Гибридный алгоритм классификации кандидатов в термины текста предметной области

И.А. Андреев
Ульяновский государственный технический университет
Ульяновск, Россия
ares-ilya@yandex.ru

В.С. Мошкин
Ульяновский государственный технический университет
Ульяновск, Россия
postforvadim@ya.ru

Н.Г. Ярушкина
Ульяновский государственный технический университет
Ульяновск, Россия
jng@ulstu.ru

***Аннотация*—в работе описывается метод классификации кандидатов в термины проблемной области при помощи лингвистических методов с использованием нейронных сетей. Приведен алгоритм работы, представлены результаты экспериментов. По результатам работы достигнут высокий показатель точности.**

Ключевые слова— термины, стемминг, нейронные сети, машинное обучение, лингвистика.

1. Введение

Выделение терминов предметной области – важная задача, которая полезна для широкого спектра задач. В классическом случае составление словаря терминов – это полностью ручной труд эксперта в данной проблемной области. Это занимает очень много времени и высокие материальные затраты, и, кроме того, на итоговый список накладывается отпечаток личного опыта эксперта.

Создание списка терминов без участия эксперта – это задача, решаемая автоматизированной системой классификации кандидатов в термины [1].

2. Гибридный метод классификации кандидатов в термины

Первый этап алгоритма составления списка кандидатов в термины из текста предметной области – получение лингвистических характеристик слов, составляющих данный текст.

Одним из вариантов такой обработки является стемминг. Для разметки текста была выбрана программа Mystem, отвечающая необходимым требованиям [2].

Предлагаемый алгоритм автоматизированного составления списка терминов предполагает результат в виде текстового документа, содержащего список терминов, составленных на основе текста при помощи лингвистического метода и отобранных нейронной сетью.

Первое, что делает пользователь – загружает текст на сервер. После загрузки начинается автоматизированная обработка текста. Для начала текст конвертируется в нужную кодировку и обрабатывается программой Mystem, которая размечает текст при помощи лингвистических метрик. После разметки загруженный в базу данных текст обрабатывается лингвистическим методом формирования кандидатов в термины [3].

Лингвистический метод извлечения терминологии можно разделить на две части: морфологическую и лексическую.

Морфологический анализ текста:

- определение морфологических признаков и частей речи словоупотреблений;
- определение канонических (начальных, нормализованных) форм слов;
- выделение значимых лексико-грамматических классов [4].

Лексический анализ текста – это извлечение терминов произвольной длины, удовлетворяющих лексико-грамматическим классам [5] [6].

Для возможности выделения терминов из текстов предметной области были разработаны лингвистические шаблоны, с помощью которых удается выделить основные термины. В русском языке синтаксическая структура терминов предметной области более чем в 90 % случаев соответствует следующим пяти шаблонам:

- одиночные сущ.;
- сущ. + сущ.;
- прил. + сущ.;
- прил. + прил. + сущ.;
- сущ. + прил. + сущ.

Вторым этапом является запуск подпрограммы обработки списка терминов при помощи нейронной сети. Нейронная сеть имеет 256 входов. Нейронная сеть представляет собой однослойный персептрон с прямыми связями, работающий с методом обратного распространения ошибки. Для нормализации входных данных используется функция активации, представленная формулой:

$$R = \frac{1}{1+ exp(-S)},$$

где R – результат функции активации, S – сумма, получаемая со входов нейронной сети с весами каждого входа [7].

Этот этап запускает пользователь в отдельном приложении. На выходе пользователь получает список прошедших отбор кандидатов в термины. Алгоритм представлен в виде блок-схемы на рисунке 1.

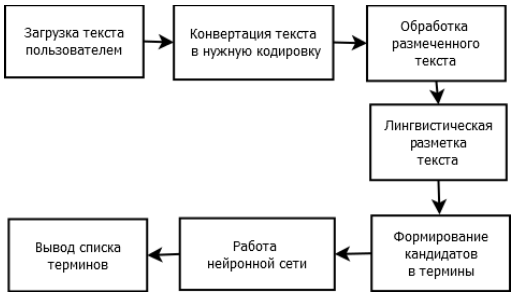

Рис. 1. Алгоритм работы системы

Следующим этапом алгоритма является извлечение терминов на основе обработанных слов текста. Входом для алгоритма является множество существительных, прилагательных, глаголов и служебных частей речи обработанного текста с морфологической информацией о них.

3. Результаты экспериментов

В качестве входных данных для выполнения эксперимента был выбран текст объемом 40000 слов по теме «Временные ряды». В результате работы системы был получен список терминов, который позволяет более детально изучить полученные результаты экспертом. Перед работой нейронная сеть была обучена на тестовой выборке.

