ИЗДАТЕЛЬСТВО ГРАМОТА

Филологические науки. Вопросы теории и практики
Philology. Theory & Practice

ISSN 2782-4543 (online)
ISSN 1997-2911 (print)

2024. Том 17. Выпуск 9 | 2024. Volume 17. Issue 9
Материалы журнала доступны на сайте (articles and issues available at): philology-journal.ru

RU

Автоматическое выделение именованных сущностей в китайско-русском корпусе параллельных и сопоставимых текстов политической тематики

Чжу Хуэй, Митрофанова О. А.

Аннотация. Цель исследования заключается в том, чтобы экспериментальным путем выявить и интерпретировать стандартные и вложенные именованные сущности в китайских и русских политических текстах, общие и специфические для сравниваемых языков, с помощью библиотек HanLP и SpaCy. В ходе исследования был создан китайско-русский корпус параллельных и сопоставимых текстов политической тематики. Научная новизна исследования состоит в том, что в нем представлены результаты распознавания различных именованных сущностей и систематизированы типы ошибок в китайско-русском корпусе параллельных и сопоставимых политических текстов. В результате исследования установлено, что наиболее частотными именованными сущностями в оригинальных китайских и русских политических текстах являются названия локаций, следующие по частоте – это названия организаций, реже всего встречаются названия персон. Большинство высокочастотных именованных сущностей в китайских оригинальных и переводных текстах в основном соответствуют друг другу. Это доказывает, что переводчики чаще всего используют дословный перевод при передаче именованных сущностей с китайского языка на русский в политических текстах. В нашем исследовании систематизируется и обобщается информация о вложенных именованных сущностях в политических текстах, выделены и проанализированы следующие их типы: [[локация]ЛОКАЦИЯ], [[локация]ОРГАНИЗАЦИЯ], [[цифра]ОРГАНИЗАЦИЯ], [[локация]ОБЪЕКТ], [[локация]ПРОЕКТ].

EN

Automatic extraction of named entities in the Chinese-Russian corpus of parallel and comparable texts on political topics

Hui Zhu, O. A. Mitrofanova

Abstract. The aim of the research is to experimentally identify and interpret standard and nested named entities in Chinese and Russian political texts, common and specific to the compared languages, using HanLP and SpaCy libraries. During the study, a Chinese-Russian corpus of parallel and comparable political texts was created. The scientific novelty of the research lies in presenting the results of recognizing various named entities and systematizing the types of errors in the Chinese-Russian corpus of parallel and comparable political texts. The study found that the most frequent named entities in original Chinese and Russian political texts are location names, followed by organization names, with person names being the least frequent. Most high-frequency named entities in original Chinese and translated texts generally correspond to each other, proving that translators often use literal translation when rendering named entities from Chinese into Russian in political texts. Our research systematizes and summarizes information on nested named entities in political texts, identifying and analyzing the following types: [[location]LOCATION], [[location]ORGANIZATION], [[number]ORGANIZATION], [[location]OBJECT], [[location]PROJECT].

Введение

Цифровая коммуникация оказывает влияние на типы разметки лингвистической информации в корпусах текстов и на параметры поиска документов в веб-пространстве. Одним из направлений исследований

Научная статья (original research article) | https://doi.org/10.30853/phil20240430

в области компьютерной лингвистики и автоматического понимания текстов, стимулируемых задачами поиска и извлечения информации, являются **распознавание, выделение и классификация именованных сущностей (Named Entity Recognition, NER)**, приобретающие высокое значение для процедур анализа текстов разных стилей, жанров и тематики (Keraghel, Morbieu, Nadif, 2024; Nadeau, Sekine, 2007). Процедуры идентификации именованных сущностей (Named Entity, NE) оказывают влияние на качество результатов машинного перевода, построения графов знаний, реферирования и аннотирования, генерации заголовков, наборов ключевых выражений, обучения моделей классификации и кластеризации документов в корпусе и т. д.

Теоретической базой исследования послужили труды таких известных ученых и разработчиков, как Д. Надо и С. Секин (Nadeau, Sekine, 2007), И. Керагель, С. Морбье, М. Надиф (Keraghel, Morbieu, Nadif, 2024), Е. И. Большакова и Н. Э. Ефремова (2017), Е. А. Филиппова (2017), посвященные типологии именованных сущностей, методам и алгоритмам их выделения. Объектом нашего исследования является именованная сущность, которая представляет собой слово или словосочетание, описывающее конкретный и четко определенный объект или явление, а также классы таких объектов. Предмет исследования – методы и алгоритмы выделения именованных сущностей в типологически различных языках, китайском и русском.

Англоязычный термин Named Entity (именованная сущность) был введен на конференции Message Understanding (MUC) (Grishman, Sundheim, 1996), и с того времени он вошел в терминологический стандарт корпусной и компьютерной лингвистики. Понятие «именованной сущности» в исследованиях по автоматической обработке естественного языка выходит за рамки традиционных трактовок имен собственных (Суперанская, 1973; Сталтмане, 1989), поскольку «именованные сущности» могут сохранять предметно-понятийную соотнесенность благодаря включению в их состав имен нарицательных (*восстание декабристов на Сенатской площади*), количественных групп (*14 (26) декабря 1825 года*), специальной лексики (термины предметной области – *перцептрон Розенблатта*) и т. д. Следуя работам (Брыкина, Файнвейц, Толдова, 2013; Большакова, Ефремова, 2017; Филиппова, 2017), в данной статье мы будем использовать русскоязычный термин «именованная сущность» для характеристики объекта исследования.

Актуальность проводимого исследования определяется необходимостью сбора и систематизации данных о типологии и функционировании именованных сущностей в различных языках и выявлением сходств и различий в их использовании в параллельных и сопоставимых корпусах текстов на языках из разных групп. По данным CLARIN (https://www.clarin.eu/resource-families/tools-named-entity-recognition), основные ресурсы и инструменты (GATE (https://gate.ac.uk/), OpenNLP (https://opennlp.apache.org/), FreeLing (https://nlp.lsi.upc.edu/freeling/node/1), NameTag (https://ufal.mff.cuni.cz/nametag) и т. д.) для выделения именованных сущностей разрабатывались для английского языка, эти результаты экстраполировались на материал других языков: немецкого, чешского, голландского, финского, русского и т. д. Важной тенденцией развития корпусной и компьютерной лингвистики является разработка специализированных корпусов текстов, в которых решаются конкретные задачи (многоуровневая разметка корпусов текстов, разрешение морфологической, лексико-семантической, синтаксической неоднозначности, разрешение анафоры, выделение терминов и терминологических сочетаний и т. д.). В этой связи создание корпусов текстов с разметкой именованных сущностей становится всё более актуальным как для общеязыковых корпусов (Li, Sun, Han et al., 2020), так и корпусов текстов определенных предметных областей – биомедицина (Zhou, Zhang, Su et al., 2004; Wang, Yang, Guan, 2018), финансы (Alvarado, Verspoor, Baldwin, 2015; Zhang, Zhang, 2022), социальные медиа (Tran, Hwang, Jung, 2015; Li, Sun, Weng et al., 2014), юриспруденция (Au, Lampos, Cox, 2022; Luz de Araujo, De Campos, De Oliveira et al., 2018). Наше исследование восполняет пробелы, связанные с автоматической разметкой именованных сущностей в текстах политической тематики на материале китайского и русского языков.

Достижение цели настоящего исследования требует решения ряда задач, в частности:

1) предварительной обработки китайско-русского корпуса параллельных и сопоставимых текстов политической тематики;

2) систематизации информации о типах именованных сущностей и методах их выявления, выбора алгоритмов и инструментов выделения именованных сущностей в корпусах текстов для русского и китайского языков;

3) проведения экспериментов по выделению именованных сущностей на материале китайско-русского корпуса параллельных и сопоставимых текстов политической тематики, обработки результатов и анализа ошибок алгоритмов.

В соответствии с поставленными задачами были определены следующие методы исследования: корпусные методы применяются на этапе разработки, предобработки и разметки корпуса, методы языкового моделирования и статистической обработки данных используются для идентификации и анализа NE, количественные методы необходимы для верификации результатов экспериментов. Материалом исследования является разработанный нами китайско-русский корпус параллельных и сопоставимых политических текстов, который состоит из двух подкорпусов: первый подкорпус – это параллельный корпус «Докладов о работе правительства в 2012-2022 гг.» (далее – ДРП), включающий в себя исходный китайский текст (далее – ДРП-К) и перевод на русский язык (далее – ДРП-Р). Второй подкорпус – это сопоставимый корпус «Послания Президента Российской Федерации Федеральному Собранию РФ 2011-2021 гг.» (далее – ППР). Основные параметры корпуса и процедуры его обработки представлены в публикации (Чжу, Захаров, 2024).

