, помеченные знаком «_» на месте пробела.

Так как для ключевых выражений, выделенных алгоритмом RAKE, назначалось ограничение по структуре n-грамм $2 \leq n \leq 3$, размечались и использовались только биграммы и триграммы. После лемматизации

[35] URL: https://www.nltk.org/book/ch05.html (дата обращения: 31.05.2022 г.)
[36] URL: https://wordnet.princeton.edu/ (дата обращения: 31.05.2022 г.)

выражения вставлялись в исходные необработанные тексты. Помимо этого, при выделении коллокаций в документах из текстов удалялись все знаки препинания, а сам текст приводился к нижнему регистру.

Пример выражений RAKE, интегрированных в русский корпус:

[*55100 55100 уважение права на всеобщую свободу передвижения и _чрезвычайный_важность_ _воссоединение_семья_ _генеральный_ассамблея_ подтверждая что все права человека и _основный_свобода_ универсальны неделимы взаимозависимы и взаимосвязаны ссылаясь на положения _всеобщий_декларация_ прав человека резолюция 217 a iii a также на статью 12 _международный_пакт_ о гражданских и _политический_право_ резолюцию 2200 a xxi приложение подчеркивая что согласно программе действий _международный_конференция_ по народонаселению и развитию _доклад_международный_конференция_ по народонаселению и развитию каир 5-13 сентября 1994 года _издание_организация_ _объединить_нация_ в продаже под r95xiii18 глава i резолюция 1 приложение _воссоединение_семья_ _зарегистрировать_мигрант_ является важным фактором международной миграции а _денежный_перевод_ _зарегистрировать_мигрант_ в их страны происхождения зачастую составляют очень важный источник валютных поступлений и способствуют _повышение_благосостояние_ оставшихся в _страна_происхождение_ родственников ссылаясь на свою резолюцию 54169 от 17 декабря 1999 года 1 вновь призывает все государства гарантировать _общепризнанный_право_ на свободу передвижения всем _иностранный_гражданин_ проживающим на их территории на _законный_основание_ 2 подтверждает что все правительства в частности правительства _принимать_страна_ должны признавать _чрезвычайный_важность_ _воссоединение_семья_ и содействовать включению этого положения в _национальный_закон_ в _цель_обеспечение_ защиты единства семей _зарегистрировать_мигрант_ 3 призывает все государства разрешать в соответствии с международно-правовыми документами свободный перевод денежных средств проживающими на их _территория_иностранный_гражданин_ родственникам в _страна_происхождение_ 4 призывает также все государства воздерживаться от принятия и обеспечить отмену уже _существовать_закон_ призванных служить в качестве _принудительный_мера_ дискриминационной по отношению к _отдельный_лицо_ или _группа_лицо_ из числа _законный_мигрант_ в _сила_отрицательный_ воздействия на _воссоединение_семья_ и право направлять _денежный_перевод_ родственникам в _страна_происхождение_ 5 постановляет продолжить рассмотрение этого вопроса на своей _пятьдесят_седьмой_сессия_ по пункту озаглавленному _вопрос_право_человек_]*

Пример выражений RAKE, интегрированных в английский корпус:

[*55/100 55/100 respect for the right to _universal_freedom_ of travel and the _vital_importance_ of _family_reunification_ the _general_assembly_ reaffirming that all*

_human_right_ and _fundamental_freedom_ are universal indivisible interdependent and interrelated recalling the provisions of the _universal_declaration_ of _human_right_ resolution 217 a iii as well as article 12 of the _international_covenant_ on civil and _political_right_ _see_resolution_ 2200 a xxi annex stressing that as stated in the programme of action of the _international_conference_ on population and development report of the _international_conference_ on population and development cairo 5-13 september 1994 _united_nation_publication_ sales no e.95.xiii.18 chap i resolution 1 annex _family_reunification_ of _documented_migrant_ is an _important_factor_ in _international_migration_ and that remittances by _documented_migrant_ to their countries of _origin_often_constitute_ a very _important_source_ of _foreign_exchange_ and are instrumental in improving the well-being of _relative_left_behind_ recalling its resolution 54/169 of 17 december 1999 1. once again _call_upon_ all states to guarantee the _universally_recognized_freedom_ of travel to all foreign nationals legally residing in their territory 2. reaffirms that all governments in particular those of _receiving_country_ _must_recognize_ the _vital_importance_ of _family_reunification_ and promote its incorporation into _national_legislation_ in order to _ensure_protection_ of the unity of families of _documented_migrant_ 3 _call_upon_ all states to allow in conformity with inter_national_legislation_ the _free_flow_ of _financial_remittance_ by _foreign_national_residing_ in their territory to their relatives in the country of origin 4. also _call_upon_ all states to refrain from enacting and to repeal if it _already_exists_ legislation intended as a _coercive_measure_ that discriminates against individuals or groups of _legal_migrant_ by adversely affecting _family_reunification_ and the right to send _financial_remittance_ to relatives in the country of origin 5. decides to continue its consideration of this question at its fifty-_seventh_session_ under the item entitled _human_right_questions_]

2.5. Обработка текстов с выделенными ключевыми выражениями

Последний шаг работы перед построением тематической модели – обработка корпусов текстов с выделенными в них ключевыми выражениями. На этом этапе все тексты приводились в нормальную форму с применением выше описанных алгоритмов лемматизации. Затем лемматизированные тексты очищались от оставшихся знаков препинания, отступов и пробелов, всех союзов, предлогов, местоимений, частиц, цифр, артиклей и специальных символов. Из русского корпуса удалялись все слова, содержащие символы латинского алфавита, из английского – все нелатинские символы. В обоих корпусах присутствовали частотные, но не значимые для будущей модели слова, например, названия месяцев и дней недели, которые можно часто

встретить в официальных документах или отчетах. Такие слова собирались в отдельные стоп-словари для каждого языка и также удалялись из текстов.

По окончании нормализации все обработанные тексты записывались в отдельные файлы, с которыми затем работал алгоритм тематического моделирования LDA, представленный в программной библиотеке для машинного обучения scikit-learn[37].

