MODELING AND ANALYSIS OF INFORMATION SYSTEMS, VOL. 30, NO. 1, 2023
journal homepage: www.mais-journal.ru

THEORY OF DATA

Name Entity Recognition Tasks: Technologies and Tools

N. S. Lagutina[1], A. M. Vasilyev[1], D. D. Zafievsky[1]

DOI: 10.18255/1818-1015-2023-1-64-85

[1]P. G. Demidov Yaroslavl State University, 14 Sovetskaya str., Yaroslavl 150003, Russia.

MSC2020: 68T50
Research article
Full text in Russian

Received October 21, 2022
After revision February 1, 2023
Accepted February 8, 2023

The task of named entity recognition (NER) is to identify and classify words and phrases denoting named entities, such as people, organizations, geographical names, dates, events, terms from subject areas. While searching for the best solution, researchers conduct a wide range of experiments with different technologies and input data. Comparison of the results of these experiments shows a significant discrepancy in the quality of NER and poses the problem of determining the conditions and limitations for the application of the used technologies, as well as finding new solutions. An important part in answering these questions is the systematization and analysis of current research and the publication of relevant reviews. In the field of named entity recognition, the authors of analytical articles primarily consider mathematical methods of identification and classification and do not pay attention to the specifics of the problem itself. In this survey, the field of named entity recognition is considered from the point of view of individual task categories. The authors identified five categories: the classical task of NER, NER subtasks, NER in social media, NER in domain, NER in natural language processing (NLP) tasks. For each category the authors discuss the quality of the solution, features of the methods, problems, and limitations. Information about current scientific works of each category is given in the form of a table for clarity. The review allows us to draw a number of conclusions. Deep learning methods are leading among state-of-the-art technologies. The main problems are the lack of datasets in open access, high requirements for computing resources, the lack of error analysis. A promising area of research in NER is the development of methods based on unsupervised techniques or rule-base learning. Intensively developing language models in existing NLP tools can serve as a possible basis for text preprocessing for NER methods. The article ends with a description and results of experiments with NER tools for Russian-language texts.

Keywords: natural language processing; text features; automated essay scoring; business letter

INFORMATION ABOUT THE AUTHORS

Nadezhda Stanislavona Lagutina correspondence author	orcid.org/0000-0002-6137-8643. E-mail: lagutinans@gmail.com PhD, associate professor.
Andrey Mikhaylovich Vasilyev	orcid.org/0000-0001-6672-0981. E-mail: andrey@crafted.su associate professor.
Daniil Dmitrievich Zafievsky	orcid.org/0000-0001-8266-2283. E-mail: zafievsky@mail.ru student.

Funding: This work was supported by initiative program VIP-016.

For citation: N. S. Lagutina, A. M. Vasilyev, and D. D. Zafievsky, "Name Entity Recognition Tasks: Technologies and Tools", *Modeling and analysis of information systems*, vol. 30, no. 1, pp. 64-85, 2023.

МОДЕЛИРОВАНИЕ И АНАЛИЗ ИНФОРМАЦИОННЫХ СИСТЕМ, ТОМ 30, № 1, 2023
сайт журнала: www.mais-journal.ru

THEORY OF DATA

Задачи в области распознавания именованных сущностей: технологии и инструменты

Н. С. Лагутина[1], А. М. Васильев[1], Д. Д. Зафиевский[1]

DOI: 10.18255/1818-1015-2023-1-64-85

[1]Ярославский государственный университет им. П. Г. Демидова, ул. Советская, д. 14, г. Ярославль, 150003 Россия.

УДК 004.912
Научная статья
Полный текст на русском языке

Получена 21 октября 2022 г.
После доработки 1 февраля 2023 г.
Принята к публикации 8 февраля 2023 г.

Задача распознавания именованных сущностей (named entity recognition, NER) состоит в выделении и классификации слов и словосочетаний, обозначающих именованные объекты, таких как люди, организации, географические названия, даты, события, обозначения терминов предметных областей. В поисках лучшего решения исследователи проводят широкий спектр экспериментов с разными технологиями и исходными данными. Сравнение результатов этих экспериментов показывает значительное расхождение качества NER и ставит проблему определения условий и границ применения используемых технологий, а также поиска новых путей решения. Важным звеном в ответах на эти вопросы является систематизация и анализ актуальных исследований и публикация соответствующих обзоров. В области распознавания именованных сущностей авторы аналитических статей в первую очередь рассматривают математические методы выделения и классификации и не уделяют внимание специфике самой задачи. В предлагаемом обзоре область распознавания именованных сущностей рассмотрена с точки зрения отдельных категорий задач. Авторы выделили пять категорий: классическая задача NER, подзадачи NER, NER в социальных сетях, NER в предметных областях, NER в задачах обработки естественного языка (natural language processing, NLP). Для каждой категории обсуждается качество решения, особенности методов, проблемы и ограничения. Информация об актуальных научных работах каждой категории для наглядности приводится в виде таблицы, содержащей информацию об исследованиях: ссылку на работу, язык использованного корпуса текстов и его название, базовый метод решения задачи, оценку качества решения в виде стандартной статистической характеристики F-меры, которая является средним гармоническим между точностью и полнотой решения. Обзор позволяет сделать ряд выводов. В качестве базовых технологий лидируют методы глубокого обучения. Основными проблемами являются дефицит эталонных наборов данных, высокие требования к вычислительным ресурсам, отсутствие анализа ошибок. Перспективным направлением исследований в области NER является развитие методов на основе обучения без учителя или на основе правил. Возможной базой предобработки текста для таких методов могут служить интенсивно развивающиеся модели языков в существующих инструментах NLP. Завершают статью описание и результаты экспериментов с инструментами NER для русскоязычных текстов.