Теоретическая значимость работы состоит в обобщении лингвистической информации об именованных сущностях в китайском и русском языках, что обогащает представления о принципах их перевода в процессе обработки текстов политической тематики. Практическая значимость исследования определяется тем, что в нем содержатся результаты идентификации NE и систематизированы типы возможных ошибок разметки NE в китайско-русском корпусе параллельных и сопоставимых политических текстов.

Обсуждение и результаты

Особенности разметки именованных сущностей в корпусах текстов

Разметка именованных сущностей в корпусах текстов, которая обычно проводится по заранее определенным семантическим категориям (тегам NE): PERSON (персона: *Александр Македонский*; *Гай Юлий Цезарь*; *Сократ* и т. д.), ORGANIZATION (организация: *Организация Объединенных Наций* (*ООН*); *Международное агентство по атомной энергии* (*МАГАТЭ*); *Межгосударственный авиационный комитет* (*МАК*) и т. д.), LOCATION (локация: *Куликово поле*; *монастырь Шаолинь*; *Великая Китайская стена* и т. д.), ADDRESS (адрес: *Петровка, 38*; *улица Лизюкова (Воронеж)*, *Бейкер-стрит, 221B (Лондон)* и т. д.), DATE (дата: *2-5 сентября 1666 г.*; *9 февраля (3 марта) 1861 г.*; *9 мая 1945 г.* и т. д.), TRADEMARK (торговая марка: *«Красный октябрь»*; *«Весёлый молочник»*; *«Зелёная линия»* и т. д.), SUBSTANCE (вещество: H_2O; *NaCl*; $CuSO_4$ и т. д.), OCCUPATION (род занятий: *лингвист*; *программист*; *переводчик* и т. д.) и другие. Дополнением процедуры выделения именованных сущностей является нормализация: *И. Е. Репин* (*И. Е. Репина, И. Е. Репину…*) = *И. Репин* (*И. Репина, И. Репину…*) = *Илья Ефимович Репин* (*Ильи Ефимовича Репина, Илье Ефимовичу Репину…*) = *Илья Репин* (*Ильи Репина, Илье Репину…*) и т. д. (Popov, Adaskina, Andreyeva et al., 2016). Усложнение задачи предполагает выделение вложенных именованных сущностей (Nested Named Entities, NNE) и установление связей между отдельными NE: *Московский университет им. М. В. Ломоносова*; *Санкт-Петербургский политехнический университет Петра Великого*; *посёлок имени Морозова* и т. д. (Loukachevitch, Artemova, Batura et al., 2021). Число и типы тегов в иерархиях именованных сущностей варьируют от десятка до нескольких сотен и отличаются в зависимости от языка, типа текстов и предметной области (Sekine, Sudo, Nobata, 2002). Так, поиск по персонам, организациям и локациям осуществляется в новостных и общественно-политических текстах, поиск по терминам, обозначающим научные понятия, по библиографическим ссылкам – в научных и учебных текстах и т. д.

Распознавание и выделение вложенных именованных сущностей является расширенным вариантом решения задачи автоматического выделения именованных сущностей. По сравнению со стандартными изолированными именованными сущностями вложенные именованные сущности требуют детализированной аннотации и дополнительного определения отношений между сущностями, что повышает сложность процедуры. Новейшие исследования на русскоязычном материале в этой области связаны с выделением вложенных именованных сущностей в наборе данных российских новостей NEREL (Loukachevitch, Artemova, Batura et al., 2021). В работе с китайскими текстами ученые уделяют особое внимание созданию специализированных алгоритмов и методов автоматического распознавания вложенных именованных сущностей в медицинских и новостных текстах (许浩亮, 李雁群, 何云琪 и др., 2018; 闫璟辉, 宗成庆, 徐金安, 2024). Тем не менее исследование вложенных именованных сущностей в китайско-русском корпусе параллельных и сопоставимых текстов политической тематики является первым в отношении рассматриваемой языковой пары.

Методы и инструменты автоматического выделения именованных сущностей

Современные технологии автоматического выделения именованных сущностей сейчас развиваются в трех основных направлениях: лингвистический подход, подход, основанный на машинном обучении, и гибридный подход (Большакова, Ефремова, 2017; Филиппова, 2017; Keraghel, Morbieu, Nadif, 2024).

Лингвистический подход предполагает проведение процедур автоматического выделения именованных сущностей с использованием правил и словарных баз данных. Правила составляются в соответствии с шаблонами лексико-грамматических конструкций, к употреблению в которых тяготеют именованные сущности (Большакова, Иванов, Сапин и др., 2016; Бабина, 2016): например, аббревиатуры *ООО*, *ОАО* и им подобные, как правило, вводят названия организаций (*ООО «ПетербургГаз»*; *ОАО «РЖД»* и т. д.), обозначения статуса лица (*господин*, *князь*, *председатель*) могут предшествовать именам собственным, соответствующим персонам (*князь Мышкин*, *зицпредседатель Фунт* и т. д.). Словари именованных сущностей составляются с учетом шаблонов на основе репрезентативных корпусов текстов целевой предметной области. Тем самым лингвистический подход обеспечивает высокую точность автоматического выделения именованных сущностей в условиях соблюдения ограничений на предметную область и тематику текстов. Недостатком правилового подхода является то, что правила и словари не всегда могут быть перенесены на другие предметные области (Zhang, Wang, 1997; Shaalan, Raza, 2009).

Подход к решению задачи автоматического выделения именованных сущностей, основанный на алгоритмах и моделях машинного обучения, оказывается более гибким по сравнению с трудозатратным и ресурсоемким лингвистическим подходом. Для автоматического выделения именованных сущностей используется обучение с учителем, без учителя и частичное обучение. Алгоритмы обучения без учителя применяются в случае отсутствия размеченных обучающих данных, при этом реализуется оценка близости и классификация

векторов синтагматически связанных групп слов – кандидатов в именованные сущности (Shinyama, Sekine, 2004; Bonnefoy, Bellot, Benoit, 2011). При выборе алгоритмов с частичным обучением (Kozareva, Bonev, Montoyo, 2005; Gao, Kotevska, Sorokine et al., 2021) допустимо использование малых наборов размеченных вручную данных и обучение семейства классификаторов для повышения качества и объема тренировочной выборки, которая пополняется за счет корректно распознанных примеров. Алгоритмы обучения с учителем могут использоваться в случае наличия репрезентативного корпуса с разметкой именованных сущностей, к которому применяются различные алгоритмы для построения классификационных моделей: скрытая марковская модель (Hidden Markov Model, HMM) (Morwal, Jahan, Chopra, 2012), метод максимума энтропии (Maximum Entropy, ME) (Borthwick, Sterling, Agichtein et al., 1998), метод условных случайных полей (Conditional Random Fields, CRF) (Shishtla, Gali, Pingali et al., 2008), машина опорных векторов (Support Vector Machine, SVM) (Yamada, Kudo, Matsumoto, 2002) и т. д. Распространение нейросетевых подходов в NLP расширило возможности решения задачи автоматического выделения именованных сущностей с помощью CNN, RNN, CNN, LSTM (Collobert, Weston, Bottou et al., 2011; Cetoli, Bragaglia, Harney et al., 2018; Huang, Li, Subudhi et al., 2020), моделей семейства Трансформер (BERT и др.) (Devlin, Chang, Lee et al., 2019), больших языковых моделей (Large Language Models, LLM) (Li, Sun, Tang et al., 2023). В прикладных разработках по автоматическому выделению именованных сущностей используются программные комплексы, среди которых более всего известны *GATE* (*General Architecture for Text Engineering*; http://www.gate.ac.uk/), *Stanford Named Entity Recognizer* (https://nlp.stanford.edu/software/CRF-NER.html) и ряд других. Эти системы были разработаны и широко применяются прежде всего для автоматического выделения именованных сущностей в корпусах англоязычных текстов, однако для русского и китайского языков они малопригодны. С учетом языково-специфичных задач и данных для автоматического выделения именованных сущностей были разработаны специализированные инструменты. Для русского языка высокие результаты показывают *Abbyy InfoExtractor SDK* (https://www.abbyy.com/flexicapture-sdk/), *Томита-парсер* (https://tech.yandex.ru/tomita/), *PullEnti SDK* (https://pullenti.ru/Download), библиотеки для Python *SpaCy* (https://spacy.io/), *Natasha* (https://github.com/natasha/natasha) и ряд других.