Пример готового текста на русском языке с ключевыми выражениями RAKE:

[*уважение право всеобщий свобода передвижение _чрезвычайный_важность_ _воссоединение_семья_ _генеральный_ассамблея_ подтверждать право человек _основный_свобода_ универсальный неделимый взаимозависимый взаимосвязанный ссылаться положение _всеобщий_декларация_ право человек _международный_пакт_ гражданский _политический_право_ приложение подчёркивать программа действие _международный_конференция_ народонаселение развитие _доклад_международный_конференция_ народонаселение развитие каир _издание_организация_ _объединить_нация_ продажа приложение _воссоединение_семья_ _зарегистрировать_мигрант_ важный фактор международный миграция _денежный_перевод_ _зарегистрировать_мигрант_ страна происхождение зачастую составлять важный источник валютный поступление способствовать _повышение_благосостояние_ остаться _страна_происхождение_ родственник ссылаться вновь призывать государство гарантировать _общепризнанный_право_ свобода передвижение _иностранный_гражданин_ проживать территория _законный_основание_ подтверждать правительство частность правительство _принимать_страна_ признавать _чрезвычайный_важность_ _воссоединение_семья_ содействовать включение положение _национальный_закон_ _цель_обеспечение_ защита единство семья _зарегистрировать_мигрант_ призывать государство разрешать соответствие международно-правовой документ свободный перевод денежный средство проживать _территория_иностранный_гражданин_ родственник _страна_происхождение_ призывать весь государство воздерживаться принятие обеспечить отмена _существовать_закон_ призвать служить качество _принудительный_мера_ дискриминационный отношение _отдельный_лицо_ _группа_лицо_ _законный_мигрант_ _сила_отрицательный_ воздействие _воссоединение_семья_ право направлять _денежный_перевод_ родственник _страна_происхождение_ постановлять продолжить рассмотрение вопрос _пятьдесят_седьмой_сессия_ озаглавить _вопрос_право_человек_]*

[37] URL: https://scikit-learn.org/stable/ (дата обращения: 31.05.2022 г.)

Пример готового текста на английском языке с ключевыми выражениями RAKE:

[*respect _universal_freedom_ travel _vital_importance_ _family_reunification_ _general_assembly_ _human_right_ _fundamental_freedom_ universal indivisible interdependent interrelate recall provision _universal_declaration_ _human_right_ article _international_covenant_ civil _political_right_ _see_resolution_ annex stress programme _international_conference_ population _international_conference_ population cairo _united_nation_publication_ sale chap annex _family_reunification_ _documented_migrant_ _important_factor_ _international_migration_ remittance _documented_migrant_ _origin_often_constitute_ _important_source_ _foreign_exchange_ instrumental improve well-being _relative_left_behind_ recall _call_upon_ guarantee _universally_recognized_freedom_ travel foreign legally reside territory reaffirms particular _receiving_country_ _must_recognize_ _vital_importance_ _family_reunification_ promote incorporation _national_legislation_ _ensure_protection_ unity family _documented_migrant_ _call_upon_ conformity inter_national_legislation_ _free_flow_ _financial_remittance_ _foreign_national_residing_ territory relative origin _call_upon_ refrain enact repeal _already_exists_ legislation _coercive_measure_ discriminates individual _legal_migrant_ adversely affect _family_reunification_ send _financial_remittance_ relative origin decides fifty-_seventh_session_ entitle _human_right_question_]*

После данного этапа обработки тексты заметно уменьшаются в объеме и содержат только значимые для построения модели слова в их начальной форме, см. таблицу 4.

Таблица 4. Объемы корпусов до и после предобработки

Корпус	**Объем до предобработки (слов)**	**Объем после предобработки (слов)**
Русский	2 424 172	735 734
Английский	2 716 043	682 366

2.6. Построение n-граммных тематических моделей корпуса на основе многоязычного алгоритма LDA

Очищенные и лемматизированные тексты подавались на вход алгоритму LDA, реализованному в библиотеке scikit-learn[38], для построения тематических моделей двух коллекций.

Модель LDA обучается на корпусе с учетом предварительно назначенных пользовательских параметров: количество выделяемых тем, количество слов в каждой теме, дополнительный стоп-словарь, количество итераций процедуры, и т.д. Для данной работы было решено выделить двадцать самых главных тем и выбрать по десять слов для каждой темы. Такое число тем и слов было выбрано в результате подбора оптимального значения меры U_mass[39]. Итерация проводилась пятьдесят раз для каждой коллекции. Отдельная тематическая модель строилась для каждого корпуса.

Алгоритм LDA работает сразу со всеми текстами корпуса: все тексты считываются и фильтруются по стоп-словарю, прошедшие фильтрацию слова попадают в список токенов. Из этого списка создается словарь, в котором каждому слову сопоставляется его уникальный номер. N-граммы из двух и трех слов, размеченные в обрабатываемых нами текстах специальным символом «_», считались алгоритмом как один токен и заносились в словарь как единый элемент. Далее из словаря строится модель корпуса, в которой реализуется идея «мешка слов». В модели «мешка слов» представлены пары «номер токена» — «его частотность», далее алгоритм переводит тексты в цифровой формат и находит характерные для текстов встречающиеся слова и выражения, относит их к темам и выводит результат. Для облегчения интерпретации результатов на вывод подается информация в следующем виде: сначала регистрируется номер темы (темы при этом упорядочиваются

[38] URL:https://scikit-learn.org/stable/modules/generated/sklearn.discriminant_analysis.LinearDiscriminantAnalysis.html (дата обращения: 31.05.2022 г.)
[39] URL: https://aksw.org/Projects/Palmetto.html (дата обращения: 31.05.2022 г.)

случайным образом), затем само слово или выражение и его вес в теме (его значимость, мера тяготения к теме); компоненты темы расположены в порядке убывания весов – таким образом отражается содержание той или иной темы.

Пример темы (слово/выражение - вес):

Topic 0:[('наркотик', 658.3639348436441), ('борьба', 619.7717619840672), ('преступность', 599.99264639793), ('помощь', 514.2203038542194), ('_уголовный_правосудие_', 379.9618936742859), ('укрепление', 315.5539662692955), ('_предупреждение_преступность_', 296.3438362991204), ('_социальный_совет_', 286.1323952433334), ('экономический', 264.2224352809701), ('протокол', 259.82498104624705)]

Напомним, что в обрабатываемом корпусе уже произведена сборка ключевых выражений на основе алгоритма RAKE, что обеспечивает построение n-граммной тематической модели, в темах которой будут как отдельные леммы, так и значимые для содержания текстов выражения (биграммы и триграммы). Однако модель LDA позволяет подключить опцию независимого выделения n-грамм (ngram_range)[40] из библиотеки scikit-learn, которая в нашем случае дополняет алгоритм RAKE. Такая модель, помимо учета словосочетаний, ранее выделенных с помощью RAKE, способна сама находить устойчивые n-граммы (коллокации) в текстовых данных. Эта процедура реализуется на основе векторного представления корпуса в модуле CountVectorizer с помощью параметра ngram_range = (x, y), в котором можно указать желаемый размер n-грамм. Модель с такими дополнительными параметрами оказалась тяжелее и медленнее обычной реализации LDA, поэтому в силу экономии времени и текстового объема было решено опробовать данный алгоритм на меньшем количестве тем. Всего было решено выводить по десять тем, содержащих десять слов.