Оцениваются два списка терминов, составленных по одному тексту:

- первый список составлен экспертом;
- второй – программной системой.

При составлении списков использованы одинаковые наборы шаблонов. При нахождении экспертом термина, не подходящего по шаблону, термин игнорируется и не заносится в результирующий список.

Для эксперимента нейронная сеть обучена предварительно на основе списка, содержащего 115 элементов, как являющимися, так и не являющимися терминами.

В результате работы разработанного программного обеспечения и эксперта были сформированы два списка терминов: результат работы нейронной сети, сформированный сервисом, и эталонный – сформированная экспертом.

Общий список терминов, составленный на основе суммы обоих списков терминов, исключая ошибочные, составил 2318 терминов. Характеристики полученных списков представлены в виде таблицы I

Таблица I. Характеристики списком терминов

Источник списка терминов	Количество терминов	Количество ошибок	Необнаруженных терминов
Эксперт	2157	0	52
Разработанное ПО	2286	97	20

На основе полученных данных было рассчитано качество работы программного обеспечения. Процент безошибочных определений терминов составил 100%. Это связано с тем, что, несмотря на некоторое количество ошибок и неопределённых терминов, программное обеспечение определило термины, пропущенные экспертом при вычитывании.

Процент неопределённых терминов, видимых экспертом при вычитывании текста составил 1%. Это позволяет сделать вывод, что разработанная система в целом может применяться как замена экспертному методу с условием проверки вывода экспертом, т.к. количество ошибочных терминов отлично от 0.

4. Заключение

В данной работе рассмотрена проблема автоматизированного формирования списка терминов по тексту предметной области, поднята проблема оценки качества сформированного списка терминов. Был описан процесс автоматизированного составления списка терминов, а также рассмотрен разработанный алгоритм.

Для оценки качества сформированного списка терминов были приглашены эксперты, которые также составили списки терминов по текстам, выбранным для эксперимента. Проведенные эксперименты показали, что результаты работы информационной системы сопоставимы с экспертными результатами, однако уступают им за счёт некоторого количества ошибок, допускаемых системой.

Перспектива исследования может заключаться в интеграции новых лингвостатистических методов и изменении характеристик нейронной сети.

Литература

[1] Ковязина, М.А. Извлечение ключевых терминов на базе корпуса текстов о разработке нефтяных и газовых месторождений // Вестник Тюменского государственного университета. Гуманитарные исследования. Humanitates. – 2016. – Т. 2, № 3. – С. 61-69.

[2] Андреев, И.А. Семантическая метрика терминологичности на основе онтологии предметной области / И.А. Андреев, В.А. Башаев, В.В. Клейн, В.С. Мошкин, Н.Г. Ярушкина // Автоматизация процессов управления. – 2014. – № 4(38). – С. 76-84.

[3] Kustikova, V.D. A Survey of Deep Learning Methods and Software for Image Classification and Object Detection / V.D. Kustikova, P.N. Druzhkov // Proc. of the 9th Open German-Russian Workshop on Pattern Recognition and Image Understanding, 2014.

[4] Мошкин, В.С. Семантическая метрика «термин/не термин» на основе онтологии проблемной области / В.С. Мошкин, И.А. Андреев, В.А. Башаев, В.В. Клейн // Методы и технологии гибридного и синергетического искусственного интеллекта: материалы I международной Поспеловской летней школы-семинара для студентов, магистрантов и аспирантов. – Калининград: Изд-во БФУ им. И. Канта, 2014. – С. 67-73.

[5] Лукашевич, Н.В. Использование методов машинного обучения для извлечения слов-терминов / Н.В. Лукашевич, М.Ю. Логачев // Компьютерная лингвистика и семантический Web: по материалам двенадцатой национальной конференции по искусственному интеллекту КИИ, 2010.

[6] Браславский, П.И. Сравнение четырех методов автоматического извлечения двухсловных терминов из текста / П.И. Браславский, Е.А. Соколов // Компьютерная лингвистика и интеллектуальные технологии. – М.: Изд-во РГГУ, 2006. – С. 88-91.

[7] Yarushkina, N.G. Hybridization of Fuzzy Inference and Self-learning Fuzzy Ontology-Based Semantic Data Analysis / N.G. Yarushkina, V.S. Moshkin, I.A. Andreev, V.V. Klein, E. Beksaeva // Proceedings of the First International Scientific Conference “Intelligent Information Technologies for Industry” (IITI’16). – Springer International, 2016.