Автоматический поиск фрагментов, содержащих биографическую информацию, в тексте на естественном языке[1]

А.В. Глазкова <a.v.glazkova@utmn.ru>
Тюменский государственный университет,
625003, Россия, г. Тюмень, ул. Володарского, д.6

Аннотация. Поиск и классификация текстовых документов применяются во многих практических приложениях и являются одними из ключевых задач информационного поиска. Методы поиска и классификации текстов находят применение в поисковых системах, электронных библиотеках и каталогах, системах сбора и обработки информации, платформах для онлайн-обучения и многих других. Существует большое количество частных применений указанных методов, однако каждая подобная практическая задача отличается, как правило, слабой формализуемостью, узкой предметностью и, следовательно, требует индивидуального изучения и собственного подхода к решению. В данной работе рассматривается задача автоматического поиска и типизации текстовых фрагментов, содержащих биографическую информацию. Ключевой проблемой при решении указанной задачи является проведение мультиклассовой классификации текстовых фрагментов в зависимости от наличия и типа содержащейся в них биографической информации. Проведя обзор научной литературы по рассматриваемому вопросу, авторы сделали вывод о перспективности и широте применения нейросетевых методов для решения подобных задач. Исходя из данного вывода, в работе проведено сравнение различных архитектур нейросетевых моделей, а также основных способов представления текстов (Bag-of-Words, Bag-of-Ngrams, TF-IDF, Word2Vec) на предварительно собранном и размеченном корпусе биографических текстов. В статье описываются этапы подготовки обучающего множества текстовых фрагментов для обучения моделей, способы представления текстов и методы классификации, выбранные для решения задачи. Также приводятся результаты мультиклассовой классификации текстовых фрагментов и показаны примеры автоматического поиска фрагментов, содержащих биографическую информацию, в текстах, не участвовавших в процессе обучения моделей.

Ключевые слова: классификация текстов; обработка естественного языка; векторные представления слов; нейронные сети; биографический текст.

[1] Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 18-37-00272 «Автоматизированное извлечение биографических фактов из текстов на естественном языке».

DOI: 10.15514/ISPRAS-2018-30(6)-12

Для цитирования: Глазкова А.В. Автоматический поиск фрагментов, содержащих биографическую информацию, в тексте на естественном языке. Труды ИСП РАН, том 30, вып. 6, 2018 г., стр. 221-236. DOI: 10.15514/ISPRAS-2018-30(6)-12

1. Введение

Постоянное увеличение количества и объема мировых информационных ресурсов требует разработки и совершенствования технологий информационного поиска. Решение задач информационного поиска связано, как правило, с обработкой естественного языка и направлено на анализ, обобщение и упорядочение неструктурированной документальной (в том числе текстовой) информации.

Одной из прикладных задач информационного поиска является задача автоматического поиска биографической информации в текстах, написанных на естественном языке. Извлечение биографической информации выполняется не только при построении текста биографической справки в поисковых системах, но и при проведении биографических исследований, которые подразумевают работу с фактами, касающимися жизни человека. Поиск биографических фактов в тексте на естественном языке имеет ряд особенностей. Так, факты сами по себе могут касаться различных сфер жизни – социальной, политической, экономической или личной. Тексты, содержащие биографические факты, включают в себя как строго документальные (автобиографии, резюме), так и нестрого документальные (воспоминания, очерки, хроники) [1]. Во втором случае какая-то часть биографической информации может встречаться в тексте в неявном виде, среди другой информации, не относящейся к биографической. Эти особенности вынуждают исследователя просматривать большое количество электронных документов в поисках значимых для его работы фактов.

В данной статье рассматривается задача автоматического поиска фрагментов, содержащих биографическую информацию, в тексте, написанном на естественном языке. Рассматриваемая задача может быть представлена как задача тематической классификации текстов. При этом текстовые фрагменты, подлежащие классификации, оцениваются с точки зрения того, содержат ли они биографическую информацию, и если содержат, то к какому тематическому классу эта информация относится.

Вопросы извлечения фактов различного характера из текста на естественном языке и их классификации довольно широко освещаются в научной литературе. При этом особенностью задачи извлечения фактов является ее слабая формализуемость и узкая предметность [2]. Существующие подходы к извлечению фактов, хотя они и охватывают довольно широкий спектр прикладных задач, сложно сравнить по качеству и результативности с подходом, представленным в данной работе. Это связано как со спецификой каждой отдельной задачи, так и с различием текстовых коллекций,

