DOI: 10.36535/0203-6460-2021-01-4

Сходство документов на основе аспекта на примере научных статей*

Мальте ОСТЕНДОРФ
(Malte OSTENDORFF),

ГЕОРГ РЕМ
(Georg REHM)

Немецкий научно-иссследовательский центр по искусственному интеллекту, г. Берлин, Германия

Терри РУАС
(**Terry RUAS**)

Вуппертальский университет, г. Вупперталь, Германия

Тилль БЛЮМЕ
(**Till BLUME**)

Кильский университет, г. Киль, Германия

Бела ГИПП
(**Bela GIPP**)

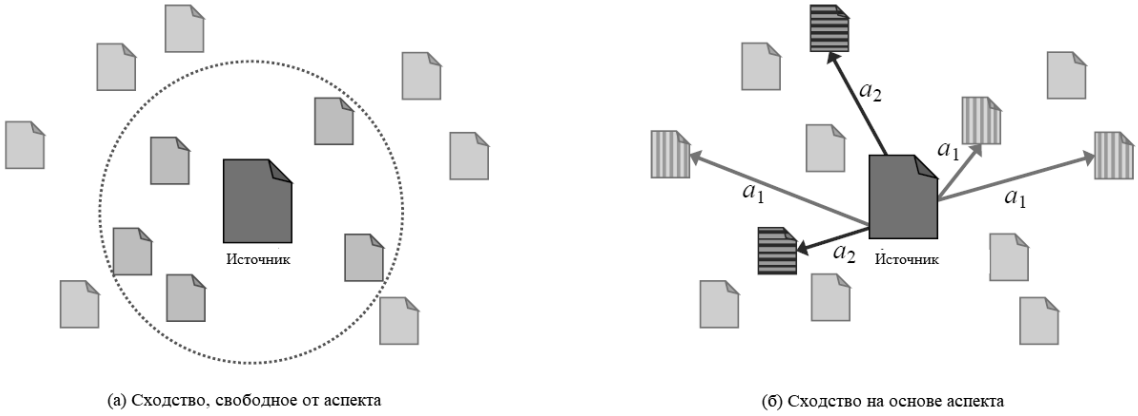
Университет г. Констанц, г. Констанц, Германия

Традиционные измерения сходства документов обеспечивают крупномодульное разграничение между схожими и несхожими документами. Обычно эти измерения не рассматривают в каких аспектах два документа являются схожими. Это ограничивает степень структурирования прикладных задач, таких как рекомендательные системы, которые полагаются на сходство документов. В статье понятие сходства расширяется аспектом информации через выполнение задачи классификации пар документов. Оценивается сходство документов на основе аспекта на примере научных публикаций. Ссылки в статьях отражают сходство по аспекту, например, часть названия, в котором встречается ссылка, выполняет функции категории для пары цитирующей и цитируемой статьи. Использовался ряд вариаций моделей Transformer, таких как ROBERTa, ELECTRA, XLNet и BERT, и они сравнивались с ведущей моделью LSTM. Наши эксперименты проводились на двух недавно созданных наборах данных, подсчитывающих 172 073 научные статьи из собраний ACL Anthology и CORD-19. Относительно выполнения результаты определяют в качестве лучшей систему SciBERT. Качественное исследование обосновывает наши количественные результаты. Выводы стимулируют проведение дальнейших исследований сходства документов на основе аспекта и разработку рекомендательных систем на основе оценки технологий. Наборы данных, коды и подготовленные модели являются публично доступными.

ВВЕДЕНИЕ

Рекомендательные системы (РС) помогают ученым в поиске релевантных статей для их работы. Когда обратная связь от пользователя осуществляется редко или недоступна, то применяются подходы на основе контента и соответствующие измерения сходства документов [1]. Рекомендательные системы советуют документ-кандидат в зависимости от его сходства или несходства по отношению к документу-источнику. Эта крупномодульная оценка сходства (похож или непохож) отрицает многие фасеты, способные сделать два документа схожими. Относительно общего понятия сходства авторы работ [2, 3] даже утверждают, что сходство является плохо определяемым понятием, если нельзя утверждать к какому аспекту относится сходство. В РС для научных статей сходство часто связано с множеством фасетов представленного исследования, например, метод, полученные данные [4]. Учитывая, что сходство документов может дифференцировать аспекты исследования, есть возможность получить определенные смоделированные рекомендации. Например, разрешается рекомендовать статью со схожими методами, но различными полученными данными.

* Перевод Ostendorff M., Ruas T., Blume T., Gipp B., Rehm G. Aspect-based document similarity for research papers. — https://arxiv.org/pdf/2010.06395.pdf



(а) Сходство, свободное от аспекта

(б) Сходство на основе аспекта

Рис. 1. Большинство РС полагается на измерения сходства между источником и *k* большинтсвом схожих целевых документов (а). Это отрицает аспекты, по которым два и более документов могут быть схожими. В сходстве документов на основе аспекта (б) документы объединены в соответствии с их внутренними, связанными с ними аспектами (a_1 и a_2).

Таким образом рекомендательная система позволяет облегчать обнаружение аналогий в научной литературе [5]. Описываем соответствующее сходство множественного аспекта в научных статьях как *сходство документов на основе аспекта.* На рис. 1 изображено сходство на основе аспекта в отличие от сходства, свободного от аспекта (традиционное). Следуя примеру научной статьи, аспект a_1 относится к полученным данным, а аспект a_2 – к методу (рис. 1б).

В предыдущей работе [6] мы предлагаем вывести аспект сходства документов, формулируя проблему как многоклассовую классификацию пар документов. В данной статье мы расширяем нашу предыдущую работу до многокатегорийного сценария и сосредотачиваемся на научной литературе, а не на общей (статьи в Wikipedia). Подобно авторам работ [7, 8] используем ссылки как учебные сигналы. Вместо использования ссылок для бинарной классификации (т.е. схожий или несхожий документ) мы включаем название раздела, в котором фигурирует ссылка, как категорию для пары документов. Названия разделов в ссылках описывают сходство по аспекту цитирующего и цитируемого документа. Наши наборы данных берутся из ACL Anthology [9] и CORD-19 [10].

В итоге наш вклад состоит в следующем: (1) расширение традиционного сходства документов до основанного на аспекте в задаче многокатегорийной многоклассовой классификации документа; (2) показ того, что сходство документов на основе аспекта хорошо подходит для научных статей; (3) оценка шести моделей Transformer и основной для задачи классификации пар документов; (4) публикация наших кодов источников, подготовленных моделей и двух наборов данных из областей компьютерной лингвистики и биомедицины для усиления дальнейшего исследования.

СВЯЗАННЫЕ РАБОТЫ

Далее обсуждается работа авторов по сходству текста, рекомендации и применения Transformer.