В ходе построения тематической модели были опробованы следующие варианты реализации n-граммной модели LDA.

[40] URL:https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html (дата обращения: 31.05.2022 г.)

LDA с биграммами и триграммами: данный алгоритм при построении тематической модели агрегирует n-граммы из уже имеющихся слов и выражений обработанного корпуса. Так, на выходе работы алгоритма можно получить целые фразы, состоящие из отдельного слова из коллекции и ключевого выражения RAKE, которые могут сочетаться вместе и формировать законченное смысловое единство, примеры таких выражений представлены в таблице 5.

Таблица 5. Биграммы и триграммы, выделенные в ходе обучения модели LDA (примеры потенциальных переводных эквивалентов)

Русский	**Английский**
*гонка вооружение _космический_пространство_*	*_arm_race_ _outer_space_*
*насилие _отношение_женщина_*	*violence woman*
*_предоставление_независимость_ _колониальный_страна_*	*grant independence _colonial_country_*

Такие выражения не были выделены как самостоятельные единицы на этапе работы алгоритма RAKE из-за наличия в них предлогов или высокой сложности фразы. Однако алгоритм LDA отлично справляется с восстановлением смысловой и грамматической связности n-грамм в полностью лемматизированном корпусе, основываясь на совместной встречаемости слов. Также данная функция позволяет выделять в отдельные группы ряды синонимов и перечислений, часто встречающихся рядом друг с другом в текстах резолюций, что позволяет представить тему коллекции в более близком к оригиналу варианте, см. таблицу 6.

Таблица 6. Биграммы и триграммы, выделенные в ходе обучения модели LDA (примеры потенциальных переводных эквивалентов)

Русский	**Английский**

расизм _расовый_дискриминация_ ксенофобия	*racism _racial_discrimination_ xenophobia*
религия убеждение	*religion belief*
пытка жестокий бесчеловечный	*torture cruel inhuman*
женщина девочка	*woman girl*

Ожидаемо, в темах для английского и русского корпусов текстов зачастую присутствуют одинаковые выражения и переводные эквиваленты, найденные как на этапе предобработки текстов, так и самим алгоритмом LDA. Однако у такого подхода оказались и недостатки, например, генерация избыточных фраз, которые могут наслаиваться одна на другую или же по цепочке повторять друг друга, примеры см. в таблице 7.

Таблица 7. Биграммы и триграммы, выделенные в ходе обучения модели LDA (примеры потенциальных переводных эквивалентов)

Русский	**Английский**
запас далеко	*highly migratory*
*далеко _мигрировать_рыба_*	*migratory _fish_stock_*
*запас далеко _мигрировать_рыба_*	*highly migratory _fish_stock_*

Результаты тематического моделирования с выделением биграмм и триграмм для русского и английского корпуса представлены в таблице 8.

Таблица 8. Темы из модели LDA с биграммами и триграммами (параллельные темы)

1 ru	*_стрелковый_оружие_ _легкий_вооружение_, торговля _стрелковый_оружие_, торговля _стрелковый_оружие_ _легкий_вооружение_, незаконный торговля, проблема _незаконный_торговля_, _вопрос_мир_ разоружение, лёгкий вооружение, незаконный торговля _стрелковый_оружие_, _уменьшение_опасность_ бедствие, решение проблема*
en	*_small_arm_ _light_weapon_, peace disarmament, eradicate eliminate, _united_nation_programme_, _combat_trafficking_, culture peace, _illicit_trade_ _small_arm_, _illicit_trade_ _small_arm_ _light_weapon_, peace disarmament, _related_report_ _advisory_committee_, _international_peace_*
2 ru	*договор нераспространение, нераспространение _ядерный_оружие_, угроза применение _ядерный_оружие_, договор нераспространение _ядерный_оружие_, участник договор нераспространение, сумма размер, _ядерный_оружие_ договор, _создание_зона_ свободный _ядерный_оружие_, договор _заключительный_документ_, укрепление доверие*
en	*proliferation _nuclear_weapon_, nuclear _weapon_state_, _review_conference_ proliferation, _review_conference_ proliferation _nuclear_weapon_, _nuclear_weapon_ _final_document_, proliferation _nuclear_weapon_ _final_document_, atomic energy, _extension_conference_ proliferation, _extension_conference_ proliferation _nuclear_weapon_, nuclear _free_zone_*
3 ru	*_третий_конференция_ _морской_право_, _региональный_рыбохозяйственный_организация_ договорённость, _различный_вид_ морепользование , разрешать суд плавать , _развиваться_государство_ _иметь_выход_ море , запас далеко, далеко _мигрировать_рыба_, запас далеко _мигрировать_рыба_, незаконный несообщаемый, людской ресурс*
en	*food _agriculture_organization, _regional_fishery_management_ arrangement, ocean law, _ocean_affair_ law, _regional_fishery_management_ arrangement, highly migratory, migratory _fish_stock_, highly migratory _fish_stock_, pollution ship, landlocked _transit_developing_country_*