использованных для оценки результатов, и особенностями обрабатываемых естественных языков. Так, работа И.М. Адамовича и О.И. Волкова [3] посвящена извлечению биографических фактов из исторических документов. Авторы описывают технологию, которая представляет факт как древовидную структуру. Корнем такого дерева является факт (например, «рождение»), связанные с данным фактом сущности сохраняются в листьях. Также извлечение биографических фактов обсуждается в статьях [4-5]. В статье [6] авторы решают задачу классификации отношений между словами текста (по сути выделение фактов и «не-фактов») с использованием сверточной нейронной сети, которая выполняет классификацию путем ранжирования (CR-CNN). В работе P. Meerkampd и Z. Zhou [7] представлена архитектура системы извлечения информации из текста, которая сочетает в себе возможности синтаксического анализа (парсинга) текста и нейронной сети. В работе Y. Homma и др. [8] описана иерархическая нейронная сеть для классификации предложений для извлечения фактов о продукте из товарных документов. Сеть классифицирует каждое предложение в документе по классам атрибутов и условий на основе последовательностей слов и предложений в документе.

Стоит отметить, что в большинстве представленных работ используются методы машинного обучения. Для решения различных задач обработки естественного языка в работах российских и зарубежных исследователей [9-14] неоднократно применялись нейросетевые технологии, основанные на рекуррентности и долгой краткосрочной памяти.

2. Представление текстов

2.1 Текстовая коллекция

В качестве текстового фрагмента в рамках данной работы рассматривается предложение. Данная языковая единица представляет собой грамматически организованное соединение слов (или слово), обладающее смысловой законченностью [15]. При этом возможны ситуации, когда несколько предложений текста в смысловом плане отражают один и тот же биографический факт. Вопросы объединения взаимодополняющих и удаления дублирующихся фактов являются дальнейшими задачами данного исследования.

В работе использовались тексты, находящиеся в открытом доступе в онлайн-энциклопедии «Википедия» [16]. Ранее [17] авторами был описан корпус биографических текстов, собранный на основе биографических текстов, представленных в «Википедии», и размеченный в полуавтоматическом режиме. Каждому предложению собранной коллекции биографических текстов сопоставлен один из классов в соответствии со следующей таксономией:

- содержит биографические факты:
  - информация о родительской семье;

- личные события;
- место жительства или пребывания;
- место работы или службы;
- национальность;
- образование;
- профессиональные события (встречи, награждения и т.д.);
- род занятий;
- рождение;
- семья (женитьба, замужество, дети);
- смерть;
- прочие биографические факты;

- не содержит биографические факты.

В корпус включены 200 биографий персоналий, живших или живущих в 20-21 веках. Основные количественные характеристики корпуса представлены в табл. 1.

Табл. 1. Количественные характеристики корпуса
Table 1. Quantitative characteristics of the corpus

Характеристика	**Значение**
Среднее количество слов в текстах	225
Среднее количество предложений в текстах	19
Доля типов биографических фактов (%)	
Информация о родительской семье	5,23
Личные события	4,17
Место жительства	3,99
Место работы	2,15
Национальность	1,51
Образование	6,46
Профессиональные события	13,80
Род занятий	14,67
Рождение	4,68
Семья	34,94
Смерть	4,49
Прочие биографические факты	3,90

Корпус биографических текстов в формате .xml доступен по ссылке [18]. Для формирования коллекции предложений, не содержащих биографические

факты, была использована выборка случайных небиографических статей из «Википедии».

Итоговая коллекция текстовых фрагментов, использованная для обучения моделей, включила в себя предложения, относящиеся к 11 классам: «Информация о родительской семье», «Личные события», «Место жительства», «Место работы», «Образование», «Профессиональные события», «Род занятий», «Рождение», «Семья», «Смерть» и «Фрагменты, не содержащие биографическую информацию». Класс «Национальность» был объединен с классом «Информация о родительской семье» ввиду семантического сходства входящих в них обучающих примеров. Класс «Прочие биографические факты» был исключен, поскольку входящие в него фрагменты содержат информацию, косвенно относящуюся к биографической, однако не принадлежащую ни к одному из конкретных классов.

Для выравнивания количества обучающих элементов в различных классах был проведен простой оверсэмплинг – дублирование элементов миноритарных классов. Итоговое количество предложений, содержащих биографические факты, составило 6773.

Предобработка текстовой коллекции включала в себя следующие этапы:

- удаление знаков препинания и специальных символов;
- перевод символов в нижний регистр;
- удаление стоп-слов;
- лемматизация (использовались средства библиотеки pymorphy2 [19]).

2.2 Признаковое пространство

В данной работе были рассмотрены четыре способа представления текстов:

- модель Bag-of-Words;
- Bag-of-Words + TF-IDF;
- Bag-of-Ngrams + TF-IDF;
- Word2Vec.

В ходе построения модели Bag-of-Words текстовая коллекция была представлена в виде матрицы, количество строк которой равно количеству документов, а количество столбцов – количеству слов в коллекции (за исключением списка стоп-слов). На пересечении строки и столбца хранится количество вхождений слова в текст конкретного документа.