Парсер *Abbyy InfoExtractor SDK*, разработанный компанией *Abbyy*, позволяет выделять и связывать между собой NE, факты, события, подключать пользовательские словари и формальные онтологии. *Томита-парсер*, созданный компанией Яндекс, работает на основе *GLR*-парсера (*Generalized left-to-right algorithm*), опирающегося на правила контекстно-свободных грамматик и словарные базы данных. *Yargy*-парсер (https://github.com/natasha/yargy) в библиотеке *Natasha* также реализует правиловый подход, но с использованием средств языка программирования Python. Помимо стандартных именованных сущностей (персоны, организации, локации) *Yargy*-парсер допускает определение конструкций, связанных с именованными сущностями в текстах специализированных предметных областей (например, названия языков программирования, библиотек, инструментов в текстах по информационным технологиям и т. д.). Лингвистический процессор *PullEnti* предназначен для извлечения информации из неструктурированных официально-деловых текстов и семантического анализа. *PullEnti* задает контрольные значения качества решения задачи автоматического выделения именованных сущностей в русском языке, что подтверждено результатами соревнований *Dialogue Evaluation* (https://www.dialog-21.ru/evaluation/).

Правиловые подходы постепенно уступают дорогу методам машинного обучения: в частности, модели автоматического выделения именованных сущностей в многоязычной библиотеке *SpaCy* позволяют извлекать именованные сущности как для русских (https://spacy.io/models/ru), так и для китайских текстов (https://spacy.io/models/zh/). *SpaCy* – это многофункциональный комплекс, который предусматривает разноплановую обработку и разметку текстов, от токенизации, лемматизации до морфологического и синтаксического анализа, процедур извлечения информации из текстов. В экспериментах с русскоязычными текстами *SpaCy* обеспечивает результаты, соответствующие стандартам NLP (Соколовский, Некрасов, Землянский и др., 2023). По эффективности автоматического выделения именованных сущностей для китайского языка *SpaCy* значительно уступает библиотеке *HanLP* (https://hanlp.hankcs.com/), которая в рамках нашего исследования является предпочитаемым инструментом для этого языка. *HanLP* – это многоязычная библиотека для обработки текстов на естественном языке для производственных целей, разработанная прежде всего для китайского языка. Она позволяет осуществлять токенизацию, морфологическую разметку, автоматическое выделение именованных сущностей, синтаксический анализ зависимостей для китайских текстов и перевод упрощенных китайских иероглифов в традиционные и наоборот. *HanLP* опирается на модели универсальных зависимостей (Universal Dependencies, UD), обеспечивая процедуры автоматического выделения именованных сущностей на основе тонкой настройки моделей английского, китайского и арабского языков. Для остальных языков, в том числе и русского, качество NER оказывается ниже. Между режимами автоматического выделения именованных сущностей в *SpaCy* и *HanLP* имеются различия в распознавании именованных сущностей. Помимо стандартных именованных сущностей (персоны, организации, локации), распознаваемых с помощью *SpaCy*, *HanLP* обладает более широким спектром возможностей и может быть полезен в работе с такими типами именованных сущностей, как дата, регион, десятичная дробь, длительность, целое число, порядковое число, длина, вес, валюта и т. д. Примеры разметки именованных сущностей приведены на Рис. 1-2.

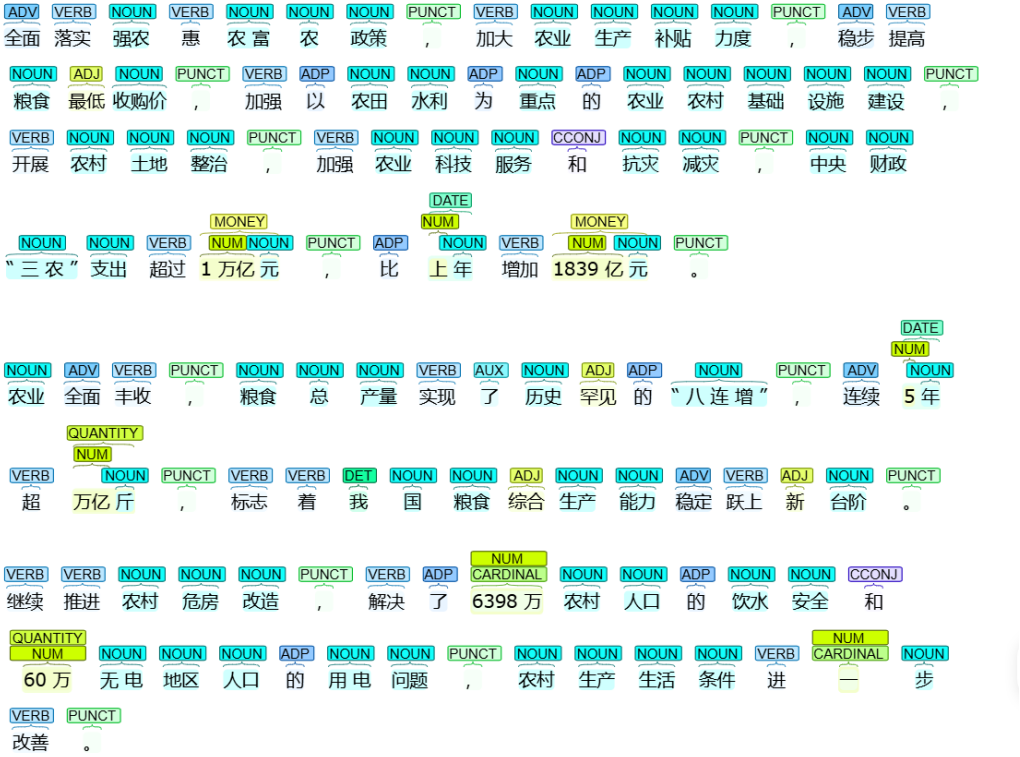

Рисунок 1. *Пример разметки именованных сущностей с помощью HanLP в китайском тексте*

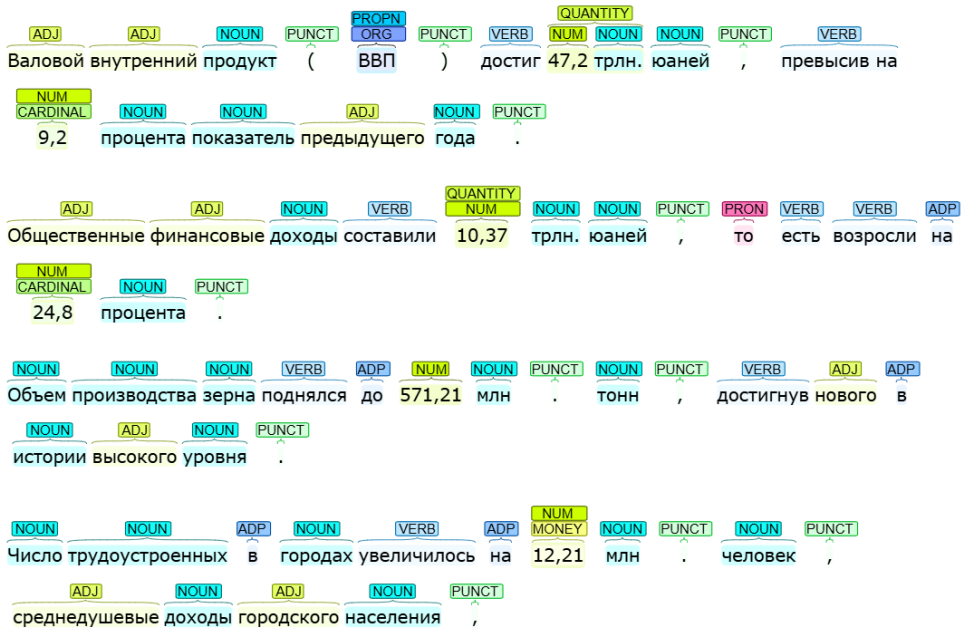

Рисунок 2. *Пример разметки именованных сущностей с помощью HanLP в русском тексте*

Описание китайско-русского корпуса параллельных и сопоставимых текстов политической тематики

В корпусной лингвистике и переводоведческих исследованиях часто используются параллельные и сопоставимые корпусы текстов (Baker, 1993, p. 233; Захаров, Богданова, 2020, с. 56-61). С конца XX века началась разработка корпусов текстов китайского языка, и до сих пор количество подобных ресурсов растет с каждым днем, включая параллельные и сопоставимые корпусы текстов (Колпачкова, 2015, с. 278): например, многожанровый русско-китайский и китайско-русский параллельный корпус (崔卫, 李峰, 2014), русско-китайский параллельный корпус научных текстов гуманитарной области (Тао, Захаров, 2015) и др. При работе с параллельными

и сопоставимыми китайско-русскими корпусами текстов особое внимание уделяется задачам, связанным с лингвистическими особенностями оригинальных и переводных политических текстов, а также с универсалиями перевода (translation universals) (王克非, 秦洪武, 2009; 李晓倩, 胡开宝, 2017; Чжу, Захаров, 2024). Однако следует отметить, что количество таких публикаций невелико, и исследование данной области требует дальнейшего углубления. Прежде всего это касается сравнительного анализа типологии именованных сущностей в китайских (Liu, Guo, Wang et al., 2022) и русских текстах (Named Entity Recognition and Fact Extraction (dialog-21.ru). 2016. https://www.dialog-21.ru/en/evaluation/2016/ner/; RuNNE. Russian Nested Named Entity Recognition Shared Task: Few-shot approach (dialog-21.ru). 2022. https://www.dialog-21.ru/en/dialogue-evaluation/competitions/dialogue-evaluation-2022/runne-2022/), результативности применения существующих методов и инструментов автоматического выделения именованных сущностей.