4 ru	*_область_предупреждение_ преступность, преступность _уголовный_правосудие_, _область_предупреждение_ преступность _уголовный_правосудие_, наркотик преступность, _предупреждение_преступность_ _уголовный_правосудие_, договор нераспространение, нераспространение _ядерный_оружие_, договор нераспространение _ядерный_оружие_, _специализированный_учреждение_ _организация_система_, участник договор*
en	*drug crime, narcotic board, _crime_prevention_ _criminal_justice_, _united_nation_office_ drug crime, _crime_prevention_ _criminal_justice_programme_, _united_nation_convention_ _transnational_organized_crime_, economic _social_council_, _narcotic_drug_ _psychotropic_substance_, counter _world_drug_problem_, _united_nation_convention_ corruption*
5 ru	*право человек, возвращенец _переместить_лицо_, отправление правосудие, _область_право_ человек, зона конфликт, беженец возвращенец, алмаз зона конфликт, алмаз зона, беженец возвращенец _переместить_лицо_, _переместить_лицо_ африка*
en	*refugee returnees, _displaced_person_, _high_commissioner_ refugee, _human_right_ _fundamental_freedom_, protection _human_right_, child _armed_conflict_, universally _human_right_, _person_belonging_, _displaced_person_ africa, _human_right_ migrant*
6 ru	*право человек _основный_свобода_, _область_право_ человек, человек _основный_свобода_, торговля человек, насилие _отношение_женщина_, женщина девочка, торговля женщина, ликвидация дискриминация, женщина ребёнок, положение женщина*
en	*_human_right_ _fundamental_freedom_, protection _human_right_, traffic woman, discrimination woman, elimination discrimination woman, violence woman, woman girl, _child_labour_, empowerment woman, _human_right_*
7 ru	*гонка вооружение _космический_пространство, _космический_пространство_, _использование_космический_пространство_ _мирный_цель_, _вопрос_образование_ наука, наука культура, _вопрос_образование_ наука культура, научно _технический_подкомитет_, исследование _использование_космический_пространство_, предотвратить гонка, исследование _использование_космический_пространство_ _мирный_цель_*

en	*_arm_race_ _outer_space_, prevention _arm_race_, prevention _arm_race_ _outer_space_* *peace disarmament, _peaceful_us_ _outer_space_ , scientific _technical_subcommittee_ , science technology, amount dollar, confidence _building_measure_, prevent _arm_race_ _outer_space_*
8 ru	*право человек, _область_право_ человек, расизм _расовый_дискриминация_, _расовый_дискриминация_ ксенофобия, ксенофобия нетерпимость, _расовый_дискриминация_ ксенофобия нетерпимость, расизм _расовый_дискриминация_ ксенофобия, насилие _отношение_женщина_, человек _основный_свобода_, борьба расизм*
en	*_racial_discrimination_ xenophobia, discrimination woman, elimination discrimination woman, xenophobia _related_intolerance_, _racial_discrimination_ xenophobia _related_intolerance_, racism _racial_discrimination_ xenophobia, protection _human_right_, _human_right_ _fundamental_freedom_, violence woman, protection _human_right_*
9 ru	*босния герцеговина, разоружение развитие, право человек, _бедный_страна_ _крупный_задолженность_, служба _внутренний_надзор_, _латинский_америка_ _карибский_бассейн_, борьба терроризм, _третий_государство_ пострадать, _облегчение_бремя_ задолженность, помощь _третий_государство_*
en	*bosnia herzegovina, _heavily_indebted_ poor, _federal_republic_ Yugoslavia, _latin_america_ caribbean, economy transition, _developing_country_ economy transition, eradication poverty, timor leste, economic _social_council_, law commit territory*

10 ru	*прекращение насилие, жестокий бесчеловечный, обращение наказание, пытка жестокий, пытка жестокий бесчеловечный, правительство территория, бесчеловечный _унижать_достоинство_вид_, бесчеловечный _унижать_достоинство_вид_ обращение, _унижать_достоинство_вид_ обращение, жестокий бесчеловечный _унижать_достоинство_вид_*
en	*violence child, torture cruel inhuman, religion belief, inhuman _degrading_treatment_ punishment, cruel inhuman _degrading_treatment_, inhuman _degrading_treatment_, torture cruel inhuman, torture cruel, _degrading_treatment_ punishment, inhuman _degrading_treatment_ punishment*

LDA с биграммами: в такой реализации на этапе построения тематической модели алгоритмом выделялись только сочетания из двух слов. Такие сочетания могли состоять из двух отдельных слов коллекции, например, см. таблицу 9.

Таблица 9. Биграммы, выделенные в ходе обучения модели LDA (примеры потенциальных переводных эквивалентов)

Русский	**Английский**
босния герцеговина	*bosnia herzegovina*
ликвидация нищета	*eradication poverty*
положение женщина	*violence woman*
вич спид	*hiv aids*

Такие выражения являют собой полноценные смысловые единства и могли бы быть выделены алгоритмом RAKE, но по какой-то причине остались им незамеченными.

Также в биграммы попали сочетания из зарегистрированного ключевого выражения RAKE и дополняющего его слова из корпуса, см. таблицу 10.

Таблица 10. Биграммы, выделенные в ходе обучения модели LDA (примеры потенциальных переводных эквивалентов)

Русский	**Английский**
*человек _основный_свобода_*	*_human_right_ defender*
*торговля _стрелковый_оружие_*	*administrative _budgetary_question_*
*внутренне _переместить_лицо_*	*league _arab_state_*
*экономический _социальный_совет_*	*economic _social_council_*

Ранее сочетания, выделенные RAKE, были размечены как отдельная лексическая единица, поэтому при построении тематической модели они воспринимаются алгоритмом как одно слово. Тем самым, среди выделенных алгоритмом LDA встречаются также сочетания, которые состоят из ранее выделенных RAKE биграмм и триграмм, см. таблицу 11.

Таблица 11. Биграммы, выделенные в ходе обучения модели LDA (примеры потенциальных переводных эквивалентов)

Русский	**Английский**
*_стрелковый_оружие_ _легкий_вооружение_*	*_small_arm_ _light_weapon_*
*_серьезный_нарушение_ _международный_гуманитарный_право_*	*_human_right_ _fundamental_freedom_*
*_оккупировать_палестинский_территория_ _восточный_иерусалим_*	*_occupied_palestinian_territory_ _including_east_jerusalem_*

В такой модели также присутствует достаточное количество переводных эквивалентов, хоть и в меньшем объеме, чем в триграммной модели. К недостаткам можно отнести повторяемость слов и выражений в темах: в одну тему могут одновременно попасть отдельные слова и n-граммы, включающие какие-то слова темы как компоненты.

Результаты тематического моделирования с выделением биграмм для русского и английского корпуса представлены в таблице 12.

Таблица 12. Темы из модели LDA с биграммами (параллельные темы)

1 **ru**	*_стрелковый_оружие_, борьба, безопасность, _легкий_вооружение_, свободный, _стрелковый_оружие_ _легкий_вооружение_, свободный _ядерный_оружие_, торговля _стрелковый_оружие_, терроризм, _ядерный_оружие_*
en	*_criminal_justice_, _crime_prevention_ _criminal_justice_, _transnational_organized_crime_, drug, _light_weapon_, _human_right_ _fundamental_freedom_, _small_arm_ _light_weapon_, _international_covenant_, protection, _small_arm_*
2 **ru**	*человек, право, _право_человек_, _расовый_дискриминация_, религия убеждение, нетерпимость, уважение, _расовый_дискриминация_ ксенофобия, ксенофобия нетерпимость, расизм _расовый_дискриминация_*
en	*_human_right_, religion belief, protection, _high_commissioner_, refugee, _racial_discrimination_, gender, child, elimination violence, racism*
3 **ru**	*развитие, _развиваться_страна_, помощь страна, _область_развитие_, _устойчивый_развитие_, ликвидация нищета, содействие достижение, координация, экономический, торговля*
en	*_social_development_, _social_council_, economic _social_council_, financing, _world_summit_ _sustainable_development_, _developing_country_, economic, _sustainable_development_ _world_summit_, _least_developed_country_, poverty*

4 **ru**	*территория, _коренной_народ_, правительство, токелау, право самоопределение, _управлять_держава_, _несамоуправляющийся_территория_, управлять держава, положение, население*
en	*_governing_territory_, self _governing_territory_, _administering_power_, independence, grant independence, _colonial_country_, independence _colonial_country_, grant, self determination, tokelau*
5 **ru**	*человек _основный_свобода_, положение женщина, право, ребёнок, право человек, жертва насилие, женщина девочка, защита, мигрант, дискриминация*
en	*_human_right_, _human_right_ _fundamental_freedom_, woman girl , child, elimination violence, discrimination, violence woman , _woman_migrant_worker_, religion belief, protection*
6 **ru**	*гонка вооружение, _космический_пространство_, предотвращение, предотвращение гонка, вооружение _космический_пространство_, _научный_комитет_, укрепление, последствие, безопасность сотрудничество, сотрудничество европа*
en	*_outer_space_, prevention _arm_race_, joint, _special_committee_, _official_record_, _peaceful_us_ _outer_space_, scientific _technical_subcommittee_, _technical_subcommittee_* *space exploration, terrorism*
7 **ru**	*помощь, афганистан, бедствие, восстановление, _стихийный_бедствие_, последствие, _международный_сообщество_, реконструкция, гуманитарный помощь, _область_право_ человек*

en	*_disaster_reduction_, _natural_disaster_, _cultural_diversity_, world, natural disaster, _humanitarian_assistance_, _biological_diversity_, provision, peace, relief*
8 **ru**	*_ядерный_оружие_, участник договор, договор нераспространение, нераспространение _ядерный_оружие_, _обладать_ядерный_оружие_, _ядерный_разоружение_, _заключительный_документ_, безопасность, опасность, _вопрос_мир_ разоружение*
en	*_nuclear_weapon_, proliferation, disarmament, proliferation _nuclear_weapon_, _nuclear_disarmament_, _final_document_, _weapon_state_, nuclear _weapon_state_, _treaty_series_, atomic energy*
9 **ru**	*_оккупировать_палестинский_территория_, развитие, израиль, право человек, _восточный_иерусалим_, _оккупировать_палестинский_территория_ _восточный_иерусалим_, _наркотический_средство_, экономический, _населить_пункт_, _палестинский_беженец_*
en	*_occupied_palestinian_territory_, Israel, _review_conference_, _treaty_series_, _palestinian_people_, _inalienable_right_, _palestine_refugee_, _human_settlement_, culture, _human_right_*
10 **ru**	*помощь, руанда, геноцид, ответственный геноцид, _международный_трибунал_, международный _гуманитарный_право_, _судебный_преследование_ ответственный, право человек, _международный_гуманитарный_право_*
en	*_international_humanitarian_law_, rwanda, _present_resolution_, _armed_conflict_, genocide, _international_tribunal_, survivor genocide, punishment crime, justice responsible, _international_criminal_court_*

LDA с ключевыми выражениями RAKE: это стандартная модель LDA, работающая на корпусе с выделенными ключевыми выражениями без дополнительного выделения n-грамм в процессе моделирования тем. В такой модели также наблюдается высокая доля выражений и пересечение русских и английских терминов внутри, например, см. таблицу 13.

Таблица 13. Ключевые выражения, выделенные RAKE (потенциальные переводные эквиваленты)

Русский	**Английский**
*_колониальный_страна_*	*_colonial_country_*
*_право_человек_*	*_human_right_*
*_основный_свобода_*	*_fundamental_freedom_*
*_палестинский_беженец_*	*_palestine_refugee_*

Такая модель, однако, более предсказуема с точки зрения выделения устойчивых выражений, так как не предполагает дополнительного поиска би- и триграмм, а использует только уже размеченные в корпусе фразы. Соответственно, в темах преобладают униграммы, отсутствует избыточность и повтор терминов, наблюдавшиеся в предыдущих реализациях, но в то же время сохраняется репрезентативность тем.

Результаты тематического моделирования с выделением ключевых выражений RAKE для русского и английского корпуса представлены в таблице 14.

Таблица 14. Результаты тематического моделирования с выражениями RAKE для русского и английского корпуса (параллельные темы)

1 ru	*наркотик борьба преступность помощь _уголовный_правосудие_ укрепление _предупреждение_преступность_ _социальный_совет_ экономический протокол*
en	*drug, _humanitarian_personnel_, crime, _narcotic_drug_, _international_law_, _united_nation_office_, traffic, prevent, _world_drug_problem_, _crime_prevention_*
2 ru	*участник _международный_право_ защита уважение помощь _основный_свобода_ положени право _защита_право_ сотрудничество*
en	*_human_right_, economic, protection, _fundamental_freedom_, _international_law_commission_, democracy, _international_cooperation_, social, political, law*
3 ru	*гонка вооружение конфликт _космический_пространство_ _мирный_цель_ _международный_сотрудничество_ исследование предотвращение космос освоение*
	*_outer_space_, _arm_race_, _peaceful_us_, _working_group_, exploration, scientific, _technical_subcommittee_, space, _international_cooperation_, _space_science_*
4 ru	*помощь пострадать _стихийный_бедствие_ последствие страна _гуманитарный_потребность_ координация глобальный _оон_хабитат_ восстановление*
en	*_natural_disaster_, _social_council_, habitat, _human_settlement_, _developing_country_, _governmental_organization_, _economic_cooperation_organization_, _disaster_reduction_ platform, equality*
5 ru	*культура человек цивилизация мир _вопрос_образование_ наука _всемирный_программа_образование_ _культура_коммуникация_ религия диалог*