Для определения наиболее характерных для классов слов использовалась мера TF-IDF. Списки слов, имеющих наибольшие значения TF-IDF для каждого класса, представлены в таблице 2. В целях удобства отображения в каждом классе показаны по 5 наиболее значимых слов и соответствующие им значения TF-IDF.

Табл. 2. Наиболее значимые слова для классов текстовой коллекции
Table 2. The most important words for text collection classes

Название класса	**Наиболее значимые слова**		**Название класса**	**Наиболее значимые слова**	
	Слово	**TF-IDF**		**Слово**	**TF-IDF**
Личные события	познакомиться	0,13	Профессио-нальные события	быть	0,19
	религия	0,08		наградить	0,16
	обвинение	0,08		работа	0,15
	уголовный	0,08		орден	0,15
	приговор	0,07		звание	0,1
Место жительства	жить	0,41	Род занятий	заместитель	0,31
	переехать	0,27		назначить	0,26
	город	0,13		начальник	0,2
	провести	0,12		должность	0,19
	вернуться	0,12		работать	0,15
Место работы	общество	0,18	Рождение	родиться	0,86
	член-корреспондент	0,14		семья	0,26
	избрать	0,14		город	0,13
	состоять	0,08		область	0,12
	являться	0,07		губерния	0,08
Образование	окончить	0,63	Семья	супруг	0,12
	факультет	0,22		замужем	0,12
	защитить	0,2		младший	0,08
	институт	0,19		брак	0,06
	диссертация	0,17		совместный	0,06
Происхождение (объединение классов «Информация о родительской семье» и «Национальность»)	мать	0,46	Смерть	умереть	0,65
	семья	0,32		кладбище	0,41
	отец	0,26		похоронить	0,35
	родиться	0,17		скончаться	0,33
	принадлежать	0,11		расстрелять	0,08

При представлении текстов в виде модели Bag-of-Words с использованием TF-IDF на пересечении строки и столбца находится значение значимости слова в данном документе, рассчитанное при помощи TF-IDF.

Представление текста в виде Bag-of-Ngrams + TF-IDF аналогично представлению Bag-of-Words + TF-IDF, однако для расчета TF-IDF вместо слов используются N-граммы символов. В данной работе лучший результат для данного способа представления текстов был получен при N=4.

Word2Vec [20] в настоящее время служит одним из наиболее популярных и эффективных способов представления слов в векторном виде, пригодном для машинного обучения (word embeddings). Этот способ построен на частоте совстречаемости слов в пределах одного контекста.

В данной работе использовались векторные представления, полученные по алгоритму Word2Vec на основе текстов русскоязычной «Википедии» за 2018 год с использованием алгоритма обучения Skip-gram. В ходе экспериментов был выбран размер результирующего контекстного вектора, равный 300.

3. Методы

Для классификации фрагментов текстов были выбраны следующие типы нейронных сетей:

- сеть прямого распространения (feedforward network, FNN);
- сеть долгой краткосрочной памяти (long short-term memory, LSTM);
- двунаправленная сеть долгой краткосрочной памяти (bidirectional long short-term memory, BLSTM).

Сети прямого распространения применялись в качестве модели для классификации текстов с использованием представлений документов в виде моделей Bag-of-Words, Bag-of-Words + TF-IDF и Bag-of-Ngrams + TF-IDF, а также в качестве фрагментов каскадных моделей. Для классификации текстов на основании векторных представлений слов использовались сети долгой краткосрочной памяти.

В отличие от классических архитектур нейронных сетей, в рекуррентном слое сети долгой краткосрочной памяти предусмотрен механизм хранения долгосрочных зависимостей, позволяющий избежать проблемы затухания градиента [21]. Архитектура ячейки сети долгой краткосрочной памяти представлена на рис. 1 [22].

Пусть xt и yt – входной и выходной сигналы соответственно в момент времени t, а ct и mt – состояние ячейки и выхода в момент t. Преобразование входного сигнала в выходной при этом происходит следующим образом:

$$i_t = \sigma\left(W_{ix}x_t + W_{im}m_{t-1} + W_{ic}c_{t-1} + b_i\right),$$
$$f_t = \sigma\left(W_{fx}x_t + W_{fm}m_{t-1} + W_{fc}c_{t-1} + b_f\right),$$
$$o_t = \sigma\left(W_{ox}x_t + W_{om}m_{t-1} + W_{oc}c_{t-1} + b_o\right),$$

$$m_t = o_t \odot h(c_t),$$
$$y_t = \phi(W_{ym} m_t + b_y),$$
$$c_t = f_t \odot c_{t-1} + i_t \odot g(W_{cx} x_t + W_{cm} m_{t-1} + b_c),$$

где W_{cx}, W_{ix}, W_{fx}, W_{ox} – веса входов, W_{cm}, W_{im}, W_{fm}, W_{om} – веса состояний ячеек, b_o, b_i, b_f – смещения, W_{ic}, W_{fc}, W_{oc} – веса связей между ячейками и слоем выходного фильтра. W_{ym} и b_y – вес и смещение для выхода. σ, g, h представляют собой некоторые нелинейные функции.

