УДК 81’33+811.581

Красикова Елизавета Александровна
кандидат филологических наук,
доцент кафедры подготовки
преподавателей редких языков
Московский государственный
лингвистический университет
г. Москва, Россия

Elizaveta Krasikova
Phd in Philology, Associate Professor
of the Department of Training Teachers
of Rare Languages
Moscow State Linguistic University
Moscow, Russia
krasikova.liza@mail.ru

ПОТЕНЦИАЛ КОРПУСНОГО МЕНЕДЖЕРА ДЛЯ ОБРАБОТКИ ЛИНГВИСТИЧЕСКИХ ДАННЫХ (на примере корпуса китайских электронных СМИ)

В исследовании тестируются возможности программного комплекса «Генератор сбалансированного лингвистического корпуса и корпусный менеджер». В несколько этапов решаются такие задачи, как отбор лингвистического материала, формирование тестового корпуса актуальных текстов электронных СМИ на китайском языке, апробация частеречной разметки на материале языка изолирующего типа, установление уровня точности работы модуля «китайский язык». В ходе анализа было установлено, что в отличие от индоевропейских языков, которые ранее служили апробационным материалом для тестирования программного комплекса, китайский язык вносит особенности в алгоритм функционирования системы в силу типологических особенностей. Полученные в ходе запросов лингвистические и статистические данные были подвергнуты анализу, в результате которого было установлено, что погрешность определения заявленных частей речи составляет 7–8 %. В качестве перспективы исследования рассматривается оптимизация поиска данных в рамках модуля «китайский язык» и формирование ряда алгоритмов поиска частей речи в заданном лингвистическом корпусе.

К л ю ч е в ы е с л о в а: *цифровая экономика; прикладная лингвистика; корпусный менеджер; искусственный интеллект; лингвистический корпус; китайский язык; электронные СМИ.*

THE POTENTIAL OF THE CORPUS MANAGER IN PROCESSING LINGUISTIC DATA (USING THE EXAMPLE OF THE CORPUS OF CHINESE ELECTRONIC MEDIA)

The article describes the capabilities of the software package “Balanced linguistic corpus generator and corpus manager”. Tasks such as the selection of linguistic material, the formation of a test corpus of relevant electronic media texts in Chinese, the testing of partial markup on the material of an isolating language, and the establishment of the accuracy level of the Chinese language module are solved in several stages. During the analysis, it was found that, unlike Indo-European languages, which previously served as an approbation material for testing the software package, the Chinese language introduces features into the algorithm of the system's functioning due to its typological features. The linguistic and statistical data obtained during the queries were analyzed, as a result of which it was found that the error in determining the declared parts of speech is 7–8 %. The optimization of data search within the framework of the “Chinese language” module and the formation of a number of algorithms for searching parts of speech in a given linguistic corpus are considered as research prospects.

K e y w o r d s: *digital economy; applied linguistics; corpus manager; artificial intelligence; linguistic corpus; Chinese language; electronic media.*

Данная статья посвящена анализу потенциала искусственного интеллекта в лингвистических исследованиях. Востребованность изучения и решения вышеуказанной задачи обусловлена стратегическими целями развития Российской Федерации на период до 2030 г. В рамках реализации Указа Президента Российской Федерации от 21.07.2020 г. № 474 «О национальных целях развития Российской Федерации на период до 2030 года» одним из направлений деятельности национального проекта «Цифровая экономика» является «Искуственный интеллект»[1].

Важнейшей составляющей технологии искусственного интеллекта являются инструменты обработки естественного языка (Natural Language Processing – NLP). Такого рода «программное обеспечение … дает широкие возможности в области лингвистических исследований, но в то же время не охватывают все особенности языковых явлений» [5, с. 171]. Ранее исследования такого рода проводились на материале индоевропейских языков, в том числе для анализа фразеологических единиц [1], единиц бытовой лексики [2], общего анализа текстов СМИ и художественной литературы [3; 9; 10]. В связи с вышеизложенным научный интерес представляет апробация искусственного интеллекта на материале китайского языка, принадлежащего к сино-тибетской группе. Примыкая к языкам изолирующего типа, китайский язык характеризуется рядом типологических свойств, которые, по нашему мнению, могут представлять определенные трудности для программного обеспечения в обработке лингвистических данных. Например, следует учитывать такие свойства китайского языка, которые отличают его от индоевропейских, как:

а) невыделимость морфемы как существующей вне слова величины, меньшей, чем слово;

б) способность выделяемой из слова части (основы или корня) к отдельному употреблению;

в) две формы существования слов;

г) функционирование односложного слова в виде нулевой, то есть в виде абсолютной формы;

д) факультативность грамматических показателей;

е) широкое распространение номинативных единиц, обладающих свойствами как слова, так и словосочетания, и т. п. [8, с. 12].