Эмпирическим материалом исследования является разработанный нами китайско-русский корпус параллельных и сопоставимых текстов политической тематики, состоящий из трех подкорпусов (ДРП-К, ДРП-Р и ППР). Тексты в корпусе были взяты с государственного сайта информационного агентства *Синьхуа* (https://russian.news.cn/index.htm). Подкорпус ДРП Китая содержит краткое описание работы правительства за прошедший год и план будущей работы, охватывающий все аспекты развития и строительства страны, отражающий яркие черты политических текстов на китайском языке, вследствие чего он имеет высокую исследовательскую ценность. Перевод «Докладов о работе правительства» на русский язык был опубликован Научно-исследовательским институтом истории партии и литературы при ЦК КПК, что обеспечило авторитетность переводного текста. Второй подкорпус – это сопоставимый корпус «Послания Президента Российской Федерации Федеральному Собранию РФ 2011-2021 гг.». ППР представляет собой высокую ценность как документ о государственной жизни России, поэтому является важным материалом для лингвистического исследования.

Для выделения именованных сущностей ДРП-К, ДРП-Р и ППР были сегментированы по годам, были определены объемы текстов в лексических токенах для каждого корпуса. Кроме того, для выделения именованных сущностей мы предварительно обработали китайские и русские тексты, включая токенизацию и лемматизацию. Для токенизации китайских текстов мы использовали инструмент *CorpusWordParser* (https://corpus.bfsu.edu.cn/TOOLS.htm), разработанный группой корпусной лингвистики Пекинского университета иностранных языков. При токенизации и лемматизации была задействована библиотека для морфологического анализа *SpaCy*.

Результаты применения библиотек SpaCy и HanLP

В данном исследовании мы используем *SpaCy* для автоматического выделения именованных сущностей в русских текстах (ДРП-Р и ППР) и *HanLP* для NER в китайских текстах (ДРП-К). В соответствии с целью данной работы и для анализа изменений в составе именованных сущностей политических текстов при переводе с китайского языка на русский в экспериментах рассматривались только три базовых типа именованных сущностей. Количественная информация о результатах автоматического выделения именованных сущностей в трех подкорпусах представлена на Рис. 3-5. На основании полученных результатов можно выявить особенности использования именованных сущностей: наиболее частотными типами именованных сущностей в оригинальных китайских и русских политических текстах (ДРП-К и ППР) являются названия локаций, следующие по частоте – это названия организаций, реже всего встречаются названия персон. В переводных политических текстах наблюдаются другие закономерности, в частности, частоты встречаемости локаций и организаций в переводных текстах (ДРП-Р) сближаются в период с 2020 по 2022 год. На наш взгляд, это может быть связано, с одной стороны, с качеством распознавания именованных сущностей, а с другой – с переводческими приемами.