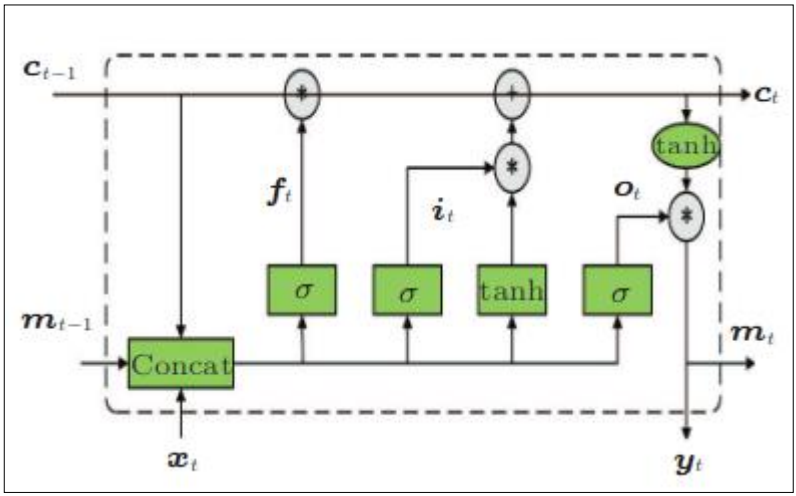

Рис. 1. Структура ячейки LSTM
Fig. 1. The structure of LSTM cell

Двунаправленная сеть долгой краткосрочной памяти комбинирует классическую LSTM-сеть, которая обрабатывает последовательность данных от её начала до конца, с другой LSTM-сетью, которая рассматривает последовательность в обратном порядке.

Анализ работ по близкой тематике показал, что указанные нейросетевые архитектуры широко применяются в задачах обработки естественного языка (см. разд. 1). Эксперименты, проведенные на используемом в работе текстовом корпусе, подтвердили эффективность использования LSTM- и BLSTM-сетей при проведении классификации на основании векторных представлений слов в сравнении с сетями прямого распространения и классическими рекуррентными сетями. В табл. 3 представлен результат сравнения нейросетевых моделей для случая бинарной классификации текстов (классификация в зависимости от того, содержит ли предложение биографическую информацию).

Табл. 3. Сравнение архитектур нейронных сетей на примере бинарой классификации
Table 3. The comparison of neural architectures for the task of binary classification

Архитектура сети	Accuracy (%)	Precision (%)	Recall (%)	F-мера(%)
FNN	86	86,01	93,89	89,78
RNN	89,5	89,86	94,66	92,19
LSTM	91,5	92,18	95,66	93,89
BLSTM	91,5	91,99	95,42	93,63

При проведении классификации фрагментов текстов в данной работе использовались нейросетевые модели, реализованные при помощи средств библиотеки Keras [23]. В качестве функций активации для рекуррентных сетей были выбраны гиперболический тангенс на внутренних слоях и функция Softmax для выходного слоя. Для сетей прямого распространения – логистическая функция на всех слоях. Размер обрабатываемых фрагментов данных (batch size) – 8. Использованный оптимизационный алгоритм – adaptive moment estimation (the Adam optimization). При обучении сетей проводилась дропаут-регуляризация с вероятностью 0,5. Количество нейронов в рекуррентных слоях варьировалось от 16 до 128 при глубине сети в 1-2 скрытых слоя. В итоге для каждой архитектуры были обучены несколько сетей, из которых по результатам на обучающей выборке была выбрана модель, допущенная до экзамена на тестовой выборке. Фрагменты исходного кода, использованные для построения моделей, можно получить по ссылке [24].

4. Результаты

В табл. 4 представлены результаты мультиклассовой классификации по 11 результирующим классам («Фрагменты, не содержащие биографическую информацию» и 10 классов фрагментов, содержащих биографические сведения).

Оценка качества классификации проводилась с использованием следующих метрик: точности (accuracy, то есть количества совпадений фактического и прогнозируемого классов, %) и F-меры (F-score, которая в случае мультиклассовой классификации определялась как средняя величина значений F-меры, рассчитанных для каждого класса по показателям точности (precision) и полноты (recall)).

Расчет точности классификации:

$$Accuracy = T / N ,$$

где T – количество фрагментов, по которым классификатор принял верное решение, N – общее количество документов.

Расчет F-меры:

$$Precision_n = TP / (TP + FP),$$

$$Recall_n = TP / (TP + FN),$$

$$F\_score_n = 2 * Precision * Recall / (Precision + Recall),$$

$$F\_score = \frac{1}{N} \sum_{n=1}^{N} F\_score_n ,$$

где TP – истинно-положительное решение, FP – ложно-положительное решение, FN – ложно-отрицательное решение, n –номер конкретного класса, N – количество классов.

Табл. 4. Результаты мультиклассовой классификации
Table 4. The results of multi-class classification

Способ представления текстов	**Архитектура сети**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F-мера (%)**
Bag-of-Words	FNN	80,2	86,64	85,17	85,9
Bag-of-Words + TF-IDF	FNN	84,16	86,87	88,27	87,57
Bag-of-Ngrams + TF-IDF	FNN	82,18	86,44	86,6	86,52
Word2Vec	LSTM	89,11	91,46	90,79	91,13
Word2Vec	BLSTM	90,1	92,49	91,8	92,15

Наилучшие результаты при проведении классификации по всем классам были достигнуты с использованием двунаправленной сети долгой краткосрочной памяти и представления текстов при помощи Word2Vec.

Результаты модели, показавшей минимальные ошибки на тестовой выборке (далее – Модель А), были сравнены с результатами двух каскадных архитектур нейросетевых моделей (рис. 2).

Во втором случае (Модель Б) на первом этапе проводится бинарная классификация фрагментов текстов. Далее предложения, классифицированные как содержащие биографическую информацию, поступают на вход модели для мультиклассовой классификации. В итоге для каждого фрагмента сначала определяется, содержит ли он биографическую информацию, и если содержит, то определяется тип этой информации.

В третьем случае (Модель В) результирующий вектор сети для мультиклассовой классификации подается на вход сети прямого распространения одновременно с представлением текста в виде модели Bag-of-Words + TD-IDF. Сравнение результатов трех моделей для мультиклассовой классификации представлено в табл. 5.

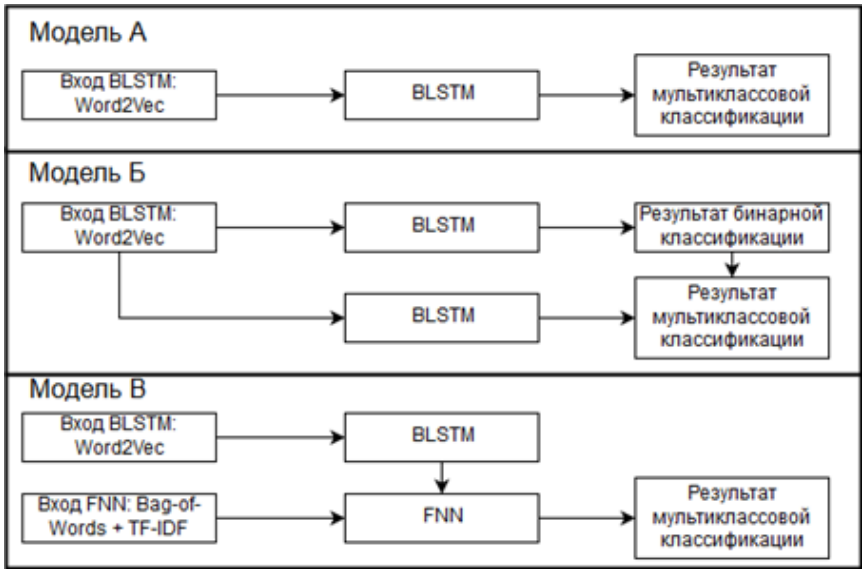

Рис. 2. Модели мультиклассовой классификации
Fig. 2. The models for multi-class classification

Табл. 5. Сравнение моделей для мультиклассовой классификации
Table 5. The comparison of multi-class classification models

Модель	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F-мера `(%)**
А	90,1	92,49	91,8	92,15
Б	93,07	95,82	93,35	94,57
В	94,06	96,37	94,36	95,36

В табл. 6 приводятся примеры автоматической классификации предложений при помощи модели, показавшей лучшие результаты на тестовой выборке. Первый текст представляет собой биографическую статью из онлайн-энциклопедии «Википедия», не входящую в корпус. Второй текст является новостью, размещенной на портале «газета.ru» [25].