Авторы [3] обсуждают понятие сходства как часто плохо определяемого в литературе и используемого в качестве «совокупности терминов, охватывающих довольно разные явления». Эти авторы [3] также формализуют то, чем является сходство текста, и предполагают, что контент, структура и стиль являются основными измерениями, присущими тексту. Что касается рекомендации литературы, то информация о контенте и пользователе является самым распространенным измерением для рассмотрения [1].

Ряд авторов [5] изучает сходство документа на основе аспекта как задачу сегментации, а не классификации. Они (авторы) делят аннотации совместных работ и работ по вычислениям на четыре класса в зависимости от их научного аспекта: описание, цель, метод и полученные данные. Сходство по косинусу, вычисленное на сегменте презентаций, позволяет провести поиск сходных статей по отдельному аспекту. Авторы работы [4] применяют тот же подход сегментации к массиву CORD-19 [10]. Совместная работа авторов [11] придерживается аналогичного подхода к рекомендациям ссылок. Эти авторы классифицируют разделы по дискурсам фасетов и строят векторы документа для каждого фасета. Однако сегментация является сверхоптимальной альтернативой, так как она нарушает когерентность документов. Что касается классификации пар документов, то сходство на основе аспекта происходит без нарушений когерентности документа.

Наши эксперименты изучают языковые модели Transformer [12]. Модели BERT [13], ROBERTa [14], XLNet [15] и ELECTRA [16] улучшают многие задачи NLP (Natural Language Processing - Обработка естественного языка), например, вывод естественного языка [17, 18] и семантическое сходство текстов [19]. Авторы работы [20] показывают, как модели BERT можно объединять в сеть Siamese [21] для создания векторных представлений, подходящих для сравнения друг с другом при помощи сходства по косинусу. Ряд исследований [22, 23] анализирует модели BERT, чтобы классифицировать одиночные документы в соответствии с восприятием или темой. Исследования авторов [8, 24] рассматривают предназначенные для определенной области модели Transformer по отношению к задачам NLP в научных документах.

Более того, ученые [8] являются первыми, кто использовал модели Transformer для кодирования названий и аннотаций статей в целях создания рекомендаций. Авторы работы [25] используют модели BERT для рекомендательных систем, но кодируют только названия статей. Иные недавние рекомендательные системы полагаются на другие технологии, такие как анализ совместного цитирования, TF-IDF или Paragraph Vectors [26, 27].

В предыдущей работе [6] мы моделируем сходство на основе аспекта как задачу многоклассовой классификации пар документов. Используем края интеллектуального графа из Wikidata как аспект информации сходства статей в Wikipedia. Применяемое определение задачи допускает только монокатегорийнную классификацию. Для научных статей это определение не вполне подходит. Две статьи могут быть схожи во многих аспектах. Соответственно мы ставим задачу многоклассовой классификации и расширяем ее до многокатегорийной.

Для наших экспериментов мы воспользуемся ссылками и разделом названий, в которых встречаются ссылки, как категориями классификации. Авторы [28] показывают соответствующий подход в контексте связывания объектов. Они утверждают, что во многих ситуациях ссылка на объект предлагает относительно крупномодульную семантическую информацию. Чтобы учитывать различные аспекты, в которых упоминается объект, эти авторы [28] связывают объекты не только с их соответвующими статьями в Wikipedia, но и с разделами, представляющими различные аспекты.

Что касается сходства на уровне сегмента и парной многоклассовой монокатегорийной классификации, то первоначальные подходы, изучающие сходство на основе аспекта, являются доступными. В частности, модели Transformer, кажется, обещают успешное решение задач относительно сходства, классификации и иных соответствующих вопросов.

ЭКСПЕРИМЕНТЫ

Представляем нашу методологию (рис. 2) классификации сходства научных статей на основе аспекта.

Наборы данных

Генерация аннотированных людьми данных для рекомендации научных статей затратна и ограничена небольшим количеством [1]. Набор данных небольшого размера мешает внедрению обучающих алгоритмов. Чтобы избежать проблемы нехватки данных, ученые полагаются на ссылки как на основу истины, т.е. когда ссылка существует между двумя статьями, обе статьи считаются схожими [7;8]. Либо ссылка существует, либо не соответствует категории в бинарной классификации. Для создания сходства на основе аспекта мы переносим эту идею в проблему многокатегорийной многоклассовой классификации. В качестве основы истины адаптируем название раздела, в котором ссылка из статьи A (источник) на B (цель) встречается как название класса (рис. 2a). Эта классификация является многоклассовой из-за множества названий разделов, а также многокатегорийной, так как статья A может цитироваться в нескольких разделах. Например, статья A, цитирующая B, в разделе Введение (Introduction) и Обсуждение (Discussion) должна соотноситься с одной выборкой набора данных.

ACL Anthology. Принимаем библиографический массив ACL Anthology [9] как набор данных. Он содержит 22 878 научных статей по вычислительной лингвистике. Помимо полных текстов массив ACL Anthology предоставляет дополнительные данные по цитированию. Ссылки аннотируются с помощью названия раздела, в котором расположены маркеры ссылок. Эта информация востребована в наших экспериментах.

CORD-19. Открытый набор научных данных по COVID-19 (CORD-19) – собрание статей по COVID-19 и относящихся к коронавирусу исследованиям из нескольких биомедицинских цифровых библиотек [10]. Ссылки и метаданные всех статей CORD-19 стандартизированы в соответствии с регулярной обработкой авторов [29]. Ссылки в CORD-19 также аннотируются с помощью названий разделов.

Предварительная обработка данных

Рассматривая ACL Anthology и CORD-19, получаем два набора данных для парной многокатегорийной многоклассовой классификации. Названия разделов из ссылок, т.е. названия классов, представлены в табл. 1. Нормализуем названия разделов (lowercase, letters-only, singular to plural) и разложим составные разделы на много простых – Conclusion and Future Work (Заключение и Дальнейшая работа) на Conclusion; Future Work (Заключение; Дальнейшая работа). Сделаем запрос в прикладной программный интерфейс DBLP [30] и Semantic Scholar [29] для соотнесения ссылок и извлечем недостающую информацию из статей, например, рефератов. Также удалим необоснованные статьи без текста или дублирующиеся. Разделим оба набора данных ACL Anthology и CORD-19 на десять классов в соотвествии с их числом выборок, посредством которых первые девять содержат наиболее популярные названия разделов, а в десятую (Прочее) сгруппируем оставшиеся. Даже если решение на основе наших десяти классов может отрицать вариации названий разделов в литературе, наша модель все еще дублирует ряд определенных авторами [4 и 5] аспектов исследования. Итоговое распределение классов является несбалансированным, но отражает истинную природу совокупностей, как показыает табл. 5. Подлинники для воспроизведения массивов данных доступны при наличии нашего кода источника.