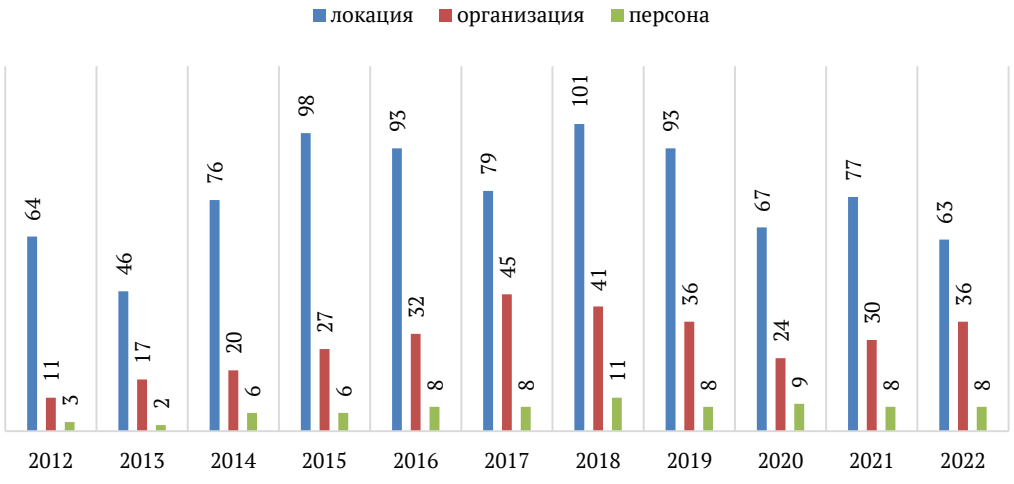

Рисунок 3. *Распределение именованных сущностей в ДРП-К по годам*

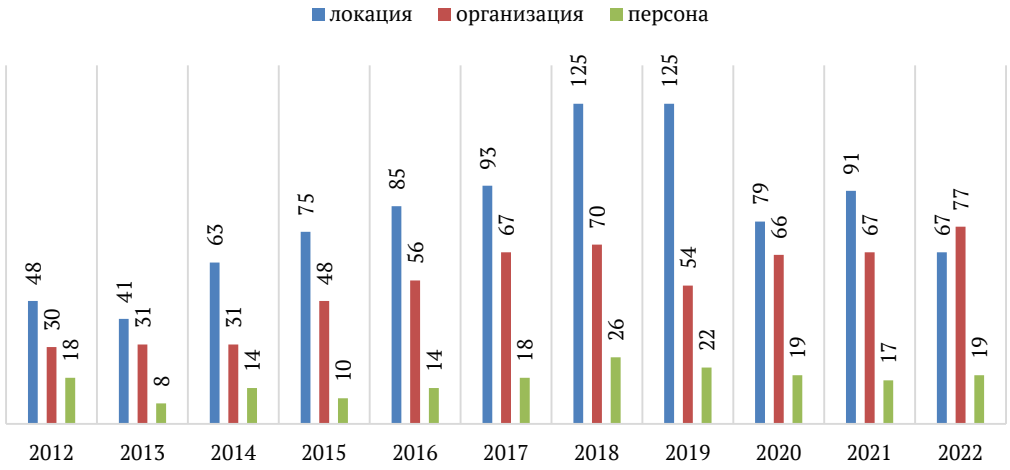

Рисунок 4. *Распределение именованных сущностей в ДРП-Р по годам*

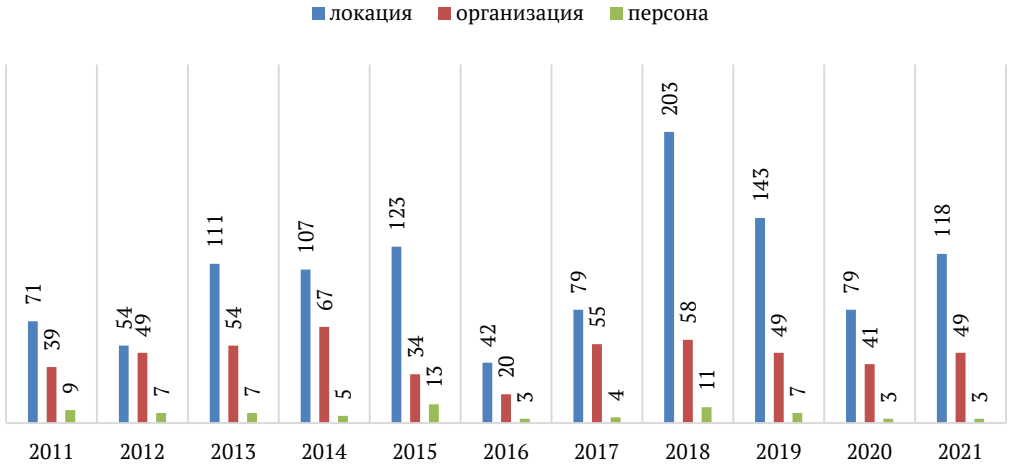

Рисунок 5. *Распределение именованных сущностей в ППР по годам*

Для дальнейшего изучения закономерностей китайско-русских переводов именованных сущностей в политических текстах и сравнения сходств и различий именованных сущностей в китайских и российских политических текстах мы выделили три типа именованных сущностей с высокой частотой встречаемости в трех корпусах, а именно: локации – 10, организации – 5 и персоны – 2, как показано в Табл. 1-3 соответственно. Сравнивая три типа высокочастотных именованных сущностей в корпусах ДРП-К и ДРП-Р, мы обнаруживаем, что они в основном соответствуют друг другу, то есть именованные сущности, часто встречающиеся в исходных текстах, в основном сохраняются и в переводных текстах, например: *香港 – Сянган*, *澳门 – Аомэнь*, *国务院 – Госсовет*, *中央 – ЦК*, *联合国 – ООН*, *习近平 – Си Цзиньпин* и т. д. Это доказывает, что переводчики чаще всего используют дословный перевод при передаче именованных сущностей с китайского языка на русский в политических текстах. Однако в то же время мы видим, что некоторые именованные сущности, которые часто встречаются в исходном тексте, отсутствуют в русском переводе, например: 两岸, 京津冀, 中华 и т. д. Мы считаем, что этот результат объясняется тем, что переводчики использовали разные варианты при передаче одной и той же именованной сущности в исходном тексте, чтобы способствовать пониманию читателем, например, 两岸 в исходном тексте может быть переведено и как «обе стороны Тайваньского пролива», и как «два берега»; 京津冀 может быть передано как «Пекин, Тяньцзинь и Хэбэй» и «Бохайский залив»; 中华 выражает практически то же значение, что и 中国, поэтому в русских переводах оно передается как «Китай».

В ходе анализа распределения именованных сущностей по типам и годам были замечены следующие закономерности в использовании именованных сущностей в ДРП-Р и ППР. В данных подкорпусах проявились сходства в использовании именованных сущностей: локации *Китай* и *Россия*, несомненно, являются наиболее важными и высокочастотными в политических текстах. Кроме того, среди локаций в китайских и российских политических текстах частотны именованные сущности, обозначающие столицы, города центрального подчинения, провинции и субъекты федерации. Из названий организаций чаще всего употребляются названия центральных политических институтов двух стран. Например, *Госсовет*, *ЦК*, *ВСНП*, *НПКС* (Китай); *Госдума*, *Совет Федерации*, *МВД*, *Федеральное Собрание* (Россия). За ними следуют международные, межрегиональные и межправительственные организации, такие как *ООН*, *Евразийский экономический союз*, *Шанхайская организация сотрудничества*,

БРИКС, *АСЕАН* и др. В категории именованных сущностей, соответствующих персонам, наиболее частотными являются имена глав государств. Результаты выделения именованных сущностей из двух подкорпусов также демонстрируют ряд отличий. Например, названия стран мира различаются по частоте, что может быть связано с целями и приоритетами развития России и Китая. Кроме того, в категории персон в корпусе ДРП-Р редко встречаются имена лиц, не являющихся политическими деятелями, в то время как в ППР упоминаются такие известные личности, как Ломоносов и Толстой, что может коррелировать со стилем речи говорящих.

Таблица 1. *Высокочастотные именованные сущности в ДРП-К по годам*

Год	Локация	Организация	Персона
2012	中国 (Китай), 两岸 (обе стороны Тайваньского пролива), 澳门 (Аомэнь), 香港 (Сянган), 东北 (Северо-Восток Китая), 上海 (Шанхай), 台湾 (Тайвань), 北京 (Пекин), 滨海新区 (Новый район Биньхай), 甘肃 (Ганьсу)	国务院 (Госсовет), 党中央 (ЦК), 中国共产党 (КПК), 十八大 (XVIII съезд), 全国政协 (ВК НПКСК)	胡锦涛 (Ху Цзиньтао), 邓小平 (Дэн Сяопин)
2013	中国、两岸、香港、澳门、舟曲 (Чжоуцюй), 玉树 (Юйшу), 汶川 (Вэньчуань), 台 (Тайвань), 上海、新疆 (Синьцзян)	国务院、十八大、党中央、残奥会 (Паралимаиада), 教育部 (Минобразования)	邓小平、习近平 (Си Цзиньпин)
2014	中国、澳门、香港、两岸、中华 (Китай), 上海、台、澳大利亚 (Австралия), 巴 (Бразилия), 冰岛 (Биндао)	国务院、党中央、亚太经合组织 (АТЭС), 东盟 (АСЕАН), 二十国集团 (Большая двадцатка)	邓小平、习近平
2015	中国、澳门、两岸、香港、韩 (Республика Корея), 丝绸之路 (Шелковый путь), 台湾、福建 (Фуцзянь), 广东 (Гуандун), 京津冀 (Пекин, Тяньцзинь и Хэбэй)	国务院、党中央、人民政协 (НПКСК), 东盟、二十国集团	习近平、邓小平
2016	中国、两岸、澳门、香港、台湾、东北、韩、京津冀、欧 (Европа), 长江 (Янцзы)	党中央、国务院、二十国集团、联合国 (ООН), 十八大	习近平、诺贝尔
2017	中国、澳门、香港、台湾、两岸、长江、中华、上海、巴黎 (Париж), 北京	党中央、国务院、人大、政协、杭州峰会 (Ханчжоуский саммит)	习近平、邓小平
2018	中国、澳门、两岸、香港、台湾、中华、北京、大陆 (материковый Китай), 京津冀、上海	党中央、国务院、十九大 (XIX съезд), 全国人大常委会 (ПК ВСНП), 博鳌 (Боао)	习近平、毛泽东 (Мао Цзэдун)
2019	中国、澳门、两岸、香港、京津冀、台湾、中美 (Китай и Америка), 北京、非 (Африка), 海南 (Хайнань)	党中央、国务院、残奥会、人民政协、十九大	习近平
2020	中国、台湾、澳门、湖北 (Хубэй), 两岸、香港、中华民族 (Китайская нация), 韩、日 (Япония), 武汉 (Ухань)	党中央、国务院、北京冬奥会 (Зимние олимпийские игры в Пекине), 二十国集团、金砖 (БРИКС)	习近平
2021	中国、澳门、台湾、香港、长江、中华、海南、黄河 (Хуанхэ), 欧、嫦娥五号 (Чанъэ-5)	党中央、国务院、人民政协、中国共产党、联合国	习近平
2022	中国、澳门、香港、中华、台湾、东北、海南、黄河、两岸、长江	党中央、二十大、国务院、人民政协、十九大	习近平

Таблица 2. *Высокочастотные именованные сущности в ДРП-Р по годам*

Год	Локация	Организация	Персона
2012	Китай, Сянган, Аомэнь, США, Тайвань, Пекин, Тяньцзинь, Шанхай, Юйшу, Гуандун	Госсовет, ГЭС, КПК, ЦК, правительство	Ху Цзиньтао, Дэн Сяопин
2013	Китай, Аомэнь, Пекин, США, Вэньчуань, Сянган, Тайванем, Шанхай, Гуанчжоу, Далянь	Госсовет, ВСНП, КПК, НИОКР, ЦК	Дэн Сяопин, Си Цзиньпин
2014	Китай, Аомэнь, Сянган, Пекин, США, Чжуцзян, Шанхай, Янцзы, Австралия, Бохайский залив	ЦК, Госсовет, АТЭС, КПК, АСЕАН,	Дэн Сяопин, Си Цзиньпин
2015	Китай, Аомэнь, США, Сянган, Шанхай, Швейцария, Шелковый путь, Шэньчжэньский, Янцзы, Гуандун	Госсовет, ЦК, КПК, НПКС, СНП	Дэн Сяопин, Си Цзиньпин
2016	Китай, Сянган, Аомэнь, Пекин, США, Тайвань, Тяньцзинь, ЕС, Хэбэй, Янцзы	КПК, ЦК, Госсовет, Большая двадцатка, НПКС, ООН	Си Цзиньпин, Ту Юю
2017	Китай, Сянган, Тайвань, Аомэнь, Пекин, Янцзы, Мировой океан, Автономный район, Азиатско-Тихоокеанский регион, Внутренняя Монголия	КПК, ЦК, Госсовет, НПКС, СНП	Си Цзиньпин, Дэн Сяопин
2018	Китай, Сянган, Аомэнь, Тайвань, Мировой океан, Пекин, Шанхай, Северо-Восток, Янцзы, Ханчжоу	КПК, ЦК, Госсовет, НИИ, ОАР	Си Цзиньпин, Мао Цзэдун
2019	Китай, Аомэнь, Пекин, США, Сянган, Янцзы, Сычуань, Тяньцзинь, Хэбэй, Тибет	КПК, ЦК, Госсовет, ЦК партии, ШОС	Си Цзиньпин
2020	Китай, Аомэнь, Сянган, Хубэй, Тайвань, Ухань, Пекин, Республика Корея, Янцзы, Япония	КПК, ЦК, НОАК, Госсовет, БРИКС	Си Цзиньпин
2021	Китай, Аомэнь, Сянган, Янцзы, ЕС, Пекин, Тайваньский пролив, Хуанхэ, Африка, Чунцин	КПК, ЦК, Госсовет, ВСНП, АТЭС	Си Цзиньпин
2022	Китай, Пекин, Аомэнь, Янцзы, Сянган, Тайвань, Хуанхэ, Африка, Большой залив, Восточная Азия	КПК, ЦК, Госсовет, Хайнаньский порт свободной торговли, ВСНП	Си Цзиньпин