Табл. 6. Примеры автоматической классификации предложений
Table 6. The examples of automatic sentences classification

Предложение	**Тип предложения (определен автоматически)**
Пример 1. «Дьяконов-Дьяченков, Георгий Иванович» (источник – «Википедия», 2018)	

Георгий Иванович Дьяконов родился 17 марта 1924 году в Москве в театральной семье.	Рождение
Отец Иван Дьяконов, родом из Оренбургской области, был артистом оперетты, который играл во многих театрах включая Москву, где у него и родился сын.	Информация о родительской семье
В 1941 году, приписав себе год, добровольцем ушёл на фронт, воевал зенитчиком в 205-м зенитно-артиллерийском полку 73-й зенитной дивизии РГК, был тяжело ранен, отличился в боях.	Род занятий
Карьеру актёра начал в своем родном городе Бугуруслане.	Место жительства
На гастролях по Украине выступал под псевдонимом Дьяченков, который позже включил в официальную фамилию.	Профессиональные события
С 1950 года выступал в Тюменском драматическом театре, где за 34 года сыграл около 200 ролей.	Место работы
Его талант отмечали на столичных гастролях Михаил Ульянов и Юрий Яковлев.	Профессиональные события
Умер 4 февраля 1991 года в Тюмени.	Смерть
Пример 2. «Зампред ЦБ Торшин уходит с поста» (источник – «газета.ru», 30.11.2018)	
Зампред Банка России Александр Торшин покидает свой пост в связи с выходом на пенсию, говорится в сообщении ЦБ.	Профессиональные события
Торшин с 1999 по 2001 годы занимал должность статс-секретаря-заместителя генерального директора государственной корпорации «Агентство по реструктуризации кредитных организаций».	Род занятий
С 2001 по 2015 годы являлся членом Совета Федерации.	Место работы
В январе 2015 года он был назначен зампредседателем Центрального банка России.	Место работы
Ранее ФБР сообщало о наличии электронной переписки россиянки Марии Бутиной, подозреваемой в шпионаже в пользу России, с высокопоставленным сотрудником Центробанка России — по сообщениям СМИ, с заместителем председателя ЦБ Александром Торшиным.	Личные события

5. Заключение

В работе предложен подход к автоматическому поиску фрагментов, содержащих биографическую информацию в тексте на естественном языке, основанный на применении нейронных сетей. Для обучения продемонстрированных в работе нейросетевых моделей был составлен корпус

биографических текстов, содержащий биографические статьи, размещённые в онлайн-энциклопедии «Википедия». На основании разработанного корпуса были протестированы различные методы представления текстов и различные архитектуры нейронных сетей. Предложенный в работе подход демонстрирует достаточно высокие результаты на тестовой выборке (F-мера – 95,36%, точность классификации – 94,06%).

В перспективе планируется провести экспериментальные исследования на других данных, в том числе на биографических текстах, не отличающихся явной хронологией изложения, а также осуществить извлечение фактов из предложений, содержащих биографическую информацию, в структурированном виде.

Список литературы

[1]. Терпугова А. В. Биографический текст как объект лингвистического исследования. Автореферат дис. кандидата филологических наук. Ин-т языкознания РАН, Москва, 2011, 26 стр.

[2]. Manning C., Raghavan P., Schütze H. Introduction to Information Retrieval. Cambridge University Press, 2008. 506 p.

[3]. Адамович И. М., Волков О. И. Система извлечения биографических фактов из текстов исторической направленности. Системы и средства информатики, том 25, вып. 3, 2015 г., стр. 235-250.

[4]. Cybulska, A., Vossen, P. Historical Event Extraction From Text. In Proc. of 5th ACL-HLT Workshop on Language Technology on Cultural Heritage, 2011, pp. 39–43.

[5]. Hienert D., Luciano F. Extraction of Historical Events from Wikipedia. Lecture Notes in Computer Science, vol. 7540, 2015, pp. 16–28.

[6]. Santos C., Xiang B., Zhou B. Classifying Relations by Ranking with Convolutional Neural Networks. In Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 2015, pp. 626-634.

[7]. Meerkamp P., Zhou Z. Information Extraction with Character-level Neural Networks and Free Noisy Supervision. Cornell University Library [электронный ресурс]. 2016. URL: https://arxiv.org/abs/1612.04118 (дата обращения 21.09.2018).

[8]. Homma Y., Sadamitsu K., Nishida K., Higashinaka R., Asano H., Matsuo Y. A Hierarchical Neural Network for Information Extraction of Product Attribute and Condition Sentences. In Proc. of the Open Knowledge Base and Question Answering (OKBQA), 2016, pp. 21-29.

[9]. Arkhipenko K., Kozlov I., Trofimovich J., Skorniakov K., Gomzin A., Turdakov D. Comparison of Neural Architectures for Sentiment Analysis of Russian Tweets. In Proc. of the International Conference “Dialogue 2016”, 2016, pp. 50-58.

[10]. Андрианов И.А., Майоров В.Д., Турдаков Д.Ю. Современные методы аспектно-ориентированного анализа эмоциональной окраски. Труды ИСП РАН, том 27, вып. 5, 2015 г., стр. 5-22. DOI: 10.15514/ISPRAS-2015-27(5)-1.