Негативная выборка

Помимо десяти положительных классов (табл. 1) введем класс, названный *None*, который будет выступать в роли отрицательного оппонента наших положительных выборок в той же пропорции [31]. Пара документов класса *None* являются случайно выбранными и непохожими. Случайная пара статей является негативной выборкой, когда статьи не сущестуют как положительная пара, не являются совместно цитируемыми, не объединены авторами и не опубликованы в одном и том же номере. Получаем 24 275 негативных выборок для набора данных ACL Anthology и 33 083 для набора данных CORD-19.Эти выборки позволят различать модели между схожими и несхожими документами.

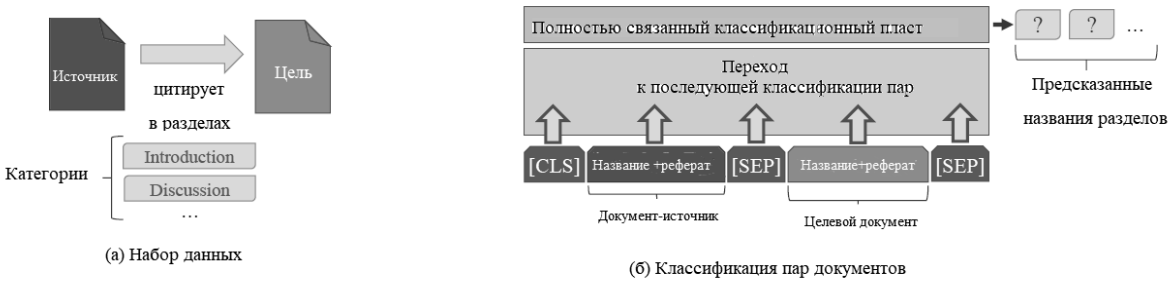


Рис. 2. Использование названий разделов из ссылок как категории для пар документов.
Эти разделы определяют аспекты сходства. Модель Transformer с названиями и рефератами в качестве входных данных используется для классификации.

Таблица 1

Распределение названий классов, извлеченных из названий цитирующих разделов в двух наборах данных

Название класса	**Подсчет**	**Название класса**	**Подсчет**		**Название класса**	**Подсчет**	**Название класса**	**Подсчет**
Введение	16 279	Заключение	1 158		Введение	15 108	Описание	454
Связанные работы	12 600	Обсуждение	1 132		Обсуждение	13 258	Материалы	420
Эксперимент	4 025	Оценка	971		Заключение	1 003	Вирус	218
Описание	1 365	Методы	719		Результаты	910	Дальнейшая работа	171
результаты	1 181	*Прочее*	22 249		Методы	523	*Прочее*	43 154
(а) ACL Anthology					(б) CORD-19			

Примечание: Верхние девять разделов-классов приводятся в убывающем порядке, оставшиеся группируются как Прочее.

Системы

Сфокусируемся на последовательной классификации пар с использованием моделей на основе архитектуры Transformer [12]. Такие модели на основе Transformer часто используются в задачах сходства текстов [7, 20]. Более того, авторы [6] обнаружили неоригинальные модели Transformer, т.е. BERT [13], XLNet [15], сети vanilla Siamese [21] и традиционные векторные представления слов (например Glo Ve [32], Paragraph Vectors [33]) в задаче классификации пар документов. Следовательно, исключаем сети Siamese и предварительные опытные модели векторных представлений слов из наших экспериментов. Вместо этого изучаем шесть вариаций Transformer и дополнительную основу для сравнения. Названия и рефераты пар научных статей, используемые как входные данные в модель, посредством которых символ [SEP] – Source Evaluation Panel, группа оценки источника – разделяет источник и целевую статью (рис. 2б). Данная процедура основана на нашей предыдущей работе [6]. В наших экспериментах мы не используем полные тексты, так как многие статьи не доступны бесплатно и отобранные модели Transformer накладывают жесткое ограничение в 512 символов.

Основа LSTM. В качестве основы используем бинаправленную LSTM [34]. Чтобы получить представления пар документов, введем названия и рефераты двух документов в LSTM, посредством которой статьи отделяются особым разделителем символов. Используем токенайзер (лексический анализатор) библиотеки SpaCy [35] и векторы слов из библиотеки fastText [36]. Векторы слов предварительно проверяются на аннотациях наборов данных ACL Anthology и CORD-19.

BERT, Covid-BERT & SciBERT. BERT- нейронная языковая модель на основе архитектуры Transformer [13]. Признано, что модели BERT предварительно проверяются на большом текстовом массиве без пересмотра. Две предварительно опытные цели – восстановление замаксированных средств идентификации пользователя (т.е моделирование языка масок) и последующее NSP (Next sentence prediction). После предварительного опыта модели BERT хорошо мотивированы для определенных задач, такие как сходство предложений [20] или классификация документов [23]. Некоторые модели BERT, предварительно проверенные на различных совокупностях, публично доступны. Для наших экспериментов мы оцениваем три вариации BERT: (1) модель BERT от авторов [13], проверенная на English Wikipedia и the BooksCorpus [37]. (2) SciBERT [24], вариация BERT, предназначенная для научной литературы, которая предварительно обработана на научных статьях по вычислительной технике и биомедицине; (3) Covid-BERT [38] – оригинальная модель BERT от авторов [13], но хорошо настроенная на CORD-19.

BioBERT [39] – другая модель BERT, специализирующаяся на биомедицинской области. Но мы исключаем BioBERT из наших экспериментов, так как SciBERT

превосходит ее в биомедицинских задачах [24]. Также опускаем вариации BERT от авторов [8], поскольку они используют цитирование на протяжении предварительно обработанной рискованной утечки данных в наш тестовый набор. Все три модели – BERT, SciBERT и Covid-BERT пожохи по своей структуре, за исключением набора, используемого на протяжении подготовки языковой модели.

RoBERTa. Авторы [14] предложили RoBERTa, которая является моделью BERT, предназначенной для более крупных массивов, более длительного времени проверки, и убирает задачу NSP из своей цели. Более того, RoBERTa использует дополнительные совокупности для предварительной обработки, главным образом такие как Common Crawl News [40], OpenWebText [42] и STORIES [42].

XLNet. В отличие от BERT модель XLNet [15] является не автокодировщиком, а авторегрессивной языковой моделью. XLNet не применяет NSP. Мы используем опубликованную ее авторами модель, XLNet, которая предварительно обработана на совокупностях – Wikipedia, BooksCorpus [37], Giga5 [43], ClueWeb 2012-B [44] и Common Crawl [45].

ELECTRA. ELECTRA [16] должна дополнительно маскировать языковое моделирование предварительно обработанной целью, заключающейся в выявлении перемещенных средств идентификации пользователя во входящей последовательности. Для этой цели авторы [16] используют генератор, который перемещает средства идентификации и дискриминатор сети, выявляющий перемещения. Генератор и дискриминатор – модели Transformer. ELECTRA не прибегает к задаче NSP. Для наших экспериментов применяется дискриминатор модели ELECTRA. Предварительно обученный дискриминатор модели ELECTRA заранее обрабатывается на тех же данных, что и BERT.