Таблица 3. *Высокочастотные именованные сущности в ППР по годам*

Год	Локация	Организация	Персона
2011	Россия, Франция, Евросоюз, Китай, Индия, Лиссабон, Москва, Америка, Азиатско-тихоокеанский регион, Африка	Госдума, Вооружённые Силы, Совет Федерации, ЕврАзЭС, МВД	Дмитрий Анатольевич Медведев, Юрий Гагарин
2012	Россия, Федерация, Беларусь, Евросоюз, Казахстан, Российское государство, Азиатско-тихоокеанский регион, Евроатлантический регион, Москва, Южная Осетия	Госдума, Вооружённые Силы, Евразийский экономический союз, Таможенный союз, ВТО	Владимир Владимирович Путин, Лао-Цзы
2013	Россия, Федерация, Дальний Восток, Сибирь, Азиатско-Тихоокеанский регион, Азия, Америка, Арктика, Москва, Европа	Госдума, Правительство, СМИ, СНГ, Совет Федерации	Лев Гумилёв, Ломоносов
2014	Россия, Дальний Восток, Российское государство, Сирия, Федерация, Восточная Сибирь, Иран, Сибирь, СНГ, Сочи	Правительство, Федеральное Собрание, Вооружённые Силы, Общественная палата, ОМС	Владимир Владимирович Путин, Николой Бердяев
2015	Россия, Украина, Крым, Севастополь, США, Европа, Евросоюз, Федерация, Дальний Восток, Херсонес	Банк России, ООН, Агентство стратегических инициатив, Министерство обороны, Минфин	Иван Ильин, Янукович
2016	Россия, Дальний Восток, Комсомольск-на-Амуре, Владивосток, Крым, Севастополь, Сочи, Азиатско-Тихоокеанский регион, Азово-Черноморский, Вьетнам	Правительство, НКО, Евразийский экономический союз, ОМС, Центральный банк	Дмитрий Анатольевич Медведев, Дмитрий Иванович Менделеев
2017	Россия, США, Дальний Восток, Евросоюз, Австрия, Италия, Португалия, Байкал, Волга, Германия	Госдума, Центральный банк, АПК, ВЭБ, Общероссийский народный фронт	Алексей Фёдорович Лосев, Лихачёв,
2018	Россия, США, СССР, Арктика, Владивосток, Дальний Восток, Казань, Москва, Азия, Китай	Вооружённые Силы, Минобороны, АПК, НАТО, Банк России	Владимир Владимирович Путин, Роман Филипов
2019	Россия, США, Федерация, Москва, Дальний Восток, Европа, Белгородская область, Новгородский, Япония, Ленинградская область	Правительство, Центральный Банк, Госдума, Верховный Суд, МВД	Александр Сергеевич Грибоедов, Владимир Владимирович Путин
2020	Россия, Федерация, Дальний Восток, Советский Союз, Австрия, Ближний Восток, Мировой океан, Северная Африка, Северный Кавказ, Швейцария	Правительство, Государственная Дума, Совет Федерации, Федеральное Собрание, Госсовет	Владимир Владимирович Путин
2021	Россия, Федерация, Белоруссия, Камчатка, Красноярск, Кузбасс, Москва, Украина, Азия, Санкт-Петербург	Госсовет, Вооружённые Силы, Единая Россия, Правительство, ВЭБ	Владимир Владимирович Путин

Обе использованные нами библиотеки *SpaCy* и *HanLP* в ряде случаев производили ошибочную разметку именованных сущностей. При распознавании именованных сущностей в китайских политических текстах (ДРП-К) *HanLP* допускает следующие типы ошибок: 1) смешение локаций и персон: например, 古田 (*Гутянь*), уезд провинции *Фуцзянь* был распознан *HanLP* как имя человека; 2) идентификация синтаксических групп, содержащих дериваты названий локаций, как именованных сущностей со значением локации или организации, например, термин **中国**梦 (***Китайская*** *мечта*) был идентифицирован как локация, а термин **西**电**东**送 (*переброска электроэнергии с* ***запада*** *на* ***восток***) – как организация; 3) при разметке организаций *HanLP* в некоторых случаях идентифицирует именованные сущности не полностью, а частично: например, теги организаций получили группы 离岸人民币 (*оффшорный жэньминьби/юань*) и 人民币海外 (*заграница в китайских юанях*), однако они сами по себе не представляют организации с точки зрения автоматического выделения именованных сущностей. При анализе расширенного контекста оказалось, что указанные словосочетания включены в состав следующих именованных сущностей: *Центр оффшорных операций в жэньминьби* и *Фонд сотрудничества с заграницей в китайских юанях*, которые, несомненно, являются локациями. В то же время *SpaCy* демонстрирует разнородные результаты при разметке именованных сущностей в русскоязычных корпусах ДРП-Р и ППР. Первые два типа ошибок, зарегистрированных нами при анализе выдачи *HanLP*, были обнаружены и в ходе использования *SpaCy* для русскоязычных текстов, являющихся переводами с китайского: например, *SpaCy* иногда распознает именованные сущности *Аомэнь* и *Сянган* как локации или как персоны, *Олимпийские игры* – как локацию, в то же время *Здоровый Китай* и *Спокойный Китай* были идентифицированы как организации. Кроме того, модель автоматического выделения именованных сущностей в *SpaCy*

очень чувствительна к капитализации: например, словоформа *Зачитан* идентифицируется как локация, *ОБЩИЙ*, *ОБЗОР* – как организации, *Бурно*, *Реформирована* – как персонал. Тем не менее при проведении процедур автоматического выделения именованных сущностей в оригинальных русскоязычных политических текстах подкорпуса ППР *SpaCy* демонстрирует большую стабильность. В данном случае основные ошибки связаны с идентификацией локаций и организаций, а также с некорректной разметкой локаций по аналогии с обработкой подкорпуса ДРП-Р. В данном подкорпусе *SpaCy* также размечает словоформы *Добавлю*, *Культура*, *Знание* и т. д. как именованные сущности. Стоит отметить, что *SpaCy* допускает особый тип ошибки при распознавании именованных сущностей в ППР, размечая как персоны некоторые словоформы, в частности *Президент*, *Послание*, *Правительство*, *Генеральный* и др.

На основе результатов выделения именованных сущностей в корпусах ДРП-К, ДРП-Р и ППР мы попытаемся обобщить встречающиеся типы вложенных именованных сущностей в китайских и российских политических текстах (Табл. 4-5).