[11]. Пархоменко П.А., Григорьев А.А., Астраханцев Н.А. Обзор и экспериментальное сравнение методов кластеризации текстов. Труды ИСП РАН, том 29, вып. 2, 2017 г., стр. 161-200. DOI: 10.15514/ISPRAS-2017-29(2)-6.

[12]. Ravuri S., Stolcke A. A Comparative Study of Recurrent Neural Network Models for Lexical Domain Classification. In Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 6075-6079
[13]. Yogatama D., Dyer C., Ling W., Blunsom P. Generative and discriminative text classification with recurrent neural networks. arXiv preprint arXiv:1703.01898, 2017.
[14]. Chen G., Ye D., Xing Z., Chen J., Cambria E. Ensemble application of convolutional and recurrent neural networks for multi-label text categorization. In Proc. of the International Joint Conference on Neural Networks (IJCNN), 2017, pp. 2377-2383.
[15]. Валгина Н.С., Розенталь Д.Э., Фомина М.И. Современный русский язык. Учебник. 6-е изд., перераб. и доп. Москва, Логос, 2002, 528 стр.
[16]. Википедия. Свободная энциклопедия. URL: https://ru.wikipedia.org/ (дата обращения: 26.11.2018).
[17]. Глазкова А. В. Формирование текстового корпуса для автоматического извлечения биографических фактов из русскоязычного текста. Современные информационные технологии и ИТ-образование, том 14, вып. 4, 2018 г.
[18]. Корпус биографических текстов, URL https://sites.google.com/site/utcorpus/ (дата обращения: 01.12.2018).
[19]. Морфологический анализатор pymorphy2, URL: https://pymorphy2.readthedocs.io/en/latest/ (дата обращения: 01.12.2018).
[20]. Mikolov T., Sutskever I., Chen K., Corrado G. S., Dean J. Distributed representations of words and phrases and their compositionality. In Proc. of the 26th International Conference on Neural Information Processing Systems, vol. 2, 2013, pp. 3111-3119.
[21]. Hochreiter S., Schmidhuber J. Long Short-term Memory. Neural computation, vol. 9, № 8, 1997, pp. 1735-1780.
[22]. Bai T., Dou H. J., Zhao W. X., Yang D. Y., Wen J. R. An Experimental Study of Text Representation Methods for Cross-Site Purchase Preference Prediction Using the Social Text Data. Journal of Computer Science and Technology, vol. 32, №. 4, 2017, pp. 828-842.
[23]. Keras: The Python Deep Learning library. URL: https://keras.io/ (дата обращения: 17.11.2018).
[24]. URL: https://github.com/oldaandozerskaya/biographical_samples.git (дата обращения: 27.12.2018).
[25]. газета.ru. URL: https://www.gazeta.ru/ (дата обращения: 09.12.2018).

Automatic search for fragments containing biographical information in a natural language text

A.V. Glazkova <a.v.glazkova@utmn.ru>
University of Tyumen,
6, Volodarsky st., Tyumen, 625003, Russia

Abstract. The search and classification of text documents are used in many practical applications. These are the key tasks of information retrieval. Methods of text searching and classifying are used in search engines, electronic libraries and catalogs, systems for collecting and processing information, online education and many others. There are a large number of particular applications of these methods, but each such practical task is characterized, as a rule, by weak formalizability and narrow objectivity. Therefore, it requires individual study and its own approach to the solution. This paper discusses the task of automatically searching

and typing text fragments containing biographical information. The key problem in solving this problem is to conduct a multi-class classification of text fragments, depending on the presence and type of biographical information contained in them. After reviewing the related works, the author concluded that the use of neural network methods is promising and widespread for solving such problems. Based on this conclusion, the paper compares various architectures of neural network models, as well as basic text presentation methods (Bag-Of-Words, TF-IDF, Word2Vec) on a pre-assembled and marked corpus of biographical texts. The article describes the steps involved in preparing a training set of text fragments for teaching models, methods for text representation and classification methods chosen for solving the problem. The results of the multi-class classification of text fragments are also presented. The examples of automatic search for fragments containing biographical information are shown for the texts that did not participate in the model learning process.

Keywords: text classification; natural language processing; word embedding; neural networks; biographical text.