Гиперпараметры и применение. Мы выбираем нужные гиперпараметры LSTM в соответствии с полученными авторами [46] следующими данными: 10 периодов для подготовки, размер группы b=8, скорость изучения $\eta = 1^{-5}$, два уровня LSTM со 100 скрытыми размерами, внимание и выдача с вероятностью d=0,1. Тогда как основа LSTM использует vanilla PyTorch, все методы на основе Transformer применяются с использованием the Huggingface API [47]. Каждая модель Transformer используется в ее версии BASE. Гиперпараметры для хорошо мотивированной Transformer строятся с помощью работы авторов [13]: четыре подготовленных периода, скорость изучения $\eta = 2^{-5}$, размер группы b=8, и оптимизатор Adam с $\varepsilon = 1^{-8}$. Проводим оценку в стратифицированной k-кратной перекрестной проверке с k=4 (т.е. класс распределения остается идентичным для каждого повторения). Это приводит в среднем к 54 618, 75 /18 206, 2^5 подготовленных/опытных образцов для набора данных ACL Anthology и 74 436/ 24 812 – для набора данных CORD-19. Код источника, массивы данных и подготовленые модели публично доступны* Мы предоставляем Google Colab для испытания подготовленных моделей на любых статьях из Semantic Scholar**.

РЕЗУЛЬТАТЫ

Наши результаты разделены на три части: полная оценка, оценка категорий классов и количественная оценка***.

Полная оценка

Подробные результаты нашей количественной оценки представлены в табл. 2. Мы проводим оценку как 4- кратную перекрестную проверку на основе наших наборов данных. Сообщаем микро- и макро- среднее для полноты, точности и значения F1, чтобы принять во внимание несбалансированное распределение по категориям и классам (см. раздел «Наборы данных»).

Таблица 2

Полное значение F1 (со стандартным отклонением), полнота и точность для макро- и микро- среднего 7 методов для набора данных ACL Anthology и CORD-19

	ACL Anthology						CORD-19					
Массив данных	macro avg			micro avg			macro avg			micro avg		
	F1 (std)	P	R	F1(std)	P	R	F1 (std)	P	R	F1 (std)	P	R
^{LSTM}baseline	.063 ±.001	.069	.058	.290 ±.004	**.761**	.179	.128 ±.001	.137	.121	.579 ±.005	.758	.469
BERT	.256 ±.002	.317	.238	.641 ±.002	.719	.578	.387 ±.011	**.619**	.357	.822 ±.002	.840	.806
Covid-BERT	.270 ±.006	.404	.253	.648 ±.005	.715	.592	.394 ±.010	.578	.364	.818 ±.001	.836	.802
SciBERT	**.326 ±.005**	**.458**	**.303**	**.678 ±.002**	.725	**.637**	**.439 ±.010**	.560	**.401**	**.833 ±.003**	**.846**	**.820**
RoBERTa	.250 ±.003	.285	.232	.626 ±.003	.703	.564	.332 ±.008	.473	.316	.820 ±.001	.840	.801
XLNet	.263 ±.011	.372	.250	.645 ±.011	.705	.595	.362 ±.025	.523	.345	.817 ±.002	.832	.804
ELECTRA	.245 ±.005	.287	.228	.616 ±.021	.693	.554	.280 ±.001	.306	.276	.820 ±.002	.840	.801

Примечание:SciBERT выдает лучшие результаты для обоих наборов данных.

* GitHub repository: https://github.com/malteos/aspect-document-similarity

** https://colab.research.google.com/github/malteos/aspect-document-similarity/blob/master/demo.ipynb

*** Оценка по категориям и количественная оценки перестают использовать один из двух наборов данных из-за пространственных ограничений, но доступны на GitHub.

С учетом полных оценок SciBERT является лучшим методом с 0,326 макро-F1 и 0,678 микро-F1 в наборе данных ACL Anthology и с 0,439 макро-F1 и 0,833 микро-F1 в наборе данных CORD-19. Все модели Transformer по всем показателям превосходят LSTM за исключением микро-точности в наборе ACL Anthology. Этот разрыв между макро- и микро- средними результатами существует из-за несоответствия категорий классов (см. раздел «Оценка категорий классов»). BERT, SciBERT и Covid-BERT в среднем лучше выполняются на ACL Anthology и CORD-19 при сравнении основы и других моделей на основе Transformer. Что касается набора данных ACL Anthology, то методы ранжируются одинаково для макро- и микро-. SciBERT представляет более высокие оценки с большим отрывом ото всех, за ней следуют Covid-BERT, XLNet и BERT. Менее эффективными являются RoBERTa (0,626 микро-F1) и ELECTRA (0,616 микро-F1). С точки зрения макро-среднего методы представляют одинаковые средние значения для массивов CORD-19 и ACL Anthology за исключением BERT, превышающего XLNet. Только для микро-среднего в массиве CORD-19 результат отличается, т. е. ELECTRA и RoBERTa достигают более высоких оценок F1, чем Covid-BERT и XLNet. Даже если Covid-BERT лучше мотивирован на наборе CORD-19, его эффективность содержит 0,818 микро-F1.

Оценка категорий классов

Делим оба набора данных ACL Anthology и CORD-19 на 11 категорий классов с положительными и отрицательными примерами (разделы «Предварительная обработка данных» и «Негативная выборка»). Каждый класс представляет различный раздел, в котором статья получает ссылку. Раздел указывает на то, в каких аспектах две статьи являются схожими. Эти аспекты могут также быть двусмысленными, затрудняя задачу классификации названий. Следующий раздел изучает эффективность классификации относительно различных категорий классов. Табл. 3 представляет оценку F1, полноту и точность набора SciBERT для всех 11 категорий. Дополнительно мы включаем полные результаты для единичных и многокатегорийных выборок (т.е. 2 и ≥ 3). Оставшиеся методы из табл. 2 представляют более низкие, но пропорционально схожие значения*.

Категория None имеет самую высокую с большим отрывом оценку F1 (0,942 для набора ACL Anthology и 0,980 для набора CORD-19). Категория Other показывает вторую лучшую оценку F1, которая в сценарии классификации сходства (похожий-непохожий) может быть интерпретирована как противоположный класс по отношению к категории None. Оставшиеся положительные категории показывают более низкие оценки, а также более низкий ряд (число) выборок. Поскольку мы проводим 4-кратную перекрестную проверку, сооотношение подготовленного и опытного образцов составляет 75/25. В наборе CORD-19 10 788 (категория Other) опытных образцов существуют относительно 3 777 образцов (категория Introduction), которая является наиболее общим названием раздела (табл. 1). Пока более низкое число опытных образцов необязательно коррелирует с низкой точностью. В наборе ACL Anthology категория Related Work (3 150 опытных образцов) создает более высокие оценки по сравнению с категорией Introduction (4 069 образцов) с оценкой F1 в 0,638 при наличии только 113 образцов. Рзультаты в табл. 3 отражают воздействие категорий классов на общую эффективность. Шесть категорий (набор ACL Anthology – Conclusion, Discussion, Evaluation и Methods; набор CORD-19 – Future Work и Virus) имеют оценки F1 от 0 до 0,05. Различие в числе образцов и трудности в раскрытии латентной информации с точки зрения аспектов способствуют снижению точности в некоторых категориях. Даже для экспертов области местоположение того, где одна статья цитирует другую, например, в Introduction или Experiment, не является тривиальным для прогнозирования.