en	*Culture, religion, peace, belief, _human_right_, _united_nation_educational_, scientific, _cultural_organization_, world, education*
6 ru	*_ядерный_оружие_ договор применение разоружение нераспространение участник _свободный_владение_ соглашение _обладать_ядерный_оружие_запрещение*
en	*_nuclear_weapon_, nuclear, proliferation, disarmament, _weapon_state_, _nuclear_disarmament_, _final_document_, _review_conference_, prohibition, _free_zone_*
7 ru	*ребёнок _акт_насилие_ право торговля _положение_женщина_ девочка _обеспечение_защит _детский_труд_ _сексуальный_насилие_ _уязвимый_положение_*
	_human_right_, _woman_girl_, migrant, traffic, _related_intolerance_, _gender_equality_, child, violence, _armed_conflict_, hiv
8 ru	*территория _колониальный_страна_ _предоставление_независимость_ _управлять_держава_ право токелау _несамоуправляющийся_территория_ собственность самоопределение _новый_каледония_*
en	*_self_governing_territory_, _administering_power_, _special_committee_, self-determination, _crime_prevention_, _grant_independence_, _colonial_country_, _regional_centre_, _criminal_justice_, Tokelau*
9 ru	*соглашение _морской_право_ сторона решение судно загрязнение сохранение _морской_среда_ плавать _морской_сектор_*
	conservation, _regional_fishery_management_, _ocean_law_, _marine_environment_, _international_trade_law_, _international_maritime_organization_, _developing_country_, safety, _sustainable_development_, amount

10 **ru**	*право человек развитие _право_человек_ положение _вопрос_право_человек_ уважение поощрение _область_право_ инициатива*
en	*_human_right_, economic, protection, _fundamental_freedom_, _international_law_commission_* *_international_law_, democracy, strengthen, _international_cooperation_, social, respect*
11 **ru**	*пытка _жертва_пытка_ _стрелковый_оружие_ бесчеловечный жестокий _унижать_достоинство_вид_ _акт_насилие_ предотвращение _жестокий_обращение_ запрещение*
en	*torture, _international_covenant_, protection, woman, inhuman, victim, _degrading_treatment_, cruel, punishment, child* *_democratic_republic_, congo, territory, rwanda, _international_tribunal_, genocide, _international_criminal_court_, _former_yugoslavia_, bosnia, Herzegovina*
12 **ru**	*_правительство_афганистан_ афганистан ликвидировать восстановление _женщина_девочка_ боевой положение талибан _палестинский_беженец_ _политический_урегулирование_*
en	*refugee, Afghanistan, _human_right_, protection, reconstruction, _displaced_person_, _internally_displaced_person_, _governmental_organization_, woman, rehabilitation*
13 **ru**	*мигрант _право_человек_ _защита_мигрант_ _область_право_ _расовый_дискриминация_ _иметь_документ_ положение ксенофобия расизм жертва*

en	*_human_right_, woman, migrant, justice, _migrant_worker_, _international_migration_, _expected_accomplishment_, issue, concern, _racial_discrimination_*
14 ru	*разоружение _легкий_оружие_ безопасность сотрудничество содействие мир _стрелковый_оружие_ _незаконный_оборот_ _массовый_уничтожение_ _ядерный_оружие_*
en	*disarmament, peace, _building_measure_, _small_arm_, strengthen, _light_weapon_, conflict, _international_community_, traffic, weapon*
15 ru	*_положение_женщина_ ребёнок _равенство_женщина_ _роль_женщина_ мужчина афганистан насилие _отношение_женщина_ девочка торговля*
en	*refugee, Afghanistan, _human_right_, protection, _international_community_, _internally_displaced_person_, woman, rehabilitation, _human_right_education_, reconstruction*
16 ru	*персонал безопасность агентство защита секреариат _преследование_персонал_ _переместить_лицо_ _международный_гуманитарный_право_ _гуманитарный_персонал_ _управление_верховный_комиссар_*
en	*_state_party_, institute, safety, crime, _humanitarian_personnel_, _international_law_, drug, _international_humanitarian_law_, law, strengthen*
17 ru	*_двухгодичный_период_ рекомендация положение бюджет работа административный комитет сведение персонал отчёт*

en	*biannial, _advisory_committee_, staff, board, _budgetary_question_, amount, provision, _related_report_, _proposed_programme_budget_, _official_record_*
18 ru	*израиль _восточный_иерусалим_ _оккупировать_палестинский_территория_ _район_ближний_восток_ оружие _палестинский_беженец_ _правительство_израиль_ регион _ядерный_оружие_ _оккупировать_держава_*
en	*_palestinian_people_, _occupied_palestinian_territory_, Israel, _including_east_jerusalem_, _occupying_power_, protection, _palestine_refugee_, _human_right_, war, peace*
19 ru	*развитие _развиваться_страна_ страна _область_развитие_ _устойчивый_развитие_ достижение решение торговля помощь _искоренение_нищета_*
en	*habitat, equality, _human_settlement_, _developing_country_, Cambodia, youth, _sustainable_development_, environment, island, _world_summit_*
20 ru	*_африканский_страна_ _африканский_союз_ алмаз _устойчивый_развитие_ _развитие_африка_ _развиваться_страна_ здравоохранение малярия вичспид абуджа*
en	*_african_union_, _small_arm_, strengthen, _african_country_, disarmament, peace, hiv, protection, malaria, girl*

Сравнивая полученные результаты, можно заметить, что разные варианты реализации алгоритма тематического моделирования могут применяться для разных целей. Так, модель LDA с выделением биграмм и триграмм отлично подходит для выявления переводных эквиваленотов в параллельных корпусах текстов. Такая модель успешно справляется с восстановлением смысловых связей между коллокациями и отдельными словами и способна генерировать полноценные выражения. К недостаткам модели можно отнести большее по сравнению с другими моделями количество времени, которое тратится на работу, и избыточность

генерируемых фраз, которые иногда могут включать в себя части уже найденных выражений, таким образом, повторяясь и расширяясь с каждой итерацией. Тем не менее, модель успешно находит содержащиеся в коллекции документов темы и наполняет их семантически точными выражениями. Такую модель можно использовать для более точного и узкого анализа небольших корпусов или для уточнения уже описанных тем.