Ключевые слова: распознавание именованных сущностей; автоматическая обработка текста; обзор; инструменты обработки естественного языка

ИНФОРМАЦИЯ ОБ АВТОРАХ

Надежда Станиславовна Лагутина
автор для корреспонденции
orcid.org/0000-0002-6137-8643. E-mail: lagutinans@gmail.com
канд. физ.-мат. наук, доцент.

Андрей Михайлович Васильев
orcid.org/0000-0001-6672-0981. E-mail: andrey@crafted.su
доцент.

Даниил Дмитриевич Зафиевский
orcid.org/0000-0001-8266-2283. E-mail: zafievsky@mail.ru
студент.

Финансирование: Работа выполнена при поддержке инициативного проекта VIP-016.

Для цитирования: N. S. Lagutina, A. M. Vasilyev, and D. D. Zafievsky, “Name Entity Recognition Tasks: Technologies and Tools”, *Modeling and analysis of information systems*, vol. 30, no. 1, pp. 64-85, 2023.

Введение

Именованная сущность (named entity, NE) является идентификатором имен людей, названий организаций, географических местоположений, событий, дат и т. п. в тексте на естественном языке. Впервые термин был использован в 1996 году [1]. С этого момента NER стало классической задачей NLP.

Большое количество исследований, посвященных NER, показывает не только актуальность этой задачи, но и необходимость обобщения и анализа накапливаемых результатов. Практически все обзоры, как ранние [2, 3], так и поздние [4, 5] систематизируют методы выделения NE. При этом методы глубокого обучения получают наибольшее внимание, так как часто дают самые лучшие результаты. Однако соревнования между исследователями в области NLP демонстрируют, что для разных задач одни и те же методы NER показывают различное качество работы [6, 7]. Поэтому авторы данного обзора решили систематизировать результаты NER, отталкиваясь от категорий задач в этой области. Анализ отдельных задач позволит более детально понять границы применения методов выделения NE, определить перспективные направления исследований, увидеть спектр областей применения инструментов NER. Обзор содержит работы за период с 2018 по 2022 годы, но включает и некоторые более ранние, показавшиеся авторам этой статьи важными и интересными.

Каждая категория задач описана в отдельном разделе статьи и снабжена таблицей, агрегирующей информацию об исследованиях: ссылку на работу, язык использованного корпуса текстов, базовый метод решения задачи, название корпуса текстов, качество решения F-мера. Второй раздел посвящен классической задаче NER, таблица 1. Третий раздел рассматривает подзадачи NER, таблица 2. Четвертый — NER в социальных сетях, таблица 3. Пятый — NER в предметных областях, таблица 5. Шестой — NER в задачах NLP, таблица 6. В седьмом разделе описаны эксперименты с тремя инструментами NLP: spaCy, Stanford's NER (Stanza) и DeepPavlov, предоставляющими NER API для русского языка. Авторы не включили в обзор большое количество работ, посвященных китайскому языку, из-за лингвистических особенностей, отличающих его от других языков.

1. Классическая задача NER

Классическая задача NER ставит цель распознавать в тексте на естественном языке такие единицы как имена людей, названия организаций, географические местоположения, числовые выражения, включая время, дату, деньги, и классифицировать их по соответствующим категориям. С момента начала исследований три категории NE получили наибольшее распространение: имена людей (PER), географические местоположения (LOC), названия организаций (ORG), иногда их называют generic NE [4]. NE, которые не соответствуют ни одному типу, иногда выделяют в отдельную категорию miscellanea (MISC). В таблице 1 приведены значения F-меры для разных категорий сущностей и среднее значение — overall. Кроме трех указанных, в таблицу 1 вошли продукты (PROD) и события (EVENT).

Показательный результат решения задачи NER для английского языка можно увидеть в работе [8]. Для построения модели текста авторы применили рекуррентные нейронные сети (RNN). Для выбора архитектуры сети использовался подход на основе дифференцированного поиска (DARTS). Эксперименты проводились с англоязычным корпусом текстов CoNLL-2003, который содержит четыре типа NE: PER, LOC, ORG, MISC. Качество NER F-мера составила 93.47 %.

Близкий по качеству результат NER F-мера 90.79 % получен для английского языка и в более ранней работе [9]. Авторы интегрировали в ансамбль четыре инструмента: Stanford Named Entity Recognizer, Illinois Named Entity Tagger, Ottawa Baseline Information Extraction и Apache OpenNLP Name Finder. Работа показала, что обучение в ансамбле может снизить частоту ошибок. Авторы указали значение F-меры для всех трех категорий NE (см. табл. 1). Однако столь высокий результат может быть обусловлен качеством подготовленного собственного корпуса текстов.