Таблица 4. *Типы вложенных именованных сущностей в китайских политических текстах*

NE	Разметка	Пример
ЛОКАЦИЯ	[[локация]ЛОКАЦИЯ]	珠三角 (дельта Жемчужной реки), 亚欧大陆桥 (евразийский континентальный мост), 天津港 (порт Тяньцзинь), 港珠澳大桥 (мост Гонконг-Чжухай-Макао), 亚太自贸区 (Азиатско-Тихоокеанская зона свободной торговли), 川藏铁路 (Сычуань-Тибетская железная дорога)
ОРГАНИЗАЦИЯ	[[локация]ОРГАНИЗАЦИЯ]	全国政协 (Всекитайский комитет Народного политического консультативного совета Китая), 海湾合作委员会 (Совет сотрудничества арабских государств Персидского залива), 全国人大常委会 (Постоянный комитет Всекитайского собрания народных представителей), 上海合作组织 (Шанхайская организация сотрудничества), 亚太经合组织 (Организация азиатско-тихоокеанского экономического сотрудничества), 丝路基金 (фонд Шёлкового пути), 亚洲基础设施投资银行 (Азиатский банк инфраструктурных инвестиций), 杭州峰会 (Ханчжоуский саммит), 上海期货交易所 (Шанхайская фьючерсная биржа)
ОРГАНИЗАЦИЯ	[[цифра]ОРГАНИЗАЦИЯ]	十八大 (XVIII съезд), 十六大 (XVI съезд), 第十一届全国人民代表大会 (Всекитайское собрание народных представителей 11-го созыва), 二十国集团 (Большая двадцатка)
ОБЪЕКТ	[[локация]ОБЪЕКТ]	辽宁舰 (авианосец Ляонин), 嫦娥四号 (спутник зондирования Луны Чанъэ-4)
ПРОЕКТ	[[локация]ПРОЕКТ]	西气东输 (переброска газа с запада на восток), 西电东送 (переброска электроэнергии с запада на восток)

Таблица 5. *Типы вложенных именованных сущностей в российских политических текстах*

NE	Разметка	Пример
ЛОКАЦИЯ	[[локация]ЛОКАЦИЯ]	Латинская Америка, Евроатлантический регион, Азово-Черноморский бассейн, Кавказское побережье
ОРГАНИЗАЦИЯ	[[локация]ОРГАНИЗАЦИЯ]	Европейский союз, АТЭС, ЕврАзЭС, Шанхайская организация сотрудничества, Евроцентробанк, Российский научный фонд, Российская академия наук, Российский экспортный центр, Южный военный округ, Общероссийский народный фронт
ОБЪЕКТ	[[персона]ОБЪЕКТ]	Семёновский полк

Согласно полученным данным, в китайских политических текстах в основном встречаются следующие пять типов вложенных именованных сущностей: [[локация]ЛОКАЦИЯ], [[локация]ОРГАНИЗАЦИЯ], [[цифра]ОРГАНИЗАЦИЯ], [[локация]ОБЪЕКТ], [[локация]ПРОЕКТ], а в российских политических текстах появляются три типа NNE: [[локация]ЛОКАЦИЯ], [[локация]ОРГАНИЗАЦИЯ], [[персона]ОБЪЕКТ]. Различия в частотности типов вложенных именованных сущностей, на наш взгляд, объясняются лингвистическими особенностями китайского языка. В то же время и в китайских, и в русских политических текстах мы видим одни и те же типы вложенных именованных сущностей, их наиболее распространенным типом в политических текстах является [[локация]ОРГАНИЗАЦИЯ], чаще всего вложенной именованной сущностью выступает название континентов и стран, а сами вложенные именованные сущности соответствуют названиям международных и межрегиональных организаций. Кроме того, в политических текстах тип вложенных именованных сущностей [[локация]ЛОКАЦИЯ] также достаточно частотен, как правило, это именованная сущность, обозначающая небольшую локацию, вложенная в именованную сущность, обозначающую более крупную локацию.

Заключение

В результате нашего исследования были получены следующие результаты:

1. Созданный в рамках проекта китайско-русский корпус параллельных и сопоставимых текстов политической тематики был проанализирован с точки зрения употребляемых в нем именованных сущностей, были проведены процедуры предобработки для проведения экспериментов с помощью специализированных инструментов.

2. Была сформирована теоретическая база исследования: изучены подходы к описанию именованных сущностей, описана типология именованных сущностей, рассмотрены алгоритмы и инструменты для автоматического выделения сущностей.

3. Выбор инструментов *SpaCy* и *HanLP* был произведен на основании того, что библиотека *SpaCy* демонстрирует высокую способность распознавать стандартные типы именованных сущностей в русских текстах, в то время как *HanLP* позволяет анализировать более широкий спектр типов именованных сущностей в китайских текстах.

4. Эксперименты дали возможность выявить следующую закономерность: среди стандартных NE в оригинальных китайских и русских политических текстах наибольшую частоту встречаемости имеет тип локации, за которым следуют организация и персона.

5. Эксперименты показали, что использование двух библиотек *SpaCy* и *HanLP* для автоматического выделения сущностей приводит к разноплановым ошибкам. В нашем исследовании дано объяснение типов ошибок и проведено их обобщение, что дает идеи для будущего совершенствования алгоритмов.

6. В данной работе проанализированы типы вложенных именованных сущностей в китайских и русских политических текстах. Это особенно важно, поскольку сравнительные корпусные исследования распознавания вложенных именованных сущностей в китайских и русских политических текстах относительно редки. Обнаружено, что типы вложенных именованных сущностей в китайских политических текстах более разнообразны и сложны, что может объясняться лингвистическими характеристиками китайского языка.

Перспективы развития исследования связаны с необходимостью совершенствования алгоритмов автоматического выделения стандартных и именованных сущностей в китайских и русских корпусах политических текстов. Планируется расширение эмпирических данных о стандартных и вложенных именованных сущностях для изучения стратегий их переводов в китайском и русском языках. Полученные результаты позволяют расширить рамки исследования и оценить вклад именованных сущностей в процедуры извлечения информации различных типов из текстов (мнения, факты, эмоции и т. д.). Решение указанных задач требует пополнения корпуса китайских и русских политических текстов.

Источники | References

1. Бабина О. И. Именованные сущности в корпусе текстов новостных сообщений: лингвистическое описание // Наука ЮУрГУ: материалы 68-й научной конференции / Министерство образования и науки Российской Федерации; Южно-Уральский государственный университет. Челябинск, 2016.
2. Большакова Е. И., Ефремова Н. Э. Извлечение информации из текстов: портрет направления // Большакова Е. И., Воронцов К. В., Ефремова Н. Э., Клышинский Э. С., Лукашевич Н. В., Сапин А. С. Автоматическая обработка текстов на естественном языке и анализ данных. М., 2017.
3. Большакова Е. И., Иванов К. М., Сапин А. С., Шариков Е. Ф. Система для извлечения информации из текстов на базе лексико-синтаксических шаблонов // Пятнадцатая национальная конференция по искусственному интеллекту с международным участием КИИ-2016: труды конференции: в 3 т. Смоленск: Универсум, 2016. Т. 1.
4. Брыкина М. М., Файнвейц А. В., Толдова С. Ю. Извлечение и идентификация именованных сущностей с использованием словарей в русском языке // Актуальные инновационные исследования: наука и практика. 2013. № 1.
5. Захаров В. П., Богданова С. Ю. Корпусная лингвистика. СПб., 2020.
6. Колпачкова Е. Н. Корпусы китайского языка: современное состояние и основные проблемы // Корпусная лингвистика – 2015: труды международной конференции. СПб., 2015.
7. Соколовский Д. Е., Некрасов В. Н., Землянский С. А., Аксёнов С. В. Оценка использования инструментов библиотеки SpaCy и DeepPavlov для задачи извлечения именованных сущностей из описаний результатов осмотров пациентов с COVID-19 // Известия Томского политехнического университета. Промышленная кибернетика. 2023. № 2.
8. Сталтмане В. Э. Ономастическая лексикография. М.: Наука, 1989.
9. Суперанская А. В. Общая теория имени собственного. М.: Наука, 1973.
10. Тао Ю., Захаров В. П. Разработка и использование параллельного корпуса русского и китайского языков // Научно-техническая информация. Серия 2: Информационные процессы и системы. 2015. № 4.
11. Филиппова Е. А. Извлечение информации // Прикладная и компьютерная лингвистика / под ред. И. С. Николаева, О. В. Митрениной, Т. М. Ландо. М.: Ленанд, 2017.
12. Чжу Х., Захаров В. П. Корпусное сравнение языка китайских и российских политических текстов // Политическая лингвистика. 2024. № 1.
13. Au T. W. T., Lampos V., Cox I. J. E-NER – an Annotated Named Entity Recognition Corpus of Legal Text // arXiv. 2022. https://doi.org/10.48550/arXiv.2212.09306
14. Baker M. Corpus Linguistics and Translation Studies: Implications and Applications // Text and Technology: In Honour of John Sinclair / ed. by M. Baker, G. Francis, E. Tognini-Bonelli. Amsterdam: John Benjamins, 1993.
15. Bonnefoy L., Bellot P., Benoit M. Mesure Non-Supervisée du Degré d'Appartenance d'une Entité à un Type // TALN 2011 (Montpellier, 27 juin – 1er juillet 2011). Montpellier, 2011.
16. Borthwick A., Sterling J., Agichtein E., Grishman R. NYU: Description of the MENE Named Entity System as Used in MUC-7 // Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, 1998. Fairfax, 1998.