Менее громоздкой является модель LDA с выделением биграмм, в которой возможны сочетания отдельных слов из коллекции и дополнения уже размеченных сторонним алгоритмом n-грамм. Такая модель хорошо справляется с задачей повышения количества n-грамм в темах, так как иногда в процессе тематического моделирования можно столкнуться с проблемой недостаточной наполненности тем ранее найденными биграмм и триграммами. Помимо этого, модель также успешно генерирует фразы-кандидаты для переводных эквивалентов, и может быть успешно задействована в задачах машинного перевода. Работа модели, к сожалению, немного замедляется из-за процесса поиска и генерации дополнительных биграмм и поэтому она, как и предыдущая, может проигрывать стандартным реализациям по скорости и простоте использования.

Наконец, модель LDA с выражениями RAKE, с помощью которой было найдено и описано двадцать тем из каждого корпуса, такова, что она быстрее справляется с поставленной задачей, поэтому лучше подходит для создания тематической модели объемных корпусов на несколько десятков тем. Она включает в себя меньшее количество выражений, однако, несмотря на это, не теряет своего реперезентативного потенциала и выделяет значимые униграммы наряду с самыми частотными биграммами и триграммами RAKE. Такая модель хорошо описывает общую тематическую наполненность корпуса и может применяться для работы с большими коллекциями данных.

Выводы к Главе 2

В данной главе мы описали исходные лингвистические данные, на которых проводились серии экспериментов по тематическому моделированию. Материалами для исследования послужил находящийся в свободном доступе корпус параллельных текстов резолюций Организации Объединенных Наций за 2000 год на русском и английском языках.

Описан процесс предобработки корпуса. Для русского языка был дополнительно введен этап морфологической разметки, что позволило улучшить качество полученных на следующем этапе с помощью алгоритма RAKE ключевых выражений. Кроме этого аспекта этапы предобработки корпусов не отличались: для каждой коллекции был выполнен процесс поиска ключевых выражений, лемматизация, фильтрация по стоп-словарю.

Далее были описаны эксперименты по построению мультимодальных двуязычных тематических моделей LDA с разными параметрами: с дополнительным выделением биграмм и триграмм, с выделением биграмм, основная модель с ключевыми выражениями RAKE. Мы построили модели для русского и английского корпуса, а затем провели выравнивание выделенных тем, охарактеризовав их наполнение.

Заключение

Данное исследование было посвящено изучению и практической реализации алгоритма мультимодального тематического моделирования в задаче семантической компрессии, осуществимого на материале корпуса параллельных текстов резолюций Организации Объединенных Наций за 2000 год.

Для этого были изучены и описаны процедуры семантической компрессии, а также ее реализация на более глобальном уровне – тематическое моделирование. Было исследовано мультимодальное тематическое моделирование, соединяющее в себе несколько алгоритмов выделения n-грамм и ключевых выражений, и основывающееся на работе с многоязычной коллекцией текстовых данных.

Был обоснован выбор алгоритма выделения ключевых слов и выражений и описана специфика его работы на многоязычном корпусе языковых данных. Далее была сформирована комбинированная методика расширения тематической модели LDA n-граммами и ключевыми выражениями, полученными на этапе предобработки или непосредственно во время построения тематической модели.

Была произведена предобработка лингвистических данных и проведен эксперимент по построению тематической модели на корпусе параллельных текстов на русском и английском языках с применением различных вариантов реализации алгоритма тематического моделирования LDA.

Оценка результатов показала, что разные тематические модели могут применяться в разных целях: для поиска кандидатов в переводные эквиваленты, в качестве источников для многоязычных словарей, в задаче расширения тематической модели биграммами и непосредственно моделирования тематической структуры корпуса текстов.

Список использованной литературы

1) Белякова А.Ю., Беляков Ю.Д. Обзор задачи автоматической суммаризации текста // Инженерный вестник Дона. 10(70). 2020. С. 142-159.
2) Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие. М.: МИЭМ, 2011.
3) Браславский П., Соколов Е. Сравнение пяти методов извлечения терминов произвольной длины // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 4–8 июня 2008 г.). М., 2008. С. 67-74.
4) Вознесенская Т. В., Леднов Д.А. Система автоматического аннотирования текстов с помощью стохастической модели // Машинное обучение и анализ данных. 2018. Т. 4. № 4. С. 266-279.
5) Воронцов К.В. Вероятностное тематическое моделирование. Электронный учебник. 2013. URL: http://www.machinelearning.ru/wiki/images/2/22/Voron-2013-ptm.pdf
6) Добров А.В. Автоматическая рубрикация новостных сообщений средствами синтаксической семантики. Дис. … канд. филол. наук. СПб, 2014.
7) Добров А.В. Автоматическая рубрикация текстов средствами комплексного лингвистического анализа // Структурная и прикладная лингвистика. Вып. 9. СПб., 2012. С. 135-147.
8) Добровольский Д.О., Кретов А.А., Шаров С.А. Корпус параллельных текстов: архитектура и возможности использования // Национальный корпус русского языка: 2003-2005. Результаты и перспективы. М., 2005. С. 263-296.

9) Захаров В.П. Хохлова М.В. Анализ эффективности статистических методов выявления коллокаций в текстах на русском языке // Компьютерная лингвистика и интеллектуальные технологии. Том 9 (16). М., 2010. С. 137-143.
10) Ирхин И.А., Булатов В.Г., Воронцов К.В. Аддитивная регуляризация тематических моделей с быстрой векторизацией текста // Компьютерные исследования и моделирование. 2020. Т. 12. №. 6. С. 1515-1528.
11) Москвина А.Д., Митрофанова О.А., Ерофеева А.Р., Харабет Я.К. Автоматическое выделение ключевых слов и словосочетаний из русскоязычных корпусов текстов с помощью алгоритма RAKE // Труды Международной конференции “Корпусная лингвистика-2017 . СПб., 2017.
12) Нестерова Н.М., Герте Н.А. Реферирование как способ извлечения и представления основного содержания текста // Вестник Пермского университета. Российская и зарубежная филология. 4(24). 2013. С. 127-132.
13) Нокель М.А. Методы улучшения вероятностных тематических моделей текстовых коллекций на основе лексико-терминологической информации: Дис. … канд. физ-мат. наук. М., 2016.
14) Нокель М.А., Лукашевич Н.В. Тематические модели: добавление биграмм и учет сходства между униграммами и биграммами // Вычислительные методы и программирование. 2000. Т. 6.
15) Седова А.Г., Митрофанова О.А. Тематическое моделирование русскоязычных текстов с опорой на леммы и лексические конструкции // Компьютерная лингвистика и вычислительные онтологии. СПб., 2017.