DOI: 10.15514/ISPRAS-2018-30(6)-12

For citation: Glazkova A.V. Automatic search for fragments containing biographical information in a natural language text. Trudy ISP RAN/Proc. ISP RAS, vol. 30, issue 6, 2018, pp. 221-236 (in Russian). DOI: 10.15514/ISPRAS-2018-30(6)-12

References

[1]. Terpugova A.V. Biographical text as an object of linguistic researchю. Author's abstract of the PhD thesis. Institute of Linguistics RAS, Moscow, 2011, 26 p. (in Russian).
[2]. Manning C., Raghavan P., Schütze H. Introduction to Information Retrieval. Cambridge University Press, 2008. 506 p.
[3]. Adamovich I.M., Volkov O.I. The system of facts extraction from historical texts. Sistemy i sredstva informatiki [Systems and Means of Informatics], vol. 25, № 3, 2015, p. 235-250 (in Russian).
[4]. Cybulska, A., Vossen, P. Historical Event Extraction From Text. In Proc. of 5th ACL-HLT Workshop on Language Technology on Cultural Heritage, 2011, pp. 39–43.
[5]. Hienert D., Luciano F. Extraction of Historical Events from Wikipedia. Lecture Notes in Computer Science, vol. 7540, 2015, pp. 16–28.
[6]. Santos C., Xiang B., Zhou B. Classifying Relations by Ranking with Convolutional Neural Networks. In Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 2015, pp. 626-634.
[7]. Meerkamp P., Zhou Z. Information Extraction with Character-level Neural Networks and Free Noisy Supervision. Cornell University Library [электронный ресурс]. 2016. URL: https://arxiv.org/abs/1612.04118 (дата обращения 21.09.2018).
[8]. Homma Y., Sadamitsu K., Nishida K., Higashinaka R., Asano H., Matsuo Y. A Hierarchical Neural Network for Information Extraction of Product Attribute and Condition Sentences. In Proc. of the Open Knowledge Base and Question Answering (OKBQA), 2016, pp. 21-29.
[9]. Arkhipenko K., Kozlov I., Trofimovich J., Skorniakov K., Gomzin A., Turdakov D. Comparison of Neural Architectures for Sentiment Analysis of Russian Tweets. In Proc. of the International Conference "Dialogue 2016", 2016, pp. 50-58.

[10]. Andrianov I., Mayorov V., Turdakov D. Modern Approaches to Aspect-Based Sentiment Analysis. Trudy ISP RAN/Proc. ISP RAN, vol. 27, №. 5, 2015 г., p. 5-22 (in Russian). DOI: 10.15514/ISPRAS-2015-27(5)-1.
[11]. Parhomenko P.A., Grigorev A.A., Astrakhantsev N.A. A survey and an experimental comparison of methods for text clustering: application to scientific articles. Trudy ISP RAN/Proc. ISP RAN, vol. 29, №. 2, 2017 г., p. 161-200 (in Russian). DOI: 10.15514/ISPRAS-2017-29(2)-6.
[12]. Ravuri S., Stolcke A. A Comparative Study of Recurrent Neural Network Models for Lexical Domain Classification. In Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 6075-6079
[13]. Yogatama D., Dyer C., Ling W., Blunsom P. Generative and discriminative text classification with recurrent neural networks. arXiv preprint arXiv:1703.01898, 2017.
[14]. Chen G., Ye D., Xing Z., Chen J., Cambria E. Ensemble application of convolutional and recurrent neural networks for multi-label text categorization. In Proc. of the International Joint Conference on Neural Networks (IJCNN), 2017, pp. 2377-2383.
[15]. Valgina N.S., Rosental D.E., Fomina M.I. Modern Russian Language. Moscow, Logos, 2002, 528 p. (in Russian).
[16]. Wikipedia. The free encyclopedia. URL: https://ru.wikipedia.org/, accessed 26.11.2018.
[17]. Glazkova A. V. Building a text corpus for automatic biographical facts extraction from Russian texts. Sovremennyye informatsionnyye tekhnologii i IT-obrazovaniye [Modern Information Technologies and IT-education], vol 14, №. 4, 2018 (in Russian).
[18]. The corpus of biographical texts, URL https://sites.google.com/site/utcorpus/, accessed 01.12.2018.
[19]. Morphological analyzer pymorphy2, URL: [19]. https://pymorphy2.readthedocs.io/en/latest/, accessed 01.12.2018.
[20]. Mikolov T., Sutskever I., Chen K., Corrado G. S., Dean J. Distributed representations of words and phrases and their compositionality. In Proc. of the 26th International Conference on Neural Information Processing Systems, vol. 2, 2013, pp. 3111-3119.
[21]. Hochreiter S., Schmidhuber J. Long Short-term Memory. Neural computation, vol. 9, № 8, 1997, pp. 1735-1780.
[22]. Bai T., Dou H. J., Zhao W. X., Yang D. Y., Wen J. R. An Experimental Study of Text Representation Methods for Cross-Site Purchase Preference Prediction Using the Social Text Data. Journal of Computer Science and Technology, vol. 32, №. 4, 2017, pp. 828-842.
[23]. Keras: The Python Deep Learning library. URL: https://keras.io/, accessed 17.11.2018.
[24]. URL: https://github.com/oldaandozerskaya/biographical_samples.git, accessed 27.12.2018.
[25]. [gazeta.ru]. URL: https://www.gazeta.ru/, accessed 09.12.2018.