17. Cetoli A., Bragaglia S., Harney A. D., Sloan M. Graph Convolutional Networks for Named Entity Recognition // arXiv. 2018. https://doi.org/10.48550/arXiv.1709.10053

18. Collobert R., Weston J., Bottou L., Karlen M., Kavukcuoglu K., Kuksa P. Natural Language Processing (Almost) from Scratch // Journal of Machine Learning Research. 2011. Vol. 12.

19. Devlin J., Chang M., Lee K., Toutanova K. Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding // arXiv. 2019. https://doi.org/10.48550/arXiv.1810.04805

20. Gao S., Kotevska O., Sorokine A., Christian J. B. A Pre-Training and Self-Training Approach for Biomedical Named Entity Recognition // PloS One. 2021. Vol. 2.

21. Grishman R., Sundheim B. Message Understanding Conference – 6: A Brief History // Proceedings of the 16th International Conference on Computational Linguistics. Copenhagen, 1996.

22. Huang J., Li C., Subudhi K., Jose D., Balakrishnan Sh., Chen W., Peng B., Gao J., Han J. Few-Shot Named Entity Recognition: A Comprehensive Study // arXiv. 2020. https://doi.org/10.48550/arXiv.2012.14978

23. Keraghel I., Morbieu S., Nadif M. A Survey on Recent Advances in Named Entity Recognition // arXiv. 2024. https://doi.org/10.48550/arXiv.2401.10825

24. Kozareva Z., Bonev B., Montoyo A. Self-Training and Co-Training Applied to Spanish Named Entity Recognition // Mexican International Conference on Artificial Intelligence. Monterrey: Springer, 2005.

25. Li Ch., Sun A., Weng J., He Q. Tweet Segmentation and Its Application to Named Entity Recognition // IEEE Transactions on Knowledge and Data Engineering. 2014. Vol. 27 (2).

26. Li J., Sun A., Han J., Li Ch. A Survey on Deep Learning for Named Entity Recognition // IEEE Transactions on Knowledge and Data Engineering. 2020. Vol. 34 (1).

27. Alvarado J. C. S., Verspoor K., Baldwin T. Domain Adaption of Named Entity Recognition to Support Credit Risk Assessment // Proceedings of the Australasian Language Technology Association Workshop. Parramatta, 2015.

28. Li P., Sun T., Tang Q., Yan H., Wu Y., Huang X., Qiu X. CodeIE: Large Code Generation Models are Better Few-Shot Information Extractors // arXiv. 2023. https://doi.org/10.48550/arXiv.2305.05711

29. Liu P., Guo Y., Wang F., Li G. Chinese Named Entity Recognition: The State of the Art // Neurocomputing. 2022. Vol. 473.

30. Loukachevitch N., Artemova E., Batura T., Braslavski P., Denisov I., Ivanov V., Manandhar S., Pugachev A., Tutubalina E. NEREL: A Russian Dataset with Nested Named Entities and Relations // Proceedings of the International Conference on Recent Advances in Natural Language Processing. RANLP, 2021.

31. Luz de Araujo P. H., De Campos T. E., De Oliveira R. R. R., Stauffer M., Couto S., Bermejo P. LeNER-Br: A Dataset for Named Entity Recognition in Brazilian Legal Text // Computational Processing of the Portuguese Language. PROPOR 2018 / ed. by A. Villavicencio, V. Moreira, A. Abad. Cham: Springer, 2018. https://doi.org/10.1007/978-3-319-99722-3_32

32. Morwal S., Jahan N., Chopra D. Named Entity Recognition Using Hidden Markov Model (HMM) // International Journal on Natural Language Computing. 2012. Vol. 1.

33. Nadeau D., Sekine S. A Survey of Named Entity Recognition and Classification // Lingvisticae Investigationes. 2007. Vol. 30. Iss. 1.

34. Popov A. M., Adaskina Yu. V., Andreyeva D. A., Charabet Ja., Moskvina A. D., Protopopova E. V., Yushina T. A. Named Entity Normalization for Fact Extraction Task // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2016”. Moscow, 2016.

35. Sekine S., Sudo K., Nobata C. Extended Named Entity Hierarchy // International Conference on Language Resources and Evaluation. Las Palmas, 2002.

36. Shaalan K., Raza H. NERA: Named Entity Recognition for Arabic // Journal of the American Society for Information Science and Technology. 2009. Vol. 8.

37. Shinyama Y., Sekine S. Named Entity Discovery Using Comparable News Articles // COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics, Switzerland. Geneva, 2004.

38. Shishtla P. M., Gali K., Pingali P., Varma V. Experiments in Telugu NER: A Conditional Random Field Approach // Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages. Hyderabad, 2008.

39. Tran V. C., Hwang D., Jung J. J. Semi-Supervised Approach Based on Cooccurrence Coefficient for Named Entity Recognition on Twitter // 2nd National Foundation for Science and Technology Development Conference on Information and Computer Science (NICS). Ho Chi Minh City, 2015.

40. Wang X., Yang Ch., Guan R. A Comparative Study for Biomedical Named Entity Recognition // International Journal of Machine Learning and Cybernetics. 2018. Vol. 9 (3).

41. Yamada H., Kudo T., Matsumoto Y. Japanese Named Entity Extraction Using Support Vector Machine // Transactions of IPSJ. 2002. Vol. 43. Iss. 1.

42. Zhang X., Wang L. Identification and Analysis of Chinese Organization and Institution Names // Journal of Chinese Information Processing. 1997. Vol. 4.

43. Zhang Y., Zhang H. 2022. FinBERT-MRC: Financial Named Entity Recognition Using BERT under the Machine Reading Comprehension Paradigm // arXiv. 2022. https://doi.org/10.48550/arXiv.2205.15485

44. Zhou G., Zhang J., Su J., Shen D., Tan Ch. L. Recognizing Names in Biomedical Texts: A Machine Learning Approach // Bioinformatics. 2004. Vol. 20 (7).

45. 崔卫, 李峰. 俄汉-汉俄平行语料库的构建设想与应用展望 // 中国俄语教学. 2014. № 1 (Цуй В., Ли Ф. Концепция построения и перспективы применения русско-китайского параллельного корпуса // Преподавание русского языка в Китае. 2014. № 1).
46. 李晓倩, 胡开宝. 中国政府工作报告英译文中主题词及其搭配研究 // 中国外语. 2017. № 6 (Ли С., Ху К. Исследование ключевых слов и их сочетаний в английских переводах «Докладов о работе правительства Китая» // Иностранные языки в Китае. 2017. № 6).
47. 王克非, 秦洪武. 英译汉语言特征探讨——基于对应语料库的宏观分析 // 外语学刊. 2009. № 1 (Ван К., Цинь Х. Исследование лингвистических особенностей перевода с английского на китайский – макроанализ на основе корпуса // Журнал иностранных языков. 2009. № 1).
48. 许浩亮, 李雁群, 何云琪, 钱龙华. 中文嵌套命名实体关系抽取研究 // 北京大学学报(自然科学版). 2018. № 4 (Сюй Х., Ли Я., Хэ Ю., Цянь Л. Исследование извлечения связей между вложенными именованными сущностями на китайском языке // Журнал Пекинского университета (естественнонаучное издание). 2018. № 4).
49. 闫璟辉, 宗成庆, 徐金安. 中文医疗文本中的嵌套实体识别方法 // 软件学报. 2024. № 6 (Янь Ц., Цзун Ч., Сюй Ц. Метод распознавания вложенных сущностей в китайских медицинских текстах // Журнал о программном обеспечении. 2024. № 6).

Финансирование | Funding

RU Публикация подготовлена в рамках проекта № 202307130002, утвержденного Советом по стипендиям Министерства образования Китая, при поддержке СПбГУ, шифр проекта 124032900006-1.

EN The publication was prepared within the framework of project No. 202307130002, approved by the Scholarship Council of the Ministry of Education of China, with the support of the Saint Petersburg State University, project code 124032900006-1.

Информация об авторах | Author information

RU **Чжу Хуэй**[1]
Митрофанова Ольга Александровна[2], к. филол. н., доц.
[1] Даляньский университет иностранных языков, Китайская Народная Республика
[2] Санкт-Петербургский государственный университет

EN **Hui Zhu**[1]
Olga Aleksandrovna Mitrofanova[2], PhD
[1] Dalian University of Foreign Languages, The People's Republic of China
[2] Saint Petersburg State University

[1] *zhuhui1230@qq.com,* [2] *o.mitrofanova@spbu.ru*

Информация о статье | About this article

Дата поступления рукописи (received): 24.07.2024; опубликовано online (published online): 04.09.2024.

Ключевые слова (keywords): распознавание именованных сущностей; вложенные именованные сущности; корпус текстов; параллельный корпус; политические тексты; named entity recognition; nested named entities; text corpus; parallel corpus; political texts.