Целью данного исследования является тестирование разрабатываемого в лаборатории фундаментальных и прикладных проблем виртуального образования Московского государственного лингвистического университета программного комплекса – корпусного менеджера [6] и генератора баз данных, которые мы условно назовем стандартным «сбалансированным лингвистическим корпусом» [4, с. 3889]. Данное исследование включало в себя не-

[1] https://digital.gov.ru/ru/activity/directions/1046/.

сколько этапов. На первом этапе был собран лингвистический корпус текстов электронных СМИ на китайском языке в период с февраля по июнь 2024 года. Объем корпуса составил 724 предложения, или 18341 токен. В качестве источника нами был выбран портал информационного агентства 新华 *Синьхуа*[1] , которое является крупнейшим официальным информационным центром правительства КНР на сегодняшний день. Выбор данного источника, прежде всего, обусловлен его информативностью, поскольку помимо изобилия информационного материала о мировых событиях «для посетителей сайта представлены эксклюзивные материалы о современном Китае в текстовом, графическом, аудио- и видеоформате» [7, с. 11].

На втором этапе исследования был сформирован ряд поисковых запросов к корпусу, в результате которых были получены количественные данные, позволившие установить уровень точности работы модуля. Полученные статистические данные можно представить в виде таблицы:

Результаты поиска по алгоритму «Части речи»

Запрос	Идентифицировано ед.	Частотность употребления, %	Погрешность
“NOUN” Нарицательное имя существительное	5602	30,54	++
“VERB” Глагол	3342	18,22	+
“PROPN” Собственное имя существительное	1487	8,10	+
“PART” Частица	1074	5,85	+
“NUM” Числительное	813	4,43	+
“ADJ” Имя прилагательное	566	3,08	++

Как видно из таблицы, наиболее частотными по употреблению оказались запросы “NOUN” и “VERB”. Среднюю частотность употребления показали такие запросы, как “PROPN”, “PART”, “NUM”, низкую частность продемонстрировал запрос “ADJ”, что может быть обусловлено повышенной погрешностью, вызванной наличием в китайском языке грамматического показателя (например, структурная частица 的), который программной системе не удалось распознать в качестве грамматического маркера.

Далее приведем несколько примеров успешной идентификации указанных запросов.

[1] http://m.news.cn/.

1. **专家** (NOUN) 但**专家**认为，在日本和美国息差难以缩小的情况下，单方面干预不可能解决根本问题。*Однако* ***эксперты*** *полагают, что в случае, когда разрыв в процентных ставках между Японией и Соединенными Штатами трудно сократить, одностороннее вмешательство не может решить фундаментальную проблему*.

2. **同意** (VERB) 美方先前不**同意**乌方用美制武器打击俄境内目标，担心引发俄乌冲突进一步升级。*Американская сторона ранее не* ***соглашалась*** *с использованием Украиной оружия американского производства для нанесения ударов по целям в России, опасаясь, что это приведет к дальнейшей эскалации российско-украинского конфликта*.

3. **习近平** (PROPN) 这是和**习近平**总书记面对面交流过的代表委员们的共同感触。*Таково общее мнение представителей и членов Совета, которые лично общались с Генеральным секретарем* ***Си Цзиньпином***.

4. **之** (PART) 一名以色列官员20日表示，以色列与莱希遇难**之**事无关。*Израильский чиновник 20-го числа заявил, что Израиль не имеет никакого отношения к прекращению деятельности «Лехи».*

5. **三** (NUM) 印尼央行上月将**三**项主要利率水平上调25个基点，以“加强印尼盾汇率稳定”。*В прошлом месяце Центральный банк Индонезии повысил* ***три*** *основные процентные ставки на 25 базисных пунктов, чтобы “укрепить стабильность обменного курса рупии*”.

6. **良好** (ADJ) 拉夫罗夫转达普京总统对习近平主席的亲切问候和**良好**祝愿。*Лавров передал сердечные приветствия и* ***добрые*** *пожелания Президента Путина Председателю КНР Си Цзиньпину*.

Вышеуказанные примеры свидетельствуют о том, что корпусный менеджер успешно обнаружил не только односложные, но и состоящие из нескольких иероглифических знаков существительные, глаголы и прилагательные. В качестве числительных системой были распознаны не только символы в виде арабских цифр, но и цифры, зафиксированные иероглифическим письмом.