Таблица 3

Результаты SciBERTотносительно наборов данных ACL Anthology и CORD-19 по категории классов, числу доступных опытных образцов (ограниченное множество), оценке F1 (со стандартным отклонением), полноте (R) и точности (P).

ACL Anthology					CORD-19				
Категория	**Опытные образцы**	**F1 (Std)**	**P**	**R**	**Категория**	**Опытные образцы**	**F1 (Std)**	**P**	**R**
Background	341	0.436 ± 0.045	0.651	0.329	Background	113	0.617 ± 0.042	0.655	0.588
Conclusion	289	0.000 ± 0.000	0.000	0.000	Conclusion	250	0.274±0.039	0.563	0.182
Discussion	283	0.000 ± 0.000	0.000	0.000	Discussion	3314	0.636 ± 0.008	0.641	0.631
Evaluation	242	0.008 ± 0.007	0.396	0.004	Future work	42	0.032 ± 0.064	0.150	0.018
Experiment	1006	0.360 ± 0.008	0.491	0.284	Introduction	3777	0.644 ± 0.004	0.669	0.620
Introduction	4069	0.527 ± 0.005	0.576	0.486	Materials	105	0.241 ± 0.038	0.552	0.157
Methods	179	0.014 ± 0.028	0.208	0.007	Methods	130	0.205 ± 0.030	0.519	0.130
Related work	3150	0.638 ± 0.012	0.660	0.617	Results	227	0.322 ± 0.021	0.558	0.227
Results	295	0.015 ± 0.011	0.475	0.008	Virus	54	0.000 ± 0.000	0.000	0.000
Other	5562	0.645 ± 0.005	0.646	0.645	Other	10788	0.876 ± 0.002	0.872	0.879
None	6068	0.942 ± 0.002	0.934	0.951	*None*	8270	0.979 ± 0.001	0.980	0.977
1 label	15652	0.721 ± 0.002	0.717	0.726	1 label	22885	0.860 ± 0.003	0.844	0.876
2 labels	1968	0.540 ± 0.003	0.738	0.425	2 labels	1632	0.656 ± 0.004	0.849	0.535
> 3 labels	585	0.492 ± 0.015	0.857	0.345	> 3 labels	295	0.590 ± 0.010	0.925	0.433

* Подробные данные по оставшимся методам доступны вмсете с подгтовленными моделями в нашем хранилище GitHub.

Таблица 4

Матирца рассеяния из выбранного множества категорий для SciBERT на наборе данных CORD-19

Основополагающая истина		**Предсказания**														
Разделы	**Выборка**	N	**B**	**C**	**D**	**I**	**O**	**R**	**C,O**	**D,I**	**D,O**	**D,R**	**I,O**	**O,R**	**D,I,O**	**D,O,R**
C,D	21	-	-	-	1	6	7	-	-	1	-	-	1	-	-	-
C,O	79	-	-	2	1	2	58	-	13	-	-	-	3	-	-	-
D,I	**459**	**1**	-	-	**163**	**146**	**17**	-	-	**103**	**7**	**2**	**9**	-	**10**	-
D,O	351	1	2	-	102	30	120	1	-	15	59	1	4	1	4	-
D,R	65	1	-	-	6	10	10	-	-	1	3	28	-	-	-	1
I,O	453	2	1	-	15	114	215	1	-	12	16	1	62	-	9	-
D,I,O	142	1	1	-	28	31	11	-	-	33	8	-	12	-	14	-
D,O,R	23	-	-	-	5	-	7	-	-	-	5	2	-	1	-	1

Примечание: (N=None, C=Conclusion, O=Other, D=Discussion, I=Introduction, R=Results). Например (**жирное выделение**), 459 опытных образцов приписаны к Discussion и *I*ntroduction (D, I), из которых 103 являются правильно класифицированными. Оставшиеся образцы в большинстве случаев классифицированы как монокатегорийные, т.е. либо Discussion (163), либо Introduction (146).

Нижние ряды в табл. 3 иллюстрируют эффект множества категорий. Значения F1 снижаются в обоих массивах по мере роста числа категорий. Это происходит из-за снижающейся полноты. Точность растет с увеличением категорий. Табл. 4 демонстрирует долю распределения многокатегорийных опытных образцов в наборе CORD-19 и соответствующих предсказаниях SciBERT (этот список лимитивирован из-за пространственных ограничений). Когда представлены две и более категорий, SciBERT часто правильно предсказывает одну из категорий, но не другие. Например, две категории из Discussion и Introduction имеют правильными только 22% опытных образцов. Пока SciBERT правильно предсказывает для оставшихся образцов одну из двух категорий, т.е. либо *Duscussion* (35%) или Introduction (31%). Мы считаем сравнимыми разультаты для других множеств категорий, таких как Discussion (D), Introduction (I) и **Others** (O).

Качественная оценка

Чтобы обосновать наши количественные полученные данные, мы качественно оцениваем предсказание от SciBERT для набора ACL Anthology. Для каждого примера в табл. 5 SciBERT предсказывает, цитирует ли источник целевую статью и в каком разделе должна встретиться ссылка. Вручную изучаем предсказания относительно их правильности.

Первым примером авторов [48] и коллектива авторов [49] является правильное предсказание. С учетом основополагающей истины этим аспектом является Other (ссылка встречается в разделе, называемом «Результаты по данным тестирования»). Мы оцениваем Introduction как верятное обоснованное предсказание, поскольку работа авторов [48] подчиняется общей, описываемой авторами [49] задаче. Поэтому можно цитировать ее во введении. Все предсказания в примере 2 - правильны. По сравнению с другими примерами мы считаем пример 2 простым случаем, так как обе статьи упоминают свою тему (т.е. сегментацию запроса, [50, 51]) в названии и в первом предложении реферата (намек на категорию «Introduction»). Оба реферата примера 2 также относятся к «взаимной информации и EM алгоритма оптимизации» как к их методам. В примере 3 ряд авторов [52, 53] не обменивается ни одной ссылкой. Следовательно, пара статьей приписывается к категории None согласно данным основополагающей истины, даже если они, как правило, связаны. Фамилии авторов [52, 53] относятся к машинному переводу с китайского языка. Пока мы не согласны с предсказанием нашей модели Experiment, так как две статьи проводят различные эксперименты, делая Experiment необоснованным предсказанием. Предсказания примера 4 правильные. Коллектив авторов [54, 1992 г.] публикуется до работы авторов [55, 2007 г.] и поэтому ссылки не существует. Тем не менее две статьи охватывают близкую тему. Таким образом, можно ожидать ссылку на авторов [54] в работе авторов [55] в разделе введение, как предсказала SciBERT. Наша модель находит это семантическое сходство, учитывая их латентную информацию по теме. Примеры 5-6 прдставляют две пары, в которрых None была правильно предсказана в соответствии с основополагающей истиной. Ряд работ из примера 6, как правило, не связан друг с другом тематически, как уже предполагают их названия. Тем не менее, авторы [56] и авторский коллектив [57] в примере 5 объединены темой разрешения многозначности. Таким образом, мы должны согласиться с предсказанием положительной категории.

Кратко, качественная оценка не противоречит нашим количественным полученным данным. SciBERT различает документы на более высоком уровне и классифицирует, какие аспекты делают их схожими. В дополнение к традиционному сходству документов предсказания на основе аспекта позволяют оценить, как две статьи относятся друг к другу на уровне семантики. Например, являются ли схожими две статьи в аспектах Introduction или Experiment, представляется ценной информацией, особенно в обзорах литературы.

ОБСУЖДЕНИЕ

В наших экспериментах SciBERT превосходит все другие методы в парной классификации документов. Мы наблюдаем, что внутри области предварительная обработка и цель NSP часто ведут к более высоким оценкам F1. Переход общих языковых моделей к определенной области, как правило, снижает эффективность в наших экспериментах. Возможным объяснением этого

является уже определенный словарь в массивах ACL Anthology или CORD-19. Ряд работ авторов [24, 39] также исследует переход обучения между областями со схожими результатами. Covid-BERT кажется исключением, так как он выдает более низкие результаты (micro-F1), чем BERT в наборее CORD-19, даже если Covid-BERT был хорошо мотивированным на набор CORD-19. Мы наблюдаем языковую модель хорошо мотивировпнную на Covid-BERT, что не гарантирует более высокую эффективность по сравнению с предварительной обработкой из случая в SciBERT. Тем не менее авторы Covid-BERT предоставляют слишком мало информации, чтобы дать собственное объяснение ее эффективности. Отдельно от предварительной обработки внутри области цель NSP имеет положительное влияние на модели. Все системы на основе BERT, использующие NSP, превосходят модели, которые исключают NSP (XLNet, RoBERTa и ELECTRA). Мы приписываем положительный эффект от NSP ее сходству с нашей задачей, поскольку обе являются следствием задач парной классификации. Табл. 2 и 3 показывают вариацию между названиями и обоими наборами данных. Большее число подготовленных опытных образцов в CORD-19 (36%) могут способствовать более высокой эффективности в сравнении с набором ACL Anthology. Несбалансированное распределение классов и различные проблемы категорий вынуждают эффективность различаться между категориями классов. Высокие оценки F1 свыше 0,9 для негативных выборок ожидаемы, поскольку категория None является неотъемлемым сходством свободным от аспекта или проблемы предсказания цитирования. Модели Transformer показаны с целью хорошего выполнения этих двух проблем [8, 20]. Помимо несбалансированного распределения подготовленных опытных образцов мы приписываем различия между положительными категориями их двусмысленности и другим проблемам, свойственным категориям классов. Авторы часто расходятся по вопросу именования их разделов (например, Results, Evaluation), таким образом усиливается проблема наименования разных аспектов статьи. Это также способствует высокому числу выборок категории Other. Некоторые разделы, также ориентированные на контент, более уникальны, чем другие. Раздел Introduction, как правило, содержит контент, отличающийся от раздела Results. Различие в содержании позволяет некоторым разделам и соответствующим категориям классов легче, чем другим, различаться и предсказывать. Предполагаем слабую эффективность для Future Work из-за нехватки или отсутствия информации в названиях или аннотациях.

Нашей основной исследовательской целью в этой статье является изучение методов, способных объединять аспект информации в традиционной классификации сходства-различия. В связи с этим мы считаем результаты перспективными. В частности, оценка micro-F1 0,86 из SciBERT для набора CORD-19 является вдохновляющей. Наша количественная оценка указывает на то, что предсказания SciBERT могут правильно идентифицировать схожие аспекты двух научных статей. В целях подтверждения, если наше первое показание обобщается, то требуется проведение большого качественного исследования. Более того, мы наблюдаем, что категории классов с небольшими подготовленными данными действуют слабо. Например, Conclusion и Discussion имеют нулевую оценку F1 для набора ACL Anthology, тогда как для большего набора данных CORD-19 Duscussion выдает 0,636 F1. Мы ожидаем, что большие подготовленные данные приведут к более правильным прогнозам.

Таблица 5

Примеры категорий пар научных статей (источник и цель), определнных в соответствии со ссылками и предсказанных с участием SciBERT

	Статья-источник	Статья-цель	Ссылка	Предсказание
1	UKP: Computing Semantic Textual Similarity by Combining Multiple Content Similarity Measures (Bar et al., 2012)	SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity (Agirre et al., 2012)	Other	Introduction x
2	Query segmentation based on eigenspace similarity (Zhang et al., 2009)	Unsupervised query segmentation using generative language models and wikipedia (Tan and Peng, 2008)	Introduction, Experiment	Introduction ✓, Experiment ✓
3	Transition-Based Parsing of the Chinese Treebank using a Global Discriminative Model (Zhang and Clark, 2009)	Enhancing Statistical Machine Translation with Character Alignment (Xi et al., 2012)	None	Experiment x
4	Experiments in evaluating interactive spoken language systems (Polifroni et al., 1992)	Evaluating information presentation strategies for spoken recommendations (Winterboer and Moore, 2007)	None	Introduction x, Other x
5	Similarity-based Word Sense Disambiguation (Karov and Edelman, 1998)	Targeted disambiguation of ad-hoc, homogeneous sets of named entities (Wang etal., 2012)	None	None ✓
6	SciSumm: A Multi-Document Summarization System for Scientific Articles (Agarwal et al., 2011)	Improving question-answering with linking dialogues (Gandhe et al., 2006)	None	None ✓

Примечание: Основанные на ограниченном множестве правильные предсказания отмечены ✓, необоснованные – x.

ЗАКЛЮЧЕНИЕ

В этой статье мы применяем к научным статьям парную многокатегорийную, многоклассовую классификацию документов, чтобы вычислить оценку сходства документа на основе аспекта. Обрабатываем названия разделов как аспекты цитирования статей и названий, соответственно встречающихся в этих разделах. Изучаемые модели обучены для предсказания цитирований и соответствующих категорий, основанных на названии статьи и ее реферате. Мы оцениваем модели Transformer BERT, Covid-BERT, SCiBERT, ELECTRA, RoBERTa и XLNet и основу LSTM относительно двух научных наборов, т.е. ACL Anthology и CORD-19. В целом SciBERT в наших экспериментах работает лучше. Несмотря на сложность задачи, SciBERT предсказала сходство документов на основе аспекта с оценкой F1 свыше 0,83. Эффективность SciBERT стимулирует дальнейшее исследование в этом направлении. Кажется обоснованным включить задачу сходства документов на основе аспекта в качестве новой цели предварительной обработки в архитектуре Transformer. Эта новая цель могла бы быть интегрирована похожим способом как двойственная цель предсказания цитирования, предложенная авторами [8]. В качестве дальнейшей работы планируем интегрировать сходство документов на основе аспекта в рекомендательную систему. Таким образом, стимулируя большее исследование пользователей, чтобы подтвердить наши первые выводы относительно того, что сходство документов на основе аспекта действительно помогает пользователям находить более релевантные рекомендации. Однако наш расширенный эмпирический анализ уже демонстрирует, что модели Transformer хорошо подходят для правильного вычисления сходства документов на основе аспекта на примере научных статей.

Благодарность. Хотим выразить благодарность всем рецензентам и Кристофу Альту за их рекомендации и ценную обратную связь. Представленное в этой статье исследование профинансировано Немецким федеральным министерством образования и исследований через проект QURATOR (Unternehmen Region, Wachstumskern, no. 03WKDAIA).

ЛИТЕРАТУРА

1. *Beel J., Gipp B., Langer S., Breitinger C.* Research-paper recommender systems: A literature survey//International Journal on Digital Libraries. — 2016. — Vol. 17, No. 4. — P. 305–338.

2. *Goodman N.* Seven strictures on similarity. Problems and projects. — 1972.

3. *Bar D., Zesch T., Gurevych I.* A Reflective View on Text Similarity//International Conference Recent Advances in Natural Language Processing (RANLP), pp. 515–520. — 2011.

4. *Huang T.-H. K., Huang C.-Y., Ding C. -Y. C., Yen-Chia Hsu Y.-C., Giles C L.* CODA-19: Reliably annotating research aspects on 10,000+ CORD-19 abstracts using a non-expert crowd. — [arXiv:2005.02367.] — 2020.

5. *Chan J., Chang J. C., Hope T., Shahaf D., Kittur A.* SOLVENT: A mixed initiative system for finding analogies between research papers// Proceedings of the ACM on Human-Computer Interaction, 2(CSCW):1–21, nov.— 2011.

6. *Ostendorff M., Ruas T., Schubotz M., Rehm G., Gipp B.* Pairwise multi-class document classification for semantic relations between Wikipedia articles//Proceedings of the 2020 ACM/IEEE Joint Conference on Digital Libraries (JCDL'20). — 2020.

7. *Jiang J.-Y., Zhang M., Li C., Bendersky M., Golbandi N., Najork M.* Semantic text matching for long-form documents // The World Wide Web Conference on - WWW '19, pages 795–806, New York, New York, USA. — ACM Press, 2019.

8. *Cohan A., Feldman S., Beltagy I., Downey D., Weld D. S.* SPECTER: Document-level representation learning using citation-informed Transformers// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL). — 2020.

9. *Steven Bird S., Dale R., Dorr B. J., Gibson B., Joseph M. T., Kan M. Y., Lee D., Powley B., Radev D. R., Tan Y. F.* The ACL Anthology reference corpus: A reference dataset for bibliographic research in Computational Linguistics// Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008, pp. 1755–1759.— 2008.

10. *Wang L. L., Lo K., Chandrasekhar Y., Reas R., Yang J., Eide D., Funk K., Kinney R., Liu Z., Merrill W., Mooney P., Murdick D., Rishi D., Sheehan J., Shen Z., Stilson B.,Wade A. D., Wang K., Wilhelm C., Xie B., Raymond D., Weld D. S., Etzioni O., Kohlmeier S.* CORD-19: The Covid-19 open research dataset. — 2020. — [arXiv:2004.10706.]

11 *Kobayashi Y., Shimbo M., Matsumoto Y.* Citation recommendation using distributed representation of discourse facets in scientific articles// Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries, pp. 243–251, New York, NY, USA, may.— ACM, 2018.

12. *Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L., Polosukhin I.* Attention is all you need// Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 6000–6010, Jun. — 2017.

14. *Liu Y., Ott M., Goyal N., Du J., Joshi M., Chen D., Levy O., Lewis M., Zettlemoyer L., Stoyanov V.* RoBERTa: A robustly optimized BERT pretraining approach. — 2019. — [arXiv:1907.11692.]

15. *Yang Z., Dai Z., Yang Y., Carbonell J., Salakhutdinov R., Le Q.V.* XLNet: Generalized autoregressive pretraining for language understanding// Advances in Neural Informati on Processing Systems 32, pp. 5754–5764. — 2019.

16. *Clark K., Luong M.-T., Le Q. V., Manning C. D.* ELECTRA: Pre-training text encoders as discriminators rather than generators//International Conference on Learning Representations, pp.1–18. — 2020.

17. *Bowman S. R., Angeli G., Potts C., Manning C. D.* A large annotated corpus for learning natural language inference// Conference on Empirical Methods in Natural Language Processing (EMNLP 2015), pp. 632–642. — 2015.

18. *Williams A., Nangia N., Bowman S.* A broad-coverage challenge corpus for sentence understanding through inference. — [arXiv:1704.05426, pages 1112–1122]. — 2018.

19. *Daniel Cer D., Diab M., Agirre E., Lopez-Gazpio I., Specia L.* SemEval-2017 Task 1: Semantic textual similarity multilingual and crosslingual focused evaluation// Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), volume 371, pp. 1–14, Stroudsburg, PA, USA.— Association for Computational Linguistics, 2017.

20. *Reimers N., Gurevych I.* Sentence-BERT: Sentence embeddings using Siamese BERT-networks//The 2019 Con-

ference on Empirical Methods in Natural Language Processing (EMNLP 2019). — 2019.

21. *Bromley J., Bentz J. W., L. Bottou L., Guyon I., Lecun Y., Moore C., Sackinger E., Shah R.* Signature verification using a Siamese time delay neural network//International Journal of Pattern Recognition and Artificial Intelligence. — 1993. — Vol. 7, No.4.

22. *Adhikari A., Ram A., Tang R., Lin J., Cheriton D. R.* DocBERT: BERT for Document Classification. — 2019. — [arXiv:1904.08398v1].

23. *Ostendorff M., Bourgonje P., Berger M., Moreno-Schneider J., Rehm G., Gipp B.* Enriching BERT with knowledge graph embeddings for document classification// Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019), pp. 305–312, Erlangen, Germany. — German Society for Computational Linguistics & Language Technology, 2019.

24. *Beltagy I., Lo K., Cohan A.* SciBERT: A pretrained language model for scientific text//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3613–3618, Stroudsburg, PA, USA. — Association for Computational Linguistics, 2019.

25. *Hassan H. A. M., Giuseppe Sansonetti G., Gasparetti F., Micarelli A., Beel J.* BERT, ELMo, USE and InferSent sentence encoders: The panacea for research-paper recommendation?// CEUR Workshop Proceedings, volume 2431, pp. 6–10. — 2019.

26. *Kanakia A., Shen Z., Eide D., Wang K.* A scalable hybrid research paper recommender system for Microsoft Academic// The World Wide Web Conference on - WWW '19, pp. 2893–2899, New York, New York, USA. — ACM Press, 2019.

27. *Collins A., Beel J.* Document embeddings vs. keyphrases vs. terms: An online evaluation in digital library recommender systems// ACM/IEEE Joint Conference on Digital Libraries (JCDL), pp. 130–133. — 2019.

28. *Nanni F., Ponzetto S. P., Dietz L.* Entity-aspect linking: Providing fine-grained semantics of entities in context// Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries, pp. 49–58, New York, NY, USA, may. — ACM, 2018.

29. *Lo K., Wang L. L., Neumann M., Kinney R., Weld D. S.* S2ORC: The Semantic Scholar open research corpus. — 2019. —[arXiv:1911.02782].

30. *Ley M.* DBLP: Some lessons learned// Proceedings of the VLDB Endowment. — 2009. — Vol. 2, No. 2. — P. 1493–1500, aug.

31. *Mikolov T., Chen K., Corrado G., Dean J.* Distributed representations of words and phrases and their compositionality// Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, pp. 3111–3119. — 2013.

32. *Pennington J., Socher R., Manning C.* Glove: Global vectors for word representation//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543. — 2014.

33. *Le Q. V., Mikolov T.* Distributed representations of sentences and documents// Proceedings of the 31st International Conference on Machine Learning. — 2014. — Vol. 32. — P. 1188–1196.

34. *Hochreiter S., Schmidhuber J.* Long short-term memory//Neural Computation. — 1997. — Vol. 9, No. 8. — P. 1735–1780, nov.

35. *Honnibal M., Montani I.* SpaCy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. — 2020. — [в печати].

36. *Bojanowski P., Grave E., Joulin A., Mikolov T.* Enriching word vectors with subword information// Transactions of the Association for Computational Linguistics. — 2017. — Vol. 5. — P. 135–146.

37. *Zhu Y., Kiros R., Zemel R., Salakhutdinov R., Urtasun R., Torralba A., Fidler S.* Aligning books and movies: Towards story-like visual explanations by watching movies and reading books//Proceedings of the IEEE International Conference on Computer Vision, 2015 Inter:19–27.— 2015.

38. *Chan B.* CORD-19 BERT Model. — 2020. — https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge/discussion/138250.

39. *Lee J., Yoon W., Kim S., Kim D., Kim S., So C. H., Kang J.* BioBERT: A pre-trained biomedical language representation model for biomedical text mining// Bioinformatics. — 2018. — P.1–8, sep.

40. *Nagel S.* 2016. Common Crawl News. — 2016. — http://commoncrawl.org/2016/10/news-dataset-available/.

41. *Gokaslan A., Cohen V.* Openwebtext corpus. — 2019. — https://skylion007.github.io/OpenWebTextCorpus/.

42. *Trinh T. H., Le Q. V.* A simple method for commonsense reasoning. — 2018. — [arXiv:1806.02847].

43. *Parker R., Graff D., Kong J., Chen K., Maeda K.* English gigaword fifth edition. — 2011. — https://catalog.ldc.upenn.edu/LDC2011T07.

44. *Callan J., Hoy M., Changkuk Yoo C., Zhao L.* Clueweb09 data set. — 2009. — https://lemurproject.org/clueweb09/.

45. *Elbaz G.* Common Crawl. — 2007. — http://commoncrawl.org

46. *Reimers N., Gurevych I.* Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 338–348, Stroudsburg, PA, USA. — Association for Computational Linguistics, 2017.

47. *Wolf T., Debut L., Sanh V., Chaumond J., Delangue C., Moi A., Cistac P., Rault T., Louf R., Funtowicz M., Brew J.* HuggingFace's Transformers: State-of-the-art natural language processing. — 2019. — [arXiv:1910.03771, oct.].

48. *Bar D., Biemann C., Gurevych I., Zesch T.* UKP: Computing semantic textual simi- larity by combining multiple content similarity measures// 1st Joint Conference on Lexical and Computational Semantics (SEM 2012), Vol. 2, pp. 435–440. — 2012.

49. *Agirre E., Cer D., Diab M., Gonzalez-Agirre A.* SemEval-2012 task 6: A pilot on semantic textual similarity - Google search// Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012), pp. 385–393. — 2012.

50. *Zhang C., Sun N., Hu X., Huang T., Chua T.S.* Query segmentation based on eigenspace similarity// ACL-IJCNLP 2009 - Joint Conf. of the 47th Annual Meeting of the Association for Computational Linguistics and 4th Int. Joint Conf. on Natural Language Processing of the AFNLP, Proceedings of the Conf., pp. 185–188.— 2009.

51. *Tan B., Peng F.* Unsupervised query segmentation using generative language models and Wikipedia// Proceed-

ing of the 17th international conference on World Wide Web - WWW '08, p. 347, New York, New York, USA. — ACM Press, 2008.

52. *Zhang Y., Clark S.* Transition-based parsing of the Chinese treebank using a global discriminative model// Proceedings of the 11th International Conference on Parsing Technologies - IWPT '09, p. 162, Morristown, NJ, USA. — Association for Computational Linguistics, 2009.

53. *Xi N., Tang C., Dai X., Huang S., Chen J.* Enhancing statistical machine translation with character alignment// 50th Annual Meeting of the Association for Computational Linguistics, ACL 2012 - Proceedings of the Conference, 2(July). — 2012. — P. 285–290.

54. *Polifroni J., Hirschman L., Seneff S., Zue V.* Experiments in evaluating interactive spoken language systems // Proceedings of the workshop on Speech and Natural Language - HLT '91, p. 28, Morristown, NJ, USA. — Association for Computational Linguistics, 1992.

55. *Winterboer A., Moore J. D.* Evaluating information presentation strategies for spoken recommendations// RecSys'07: Proceedings of the 2007 ACM Conference on Recommender Systems, pages 157–160. — 2007.

56. *Agarwal N., Reddy R. S., Gvr K., Rose C. P.* SciSumm: A multi-document ´summarization system for scientific articles// 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of Student Session (ACL HLT 2011), pp. 115–120. — 2011.

57. *Gandhe S., Gordon A. S., Traum D.* Improving question-answering with linking dialogues// International Conference on Intelligent User Interfaces, Proceedings IUI.— 2006. — P. 369–371.— 2006.

58. *Karov E., Edelman S.* Similarity-based word sense disambiguation// Computational Linguistics. — 1998. — Vol. 24, No. 1.

59. *Wang C., Chakrabarti K., Cheng T., Chaudhuri S.* Targeted disambiguation of ad-hoc, homogeneous sets of named entities// WWW'12 - Proceedings of the 21st Annual Conference on World Wide Web, pp. 719–728. — 2012.