Table 1. Research devoted to the classical task of NER

Таблица 1. Исследования, посвященные классической задаче NER

Авторы	Язык	Метод	Корпус текстов	F-мера, %					
				Overall	PER.	LOC.	ORG.	PROD.	EVENT
[8]	английский	RNN+DARTS	CoNLL-2003	93.47	-	-	-	-	-
[9]	английский	ensemble learning	авторский	90.79	95.91	91.27	85.18	-	-
[10]	английский	BiLSTM-CRF	OntoNotes5	87.95	-	-	-	-	-
	английский	BiLSTM-CRF	CoNLL-2003	91.73	-	-	-	-	-
[11]	английский	BiLSTM-CRF	CoNLL-2003	92.42	-	-	-	-	-
	немецкий	BiLSTM-CRF	CoNLL-2003	86.21	-	-	-	-	-
	голландский	BiLSTM-CRF	CoNLL-2003	88.25	-	-	-	-	-
[12]	немецкий	BiLSTM-CRF	CoNLL-2003	82.99	-	-	-	-	-
	немецкий	BiLSTM-CRF	GermEval	81.83	-	-	-	-	-
[13]	бразильский	LSTM-CRF	LeNER-Br	92.53	93.47	60.54	88.38	-	-
[14]	испанский	BiLSTM-CRF	CoNLL-2002	87.08	-	-	-	-	-
[15]	испанский	(NLTK+GATE)-LSTM	CoNLL-2002	92.0	-	-	-	-	-
[16]	русский	BiLSTM-CRF	Gareev's dataset	87.17	95.08	-	84.41	-	-
	русский	BiLSTM-CRF	Persons-1000	99.26	99.26	-	-	-	-
	русский	BiLSTM-CRF	FactRuEval 2016	82.10	-	-	-	-	-
[17]	датский	Danish BERT	DaNE	83.76	93.52	87.30	78.27	-	-
[18]	чешский	BERT-CRF	авторский	93.9	-	-	-	-	-
	русский	BERT-CRF	авторский	87.3	-	-	-	-	-
	болгарский	BERT-CRF	авторский	87.2	-	-	-	-	-
	польский	BERT-CRF	авторский	93.2	-	-	-	-	-
[19]	болгарский	BERT-CRF	авторский	89.6	93.9	94.8	85.1	59.5	50.0
	чешский	BERT-BiLSTM-CRF	авторский	94.4	95.7	98.3	95.0	79.6	55.0
	польский	BERT-CRF	авторский	93.7	93.0	95.5	95.5	54.1	100.0
	русский	rule-based+ ML-techniques	авторский	86.1	93.3	98.7	92.5	65.1	-
[20]	арабский	BERT	AQMAR	65.5	74.4	68.4	34.6	-	-
	арабский	BERT	NEWS	78.6	89.5	80.5	60.8	-	-
[21]	английский	BERT+self-training	CoNLL-2003	81.48	-	-	-	-	-
	английский	BERT	Tweet	48.01	-	-	-	-	-
	английский	BERT	OntoNotes5	68.35	-	-	-	-	-
	английский	BERT	Webpage	65.74	-	-	-	-	-
	английский	BERT	Wikigold	60.07	-	-	-	-	-
[22]	португальский	Local Grammar	PrimeiroHarem	59.0	-	-	-	-	-
[23]	турецкий	rule-based	авторский	88.71	88.58	84.71	91.47	-	-
[24]	финский	rule-based	Digitoday	86.82	85.32	93.51	89.35	76.41	100.00
	финский	BiLSTM-CRF	Digitoday	84.59	80.90	89.76	88.62	74.23	80.00
[25]	малайский	ELMO+MTBR	Malay NER	96.32	-	-	-	-	-

Более надежно выглядит метод, показывающий стабильно высокие результаты на разных корпусах [10]. Качество результата показало F-меру 87.95 % для набора данных OntoNotes с 18 типами сущностей и 91.73 % для набора данных CoNLL-2003. Архитектура классификатора представляет собой популярный Bi-LSTM-CRF. Основой вектора числовых характеристик текста являются эмбеддинги слов, а также параметры символьного уровня.

Такая же архитектура нейронных сетей Bi-LSTM-CRF лежит в основе исследований NER для трех европейских языков [11]. Лучший результат оказался для текстов на английском: F-мера 92.42 %, немного ниже на голландском: F-мера 88.25 % и на немецком F-мера 86.21 %. Аналогичный подход использовался для немецкого языка в работе [12]. F-мера оказалась ниже — около 82 %, но почти одинакова для двух разных наборов данных.

Следует отметить очевидную популярность нейронных сетей, моделирующих долгую кратковременную память (LSTM), в частности двунаправленную (Bi-LSTM), в решении задачи NER. Их комбинация с условными случайными полями, вероятностными моделями для сегментации и мар-

кировки последовательностей данных (CRF), стала стандартом для получения глубоких контекстно-зависимых представлений текста [4]. Эта технология показывает высокий результат NER для бразильских [13] F-мера 92.53 % и испанских [14] F-мера 87.08 % текстов. Интересно, что для испанского языка более высокий результат (F-мера 92.0 %) получился после предварительного отбора NE с помощью стандартных инструментов для обработки текста: NLTK и GATE [15]. Для русского языка эта технология также дает хорошее качество решения задачи [16].

Нейронные сети лежат в основе нейросетевой модели естественных языков BERT. В последнее время BERT широко используется в задачах NLP, в том числе NER. Эта модель показала хорошее качество NER для датского языка F-мера 83.76 % [17]. Комбинация технологий BERT-CRF оказалась эффективна для четырех славянских языков [18, 19]. Детальные результаты этих исследований размещены в таблице 1.

Однако применение нейронных сетей невозможно без предварительно размеченных корпусов данных большого объема. Этот факт накладывает существенные ограничения на использование указанной технологии в случае малого количества текстов, в следствие дорогостоящего сбора и разметки. Поэтому менее требовательные к ресурсам методы остаются привлекательными для исследователей.

Исследователи [21] изучали проблему распознавания именованных объектов (NER) с помощью технологии опосредованного обучения. Опосредованное обучение не требует большого количества ручных аннотаций и позволяет получать маркеры текста через внешние базы знаний, однако в ряде случаев маркеры слов оказываются ошибочными или отсутствуют. Авторы предлагают новую вычислительную структуру — BOND, использующую модели BERT и RoBERTa. BOND использует предварительно обученные языковые модели общего назначения, что позволяет ей работать в условиях ограниченных вычислительных ресурсов и обучающих корпусов. Однако хорошее качество решения оказалось только для одного корпуса CoNLL-2003 F-мера 81.48 %.