Далее представляется необходимым отметить ряд трудностей, с которыми столкнулось программное обеспечение при идентификации и обработке тестируемых запросов. Наиболее затруднительными оказались случаи выявления существительных и прилагательных. К примеру, система распознает сочетание иероглифических знаков 人民 как существительное, однако в нижеуказанной контекстной реализации данная лексема представляет собой имя прилагательное, которое занимает позицию определения к последующему определяемому слову:

7. **人民**海军忠于党，舰行万里不迷航。***Народный*** *военно-морской флот верен партии, и его корабли проходят тысячи миль, не сбиваясь с курса*.

При апробации запроса “PROPN” в качестве имен собственных система идентифицировала идиоматическое выражение 螳臂当车 (досл. *богомол лапками задерживает колесницу*):

8. “面对强大的人民军队，‘台独’分子的任何分裂行径都是螳臂当车，自欺欺人。” *Перед лицом мощной народной армии любое сепаратистское поведение сторонников независимости Тайваня – это* ***напрасные потуги***.”

Запрос “NUM” также выявил небольшой процент погрешности. Хотя программному обеспечению удалось выявить количественные, порядковые, дробные числительные и проценты, записанные как арабскими цифрами, так и иероглифами, однако наряду с вышеуказанными единицами к данной части речи системой были отнесены счетные слова (например, 个; 幅; 轮; 种), наречие 多 (‘*много’*), отрицательная частица+ наречие 不少 (‘немало’).

Таким образом, на основе полученных данных можно сделать вывод о том, что качество работы модуля «китайский язык» программного комплекса «Генератор сбалансированного лингвистического корпуса и корпусный менеджер» находится на высоком уровне. Хотя запросы были выполнены без технических сбоев, однако была установлена небольшая погрешность, которая находится в пределах 7–8 %. Выявленные случаи погрешности позволяют сформировать ряд рекомендаций по улучшению функционирования системы: 1) формирование списка «стоп-слов» для уменьшения погрешности; 2) тестирование системы при помощи алгоритмов РСЗ (ручной запрос специальный); 3) расширение списка запросов по частям речи на материале китайского языка с целью выявления наиболее «слабых мест» для системы.

ЛИТЕРАТУРА

1. *Бахтигозина В. С.* Проблема поиска фразеологизмов в лингвистическом корпусе, сформированном по правилам Spacy // Человек – язык – компьютер. Исследователи будущего : материалы научно-практической (заочной) конференции с международным участием, Москва, 25 декабря 2023 года. М., 2024. С. 111–117.

2. *Бондарчук Г. Г.* Семиотические функции английских наименований одежды в публицистическом тексте (корпусное исследование) // Вестник Московского государственного лингвистического университета. Гуманитарные науки. 2024. № 4 (885). С. 23–29. EDN BXILCR.

3. *Горожанов А. И.* Алгоритмы поиска фразеологизмов в лингвистическом корпусе с морфологической разметкой (индоевропейские языки) // Филологические науки. Вопросы теории и практики. 2024. Т. 17, №1. С. 132–138.

4. *Горожанов А. И.* Расширение стандартного сбалансированного лингвистического корпуса, построенного по правилам spaCy, коннотативными характеристиками // Филологические науки. Вопросы теории и практики. 2023. Т. 16, № 11. С. 3888–3893. DOI 10.30853/phil20230594.

5. *Кириллина Е. В.*, *Иванов Н. Н.* Особенности лингвистического анализа текста компьютерными программами (на примере обработки естественного языка) // Казанская наука. 2020. № 12. С. 171–173.

6. *Лемешко Ю. Г., Лютова Ю. А* Специфика функционирования информационного агентства «Синьхуа» в эпоху глобализации // Вестник Амурского государственного университета. Серия: Гуманитарные науки. 2012. № 58. С. 10–13.

7. *Свидетельство* о государственной регистрации программы для ЭВМ № 2023683209 Российская Федерация. «Генератор сбалансированного лингвистического корпуса и корпусный менеджер»: № 2023682269 : заявл. 25.10.2023 : опубл. 03.11.2023 / А. И. Горожанов ; заявитель – федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный лингвистический университет». EDN JHFXUV.

8. *Солнцев В. М.* Типологические свойства изолирующих языков (на материале китайского и вьетнамского языков) // Языки Юго-Восточной Азии. Проблемы морфологии, фонетики и фонологии. М., 1970. С. 11–19.

9. *Степанова Д. В*. Программный комплекс для генерации динамического корпуса текстов СМИ // Вестник Минского государственного лингвистического университета. Серия 1: Филология. 2023. № 6 (127). С. 123–130.

10. *Gorozhanov A. I,. Guseynova I. A., Stepanova D. V*. Natural Language Processing and Fiction Text: Basis for Corpus Research // RUDN Journal of Language Studies, Semiotics and Semantics. 2024. Vol. 15, No. 1. P. 195–210.