16) Ягунова Е.В., Пивоварова Л.М. Природа коллокаций в русском языке. Опыт автоматического извлечения и классификации на материале новостных текстов // Научно-техническая информация. Серия 2. 2010. С. 30-40.
17) Bird S., Klein E., Loper E. Natural language processing with Python: analyzing text with the natural language toolkit. O'Reilly Media, Inc., 2009.
18) Blei D.M, Ng A., Jordan M. Latent Dirichlet Allocation // Journal of Machine Learning Research. 2003. Vol. 3.
19) Campos R., Mangaravite V., Pasquali A., Jorge A., Nunes C., Jatowt A. A Text Feature Based Automatic Keyword Extraction Method for Single Documents // Proceedings of the 40th European Conference on Information Retrieval (ECIR'18), Grenoble, France, 2018.
20) Campos R., Mangaravite V., Pasquali A., Jorge A., Nunes C., Jatowt A. YAKE! Collection-independent Automatic Keyword Extractor // Proceedings of the 40th European Conference on Information Retrieval (ECIR'18), Grenoble, France, 2018.
21) Hofmann T. Probabilistic latent semantic analysis // Proceedings of the 15th Annual Conference on Uncertainty in Artificial Intelligence (UAI), Stockholm, Sweden, July 30-August 1, 1999.
22) Huang J. et al. Improving biterm topic model with word embeddings // World Wide Web . 23.6. 2020. P. 3099-3124.
23) Korobov M. Morphological Analyzer and Generator for Russian and Ukrainian Languages // Khachay, M., Konstantinova, N., Panchenko, A., Ignatov, D., Labunets, V. (eds.) Analysis of Images, Social Networks and Texts. AIST 2015. Communications in Computer and Information Science. Vol 542. Springer, 2015.

24) Lau J.H., Baldwin T., Newman D. On Collocations and Topic Models // ACM 131 Transactions on Speech and Language Processing. ACM Press. Vol. 10, №3. 2013.

25) Loukachevitch N., Nokel M., Ivanov K. Combining Thesaurus Knowledge and Probabilistic Topic Models // International Conference on Analysis of Images, Social Networks and Texts. Springer, Cham, 2017. P. 59-71.

26) Nokel M., Loukachevich N. Accounting ngramms and multi-word terms can improve topic models // Proceedings of the 12th Workshop on Multiword Expressions, Berlin, Germany, August 7-12, 2016. P. 44–49.

27) Roller S., Im Walde S. A multimodal LDA model integrating textual, cognitive and visual modalities // Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013.

28) Rose S. et al. Automatic keyword extraction from individual documents // Text mining: applications and theory. 1 (2010). P. 1-20.

29) Rose S.J., Cowley W.E., Crow V.L., Cramer N.O. Rapid Automatic Keyword Extraction for Information Retrieval and Analysis. 2009. URL: http://www.google.co.ve/patents/US8131735

30) Rosen-Zvi M. et al. The author-topic model for authors and documents // arXiv preprint arXiv:1207.4169 (2012).

31) Segalovich I. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine // Proceedings of the International Conference on Machine Learning; Models, Technologies and Applications. MLMTA'03, June 23 –26, 2003. Las Vegas, Nevada, USA, 2003.

32) Sha H. et al. Dynamic topic modeling of the COVID-19 Twitter narrative among US governors and cabinet executives // arXiv preprint arXiv:2004.11692 (2020).

33) Sokolova E.V., Moskvina A.D. Mitrofanova O.A. Keyphrase extraction from the Russian corpus on Linguistics by means of KEA and RAKE algorithms // Data Analytics and Management in Data Intensive Domains: Proceedings of the XX International Conference. DAMDID/RCDL'2018, October 9-12, 2018, Moscow. P. 369-372.

34) Vulić I. et al. Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications // Information Processing & Management. 51(1). 2015. P. 111-147.

35) Wallach H.M. Topic modeling: beyond bag-of-words // Proceedings of the 23rd International conference on Machine learning. 2006.

36) Witten I.H., Paynter G.W., Frank E., Gutwin C., Nevill-Manning C.G. KEA: Practical Automatic Keyphrase Extraction // Proceedings of the 4th ACM conference on Digital libraries. 1999. URL: http://www.cs.waikato.ac.nz/~eibe/pubs/chap_Wit ten-et-al_Windows.pdf

37) Yan X. et al. A biterm topic model for short texts // Proceedings of the 22nd international conference on World Wide Web. 2013.

38) Ziemski M., Junczys-Dowmunt M., Pouliquen B. The United Nations Parallel Corpus v1.0. // Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), Portorož, Slovenia. European Language Resources Association (ELRA). 2016. P. 3530–3534.

39) Zosa E., Granroth-Wilding M. Multilingual dynamic topic model // RANLP 2019-Natural Language Processing a Deep Learning World Proceedings. 2019.

Список электронных ресурсов

1) Универсальная научно-популярная энциклопедия «Кругосвет» (https://www.krugosvet.ru/)

2) Образовательный портал по машинному обучению «Machine Learning Plus» (https://www.machinelearningplus.com/)

3) Сайт по научной литературе «Science Direct» (https://www.sciencedirect.com/)

4) Сообщество IT-специалистов «Хабр» (https://habr.com/ru/)

5) Студенческий практикум программистов при Московском Авиационном Институте «Лямбда» (https://lambda-it.ru/about)

6) Портал о машинном обучении (http://www.machinelearning.ru/wiki/index.php?title=%D0%97%D0%B0%D0%B3%D0%BB%D0%B0%D0%B2%D0%BD%D0%B0%D1%8F_%D1%81%D1%82%D1%80%D0%B0%D0%BD%D0%B8%D1%86%D0%B0)