Оливера и др. [22] разработали локальную грамматику для NER на португальском языке (Local Grammar). Однако качество решения задачи оказалось низким, F-мера 59.0 %. Гораздо лучше показали себя методы, основанные на правилах (rule-based) для турецких текстов, F-мера 88.71 % [23], русских, F-мера 86.1 % [19] и финских F-мера 86.82 % [24]. Для финского языка метод, основанный на правилах, даже превзошел популярную технологию BiLSTM-CRF, F-мера 84.59 %.

Таким образом, задача NER достаточно хорошо решена для многих естественных языков. Однако это можно сказать уверенно только для небольшого количества категорий сущностей, в первую очередь для PER, LOC, ORG. Другие категории, такие как PROD, EVENT, DATE исследованы намного меньше даже для английского языка. Частично это связано с отсутствием размеченных корпусов или их малым объемом.

Исследователи в поисках лучшего решения часто проводят широкий спектр экспериментов с разными корпусами и разными технологиями. Сравнение результатов этих экспериментов показывает расхождение в качестве NER. Этот факт ставит проблему определения условий и границ применения используемых технологий, а также поиска новых решений. Один из путей преодоления таких проблем: анализ ошибок. К сожалению, понять и объяснить процессы происходящие в нейронных сетях трудно, более перспективными для этой цели кажутся методы основанные на правилах.

2. Подзадачи NER

Качественное выделение стандартных типов NE может внести существенный вклад в анализ естественного языка и стать основой решения задач в области информационного поиска. Повышение эффективности решения сложной задачи невозможно без подробного анализа ее деталей. Поэтому в этом разделе мы обсудим подзадачи, возникающие в сфере NER.

Table 2. NER subtasks research

Таблица 2. Исследования, посвященные подзадачам NER

Авторы	Задача	Язык	Метод	Корпус текстов	F-мера, %
[26]	вложенные NER	английский	BiLSTM-CRF	GENIA	74.79
	вложенные NER	английский	BiLSTM-CRF	ACE 2005	72.25
[27]	вложенные NER	английский	BiLSTM-Detection Network-Classifier	ACE 2005	78.2
[28]	вложенные NER	английский	Span-level Graph	ACE 2005	85.11
[29]	вложенные NER	английский	CNN-BERT-BiLSTM	OntoNotes5	91.3
	вложенные NER	английский	CNN-BERT-BiLSTM	CoNLL-2003	93.5
	вложенные NER	немецкий	CNN-BERT-BiLSTM	CoNLL-2003	86.4
	вложенные NER	испанский	CNN-BERT-BiLSTM	CoNLL-2002	90.3
	вложенные NER	датский	CNN-BERT-BiLSTM	CoNLL-2002	93.7
[30]	вложенные NER	английский	BiLSTM	GENIA	77.1
	вложенные NER	английский	BiLSTM	JNLPBA	78.4
[31]	границы NE	английский	CNN-BiLSTM-GRU	OntoNotes5	93.94
[32]	границы NE	иврит	BiLSTM-CRF	HAARETZ	85.22
[33]	нормализация NE	английский	BERT+NN	NCBI, BC5CDR	89.1
[34]	нормализация NE	английский	BioBERT	NCBI, BC5CDR	86.6

NER в условиях ограниченных наборов обучающих данных можно рассматривать как отдельную важную подзадачу выделения NE. В работе [35] эта проблема решается с помощью адаптации прототипической нейронной сети (prototypical network), специальной модели, разработанной для классификации в условиях, когда размеченных примеров мало. Эта сеть отображает объекты в векторное пространство, в котором классификация может быть выполнена путем вычисления расстояний до прототипных представлений каждого класса. В экспериментах авторы строят корпус текстов на основе данных из Ontonotes с 18 размеченными категориями NE и уменьшают количество примеров обучающей выборки до 20 для отдельной категории. Качество распознавания при этом сильно отличается для каждого типа NE. Один из лучших результатов получается для PER: F-мера 82.32 %, ORG: 69.2 %, LOC: 52.0 %, EVENT: 45.2 %.

Среди других подзадач самое большое внимание уделяется выделению вложенных именованных сущностей (nested named entity recognition). Например, фразы «Банк Китая» и «Университет Вашингтона» являются вложенными именованными объектами, где в название организации вложено местоположение [36]. Учёные [26] предложили новую нейронную модель для идентификации вложенных объектов путем динамического наложения слоев flatNER. Каждый слой включает LSTM и CRF, которые создают новое представление для обнаруженных объектов, а затем передают их в следующий слой. Модель динамически складывает слои flatNER до тех пор, пока объекты извлекаются. Оценка показывает, что предлагаемый подход превосходит системы, основанные на стандартных методах NER, достигая F-меры 74.7 % и 72.2 % в наборах данных GENIA и ACE2005 соответственно.

Авторы статьи [27] представляют инструмент MGNER для многозначного распознавания именованных сущностей. Он справляется со случаями, когда несколько сущностей в предложении могут быть перекрывающимися или полностью вложенными. Текст моделируют эмбеддинги слов, полученные на основе их морфологических характеристик (POS-тегов), информации о словах на уровне символов и встраивания языковой модели ELMo [37]. Качество этого решения немного выше, чем у предыдущей работы. Еще выше показывает результат алгоритм, основанный на графах [28].

В работах [29] и [30] для выявления вложенных NE определяют все возможные фразы, которые потенциально могут быть именованными сущностями, и классифицируют их с помощью глубоких нейронных сетей. В обоих случаях применяется сеть BiLSTM. Однако авторы первой работы [29] применяют этот подход для классической задачи NER для четырех европейских языков с использованием эмбеддингов на основе CNN и BERT, а второй [30] – работают со специфическими NE

медицинской области. Качество выделения NE оказывается выше в среднем на 13 %, это видно из таблицы 2.

Подход, использованный в двух предыдущих работах, связан с еще одной подзадачей NER: выделение границ NE. Исследователи [31] отмечают, что ошибки, допущенные при обнаружении границ объектов, отрицательно влияют на последующие системы выделения и классификации NE. Кроме того, они обращают внимание, что инструмент, качественно определяющий границы NE без явной детализации категории, может быть полезен в качестве предварительной обработки текста для последующего анализа в зависимости от предметной области. Авторы строят ансамбль нейронных сетей для детекции границ NE и получают высокое качество результата: F-мера 93.94 %.

Израильские учёные [32] также обращают внимание на проблему выделения границ NE, но в контексте морфологически богатых языков. Они предлагают показатель, который содержит аннотации NER на уровне лексем и морфем для текстов на иврите. Исследователи используют стандартные технологии NER BiLSTM-CRF, добавляя к характеристикам уровня символов и слов морфологические параметры. Свой подход авторы оценивают для корпуса еврейских исторических текстов и получают F-меру 85.22 %.

Еще одной подзадачей в области NER является нормализация именованных сущностей (NEN), т. е. приведение выделенной NE к некоторой начальной форме. Это необходимо для сопоставления и идентификации объектов в тех случаях, когда одна и та же NE имеет синонимы, полное название и аббревиатуру, разные названия. Авторы работы [33] предлагают автоматизированную модель нормализации именованных сущностей на основе новой нейронной сети, строящей и сопоставляющей графы отношений между NE. Сопоставление и обновление весов ребер позволяет получить информацию о семантическом сходстве между совпадающими сущностями. В качестве параметров сущностей используется модель BERT. Та же задача для медицинских текстов решается учеными в работе [34] с использованием BioBERT и математических классификаторов.

Одно из направлений развития методов NER лежит в области применения сложных морфологических, синтаксических и семантических параметров текста. Авторы работы [38] указывают, что в области NER мало исследованы лексические характеристики текстов. Они добавляют к параметрам текста из обычной модели LSTM-CRF характеристики сходства слов с типами сущностей. Исследователи показывают, что качество NER сравнимо со стандартными подходами, но производительность их системы выше. На необходимость использования синтаксической информации обращается внимание и в работе [39].

3. NER в социальных сетях

Table 3. Research about NER in social media

Таблица 3. Исследования, посвященные NER в социальных сетях

Авторы	Язык	Метод	Корпус текстов	F-мера, %				
				Overall	PER.	LOC.	ORG.	MISC
[40]	английский	CNN-BiLSTM-CRF	WNUT2017	41.86	58.66	52.86	26.55	-
[41]	английский	CNN-BiLSTM-CRF	TWITTER-2015	70.69	81.98	78.95	53.07	34.02
[42]	испанский	BiLSTM-CRF	ProfNER Gold Standard	83.9	-	-	-	-
[43]	английский	InceptionV3-BERT-LSTM-CRF	TWITTER-2015	73.47	86.44	77.16	52.91	36.05
[44]	английский	BERT-MMI-CRF	TWITTER-2015	73.41	85.24	81.58	63.03	39.45
	английский	BERT-MMI-CRF	TWITTER-2017	85.31	91.56	84.73	82.24	70.10
[20]	арабский	BERT	авторский	57.3	-	-	-	-
[45]	персидский	BERT	ParsNER-Social corpus	89.65	-	-	-	-
[46]	индийский	Transformer+CRF	авторский	49.79	-	-	-	-
[47]	английский	CRF + PLP	комментарии Yahoo!	71.0	-	-	-	-
[15]	английский	(NLTK+GATE)-LSTM	авторский	90.0	-	-	-	-

Среди всех работ в области NER как отдельную задачу можно выделить NER в социальных сетях. С одной стороны, ее постановка совпадает с классической, с другой стороны, особенности текстов в этой области приводят к отличиям в технологиях и качестве решений.

Использование самой распространенной комбинации BiLSTM-CRF для выделения четырех стандартных категорий сущностей: PER, LOC, ORG, MISC, не приводит к хорошим результатам. Это показывают меры качества из работ [40, 41], приведенные в таблице 3. Скорее всего, это объясняется качеством текстов социальных сетей и специфическим стилем их авторов. В частности на эту особенность указывают исследователи [40], описывая корпус WNUT2017. Однако для испанского языка технология BiLSTM-CRF показала результат гораздо лучше, чем для английского [42]: F-мера 83.9 % против 70.69 %. Возможно это также связано с особенностями корпуса текстов.

Языковая модель BERT и дополнительные характеристики, извлеченные из изображений, сопровождающих тексты, позволили повысить результаты для английского языка. Для анализа изображений использовались специальные инструменты: InceptionV3 [43] (F-мера 73.47 %) и MMI [44]. F-мера оказалась равна 73.41 %.

BERT стал основой NER в социальных сетях для арабских и иранских исследователей. Однако качество распознавания арабских текстов низкое: F-мера 57.3 % [20], а иранских вполне хорошее: F-мера 89.65 % [45].

Исследователи [46] обратили внимание, что индийские твиты включают много английских слов на латинице, в частности, при обозначении именованных сущностей. Для NER ученые применили новую для момента выполнения работы архитектуру нейронной сети Transformer [48] вместо обычной BiLSTM с целью повышения производительности. Однако качество решения задачи оказалось совсем низким F-мера 49.79 %. Скорее всего это связано с сильно специфическим стилем корпуса текстов.

Нестандартный подход к NER в комментариях пользователей описан в работе [47]. Предлагаемый инструмент основан на графе кореферентности меток NE и алгоритме распространения меток для выявления именованных сущностей. F-мера 71 % этого решения сравнима с показателями области в целом.

Самый хороший результат был получен турецкими учёными [15] с показателем F-меры 90.0 %. Их метод состоит из двух этапов. Во-первых, обработка текста инструментами GATE и NLTK, в результате которой обеспечиваются входные данные для второго этапа. Во-вторых, использование обученной рекуррентной нейронной сети для определения NE.

Одной из основных проблем NER в социальных сетях является зашумлённость исходного текста [40, 47]. Однако, на наш взгляд, не менее важной проблемой является постоянное возникновение новых сущностей [15], так как многие стандартные технологии базируются на лексических ресурсах: словарях, Википедии и т. п. Эти ресурсы обновляются медленно. Поэтому решением проблемы может быть дорогостоящее постоянное переобучение или создание методов не зависящих от обучающих корпусов, например, тщательно проработанных методов, основанных на правилах и использующих языковые модели, определяющие маркеры слов с высоким качеством.

4. NER в предметных областях

Распознавание именованных сущностей в предметных областях существенно отличается от классической задачи NER. В последнем случае хорошо работают методы, фиксирующие семантику слов на основе простых текстов и общих особенностей естественного языка. В случае NER в предметной области этого недостаточно [76]. Выявление самих NE и их категорий требует использования информации специфичной для конкретной области.

Самое большое количество исследований NER посвящено биомедицине. Это связано со значительным количеством доступных словарей и размеченных корпусов текстов: о заболеваниях NCBI-Disease, с названиями генов и белков BC2GM and JNLPBA, о химических веществах BC4CHEMD,

Table 4. Biomedical NER research

Таблица 4. Исследования, посвященные NER в биомедицине

Авторы	Язык	Метод	Корпус текстов	F-мера, %
[49]	английский	BiLSTM-CRF	BC2GM	80.74
	английский	BiLSTM-CRF	BC4CHEMD	89.37
	английский	BiLSTM-CRF	JNLPBA	73.52
	английский	BiLSTM-CRF	BC5CDR	88.78
	английский	BiLSTM-CRF	NCBI-Disease	86.14
[50]	английский	BiLSTM-CRF	BC2GM	79.73
	английский	BiLSTM-CRF	BC4CHEMD	88.85
	английский	BiLSTM-CRF	JNLPBA	77.58
	английский	BiLSTM-CRF	BC5CDR-chem	93.31
	английский	BiLSTM-CRF	BC5CDR-disease	84.08
	английский	BiLSTM-CRF	NCBI-Disease	86.36
[51]	английский	BiLSTM-CRF	CRAFT	72.31
	английский	BiLSTM-CRF	BioNLP CG	78.20
	английский	BiLSTM-CRF	PDR	83.44
	английский	BiLSTM-CRF	JNLPBA	77.78
	английский	BiLSTM-CRF	BC5CDR	90.57
	английский	BiLSTM-CRF	NCBI-Disease	87.47
[52]	английский	BERT+rule-based	BC2GN	86.7
	английский	BiLSTM-CRF	BC3GN	50.1
	английский	BiLSTM-CRF	OSIRISv1.2	88.14
	английский	BiLSTM-CRF	Thomas corpus	89.08
[53]	английский	Multi-BERT	EN EHR	92.39
	английский	Multi-BERT	EN UGT	84.88
	русский	Multi-BERT	RU UGT	60.45
	русский	Multi-BERT	RU EHR	85.00
[54]	итальянский	BiLSTM-CRF	SIRM COVID-19	89.01
	итальянский	BERT	SIRM COVID-19	88.58

BC5CDR. Первые три исследования, приведенные в таблице 4, используют классическую технологию NER BiLSTM-CRF для указанных корпусов текстов. Качество распознавания NE стабильно высокое: F-мера 73 % - 93 %. Интересно, что наблюдается корреляция между F-мерой в разных работах и соответствующими корпусами. Это показывает, что технология NER, базирующаяся на глубоком обучении, вполне подходит для предметной области и зависит от качества и объема обучающего корпуса текстов.

Другая технология, сочетающая языковую модель BERT и правила, также показывает хорошие результаты: F-мера 86 % - 89 % [52]. Как основа решения задачи BERT использован и в работе [53]. Исследователи определяли названия заболеваний и медикаментов для текстов на английском и русском языках. Лучшие результаты получились для английского: F-мера 92.39 %. Авторы сравнили технологии BERT и BiLSTM-CRF и показали преимущество первого, в частности для русского языка. Однако технология BiLSTM-CRF была успешно применена для итальянского [54]. В работе [77] также выявлено преимущество модели BERT. Однако реальное применение методов в промышленных программах требует оценки эффективности с точки зрения времени работы и используемой памяти [64].

Хорошие результаты NER дает классический подход BiLSTM-CRF в химической области [55, 56, 78], см. таблицу 5. А также в извлечении NE из научных статей [57] и в материаловедении [58, 60].

Интересно, что в текстах о еде лучшие показатели F-меры, более 95 %, оказались у методов, основанных на правилах [62, 65]. Возможно это связано с особенностями структуры текстов, используемых в основе корпусов, например, рецептов [63]. Хотя и стандартный подход показывает высокое качество F-меры: 92.5 % [61]. Методы, основанные на правилах, были использованы

Table 5. Domain NER research

Таблица 5. Исследования, посвященные NER в предметных областях

Авторы	Область	Язык	Метод	Корпус текстов	F-мера, %
[55]	химия	английский	Att-BiLSTM-CRF	CHEMDNER	90.84
[56]	химия	английский	BiLSTM-CRF	CHEMDNER	90.0
[57]	наука	английский	BiLSTM-CRF	Macromolecules+ICPSR	90.9
[58]	материаловедение	английский	BiLSTM-CRF	авторский	87.04
[59]	производство	английский	BiLSTM-CRF	авторский	88.0
[60]	материаловедение	английский	BiLSTM-CRF	авторский	91.66
[61]	еда	английский	BERT-BiLSTM-CRF	YELP	92.5
[62]	еда	английский	rule-based	Allrecipes+MyRecipes	96.05
[63]	еда	английский	BERT+BiLSTM+Dict	авторский	81.31
[64]	продукты	английский	Dict+CRF	FoodBase	97.4
[65]	диетология	английский	rule-based	Documents from USDA	97.0
[66]	садоводство	английский	rule-based+CRF	article abstracts of journals	81.04
[67]	кибербезопасность	английский	rule-based+BiLSTM-CRF	авторский	75.05
[68]	кибербезопасность	английский	BERT-BiLSTM-CRF	Bridges dataset	96.87
[69]	кибербезопасность	русский	BERT	авторский	72.74
[70]	юриспруденция	английский	BERT-CRF	EDGAR-NER	96.1
[71]	юриспруденция	турецкий	BiLSTM	авторский	92.27
[72]	археология	голландский	BERT-CRF	авторский	73.5
[73]	география	французский	(CasEN+CoreNLP)-CRF	GeoNER-corpus	85.2
[74]	финансы	английский	dependency parse trees+reg_exp	TU Delft's repository	80.0
[75]	биомедицина	английский	CNN-BiGRU-CRF	BioNLP13PC	85.11
	кино	английский	CNN-BiGRU-CRF	MIT Movie	86.41
	рестораны	английский	CNN-BiGRU-CRF	MIT Restaurant	78.05
	безопасность	английский	CNN-BiGRU-CRF	Re3d	68.52
	финансы	английский	CNN-BiGRU-CRF	SEC	83.44
[76]	биомедицина	английский	entity selector-CRF	NCBI	86.12
	общие	английский	entity selector-CRF	CoNLL03	92.40
	биология	английский	entity selector-CRF	GENIA	75.59
	финансы	английский	entity selector-CRF	SEC	75.59

также для отбора кандидатов на роль NE для анализа аннотаций статей в области садоводства на английском: F-мера 81.04 % [66].

Авторы работы [70] выделяли NE в юридических текстах. Они использовали модель BERT обученную на общем англоязычном корпусе CoNLL-2003 и на специализированном EDGAR-NER и сравнивали результаты с более ранними моделями. Ранние модели показали невысокую эффективность в обоих случаях. Для более продвинутых моделей CRF и BERT, обученных на юридических текстах, качество распознавания существенно превышает случай обучения тех же моделей на общих английских текстах. Однако авторы не могут сказать, связано ли с различиями в моделях или различиями в тестовых коллекциях.

Сочетание методов, основанных на правилах, и технологий глубокого обучения было использовано для NER в области кибербезопасности [67]. Однако качество оказалось не слишком высоким: F-мера 75.05 %. Намного лучше результаты в этой области дало использование BERT для текстов на английском языке: F-мера 96.87 % [68]. Для русского языка учёные отметили, что аннотированный набор данных содержит небольшое количество объектов этих типов [69], и предложили автоматический метод увеличения набора данных с использованием BERT. В итоге качество NER оказалось 72.74 %.

Близкий по качеству результат получили исследователи археологических текстов на датском языке: F-мера 75.0 % [72]. В этой работе авторы применили BERT для получения характеристик слов и классификатор CRF для определения категории NE. Похожий подход был использован для выявления географических названий во французских текстах, но с более высоким результатом: F-мера

85.2 % [73]. Определение кандидатов NE осуществлялось с использованием существующих инструментов CasEN, CoreNLP, Perdido, SEM and spaCy. Дерево синтаксических зависимостей, построенное с помощью SpaCy для получения текстовых шаблонов, используется в работе [74]. Авторы объединили эту информацию с регулярными выражениями и разработали AckNER, инструмент для выделения финансовой информации.

Из анализа статей мы видим большое разнообразие предметных областей, выбранных для NER. Так же как и в классической задаче, методы на основе глубокого обучения дают хорошие результаты при наличии больших размеченных корпусов текстов, используемых для обучения. В большинстве предметных областей таких данных мало [75]. Интересно, что во многих исследованиях источником текстов являются научные статьи. Скорее всего это связано с эффективностью сбора и разметки данных. Для небольших корпусов лучше работают методы, основанные на правилах, но они сильно зависят от особенностей языка и структуры текста в целом.

Проблемы разнообразия предметных областей исследователи пытаются решить созданием универсальных инструментов, пригодных для использования в разных сферах жизни. В работе [75] предлагается систему, уже обученную для NER, адаптировать к новой предметной области. Другие исследователи [76] строят инструмент на основе базы знаний. Интересно, что в обоих случаях средняя F-мера похожи: 80.3 % и 84.82 % соответственно, но во втором всё-таки выше. Качество работы этих двух систем для разных предметных областей приведено в конце таблицы 5.

Кроме того, в последних по времени работах приобретают популярность уже существующие инструменты для NLP. Это связано с развитием качества их работы, оно существенно увеличилось буквально за последние несколько лет. Заметно так же, что подавляющее большинство работ выполнено для текстов на английском языке. Это дает богатый спектр исследовательских задач для национальных языков.

5. NER в задачах NLP

Table 6. Research devoted to NER in NLP tasks

Таблица 6. Исследования, посвященные NER в задачах NLP

Authors	Task	Language	Method	Text corpora	F-мера, %
[79]	аннотирование	английский	spaCy	English-Inshorts	-
[60]	аннотирование	английский	BiLSTM	авторский	91.4
[80]	аннотирование	английский	BiLSTM-CRF	авторский	89.35
[81]	аннотирование	русский	-	авторский	-
[82]	тематическая классификация	английский	weighting scheme NE	авторский	-
[83]	тематическая классификация	английский	BiLSTM + CRF	авторский	88.0
[84]	машинный перевод	английский немецкий	flairNLP	субтитры	-
[85]	анализ тональности	английский	Stanford Named Entity Tagger	SES, MSR2016	-
[86]	анализ тональности	английский	LSTM	авторский	-
[63]	онтология	английский китайский	BERT-BiLSTM-CRF	авторский	81.31
[87]	вопросно-ответные системы	английский	BERT	авторский	-
[88]	определение лжи	английский	spaCy, Stanford's NER	отзывы	-

С одной стороны, NER является классической самостоятельной задачей NLP. С другой стороны, есть ряд задач в области обработки текста и извлечения информации, в которые NER входит как подзадача или часть технологии решения. В этом разделе мы рассмотрим примеры таких задач. Для тех исследований, в которых NER рассматривается как выделенная подзадача, в таблице 6 приведена оценка качества ее решения. Остальные работы имеют другую цель и используют инструменты NER как часть технологии достижения этой цели и качество непосредственно NER не оценивают.

Сингх и др. [79] решают задачу аннотирования текста и отмечают, что одной из основных проблем классических алгоритмов является потеря именованных объектов. В работе предложен метод сохранения именованных объектов, которые полезны при создании точных аннотаций. В качестве инструмента NER исследователи используют spaCy. В другой работе [60] рассматривают NER как отдельную подзадачу аннотирования текста и посвящают ей свое исследование. Для решения используется нейронная сеть Bi-LSTM. Аналогичную задачу в коммерческих документах решают авторы [80]. Именованные сущности играют важную роль в системе аннотирования русскоязычных текстов [81].

Авторы работы [82] отмечают важную роль именованных объектов в задаче классификации текстов по тематике. В статье рассматривается байесовская версия алгоритма LDA. Исследователи модифицируют алгоритм, увеличивая веса NE и показывают, что именованные сущности хорошо подходят для использования в качестве терминов, специфичных для предметной области. Kumar A., Starly B. [83] отмечают, что одной из серьезных проблем при анализе текстов промышленного производства является отсутствие модели распознавания специфичных именованных объектов. Они предлагают собственную модель NER для классификации производственных текстов по категориям.

Участники конференции по переводу речи IWSLT 2021 [84] используют инструмент flairNLP для распознавания именованных сущностей, чтобы повысить качество машинного перевода субтитров. Другая специализированная система Stanford Named Entity Tagger, как часть решения задачи аспектного анализа тональности, применяется в работе [85]. Исследователи [86] для той же задачи строят собственную систему распознавания на основе LSTM.

Китайские учёные [63] использовали стандартную технологию, сочетающую BERT, BiLSTM и CRF для NER как часть системы построения графа предметной области. Разработанный AliMe KG, граф знаний в электронной коммерции, помогает понять потребности пользователей, ответить на их вопросы и создать поясняющие тексты, в том числе руководство по покупкам, ответы на вопросы о недвижимости, создание точек продаж. Другое исследование [87] напрямую изучает вопрос NER в вопросно-ответных системах.

Сложная цель построения системы автоматической детекции лжи была поставлена в работе [88]. Авторы используют именованные сущности для моделирования трех принципов: правдивые рассказчики предоставляют более подробные отчеты, правдивые тексты содержат больше упоминаний конкретных лиц, мест и дат, а обманщики склонны скрывать потенциально проверяемую информацию. Основой моделирования является доля именованных объектов, полученных с помощью двух инструментов spaCy и Stanford's NER.

Описанные работы показывают очень широкие возможности применения технологий NER. Особенно стоит обратить внимание на уже существующие инструменты NLP. Так же как и для извлечения именованных сущностей в предметных областях, они часто используются современными исследователями. Этот факт послужил мотивацией для сравнительного анализа качества работы spaCy, Stanford's NER (Stanza) и DeepPavlov для русского языка, родного для авторов обзора.

6. Эксперименты

6.1. Инструменты обработки естественного языка

Для анализа авторы выбрали инструменты, которые активно используются в научных исследованиях академическими кругами и в коммерческих приложениях. Актуальность и качество инструментов обеспечивается активностью их разработчиков, которая выражается в постоянном обновлении программной реализации.

spaCy — это библиотека Python с открытым исходным кодом, которая предоставляет инструменты для многих задач NLP, включая подсистему NER [89]. Библиотека достаточно гибкая и позволяет использовать для анализа текста как подходы основанные на правилах, так и на нейронных сетях.

Пользователь может либо использовать существующую текстовую модель, либо обучить новую. spaCy предоставляет три предварительно обученные модели для русского языка, которые нацелены на различные варианты баланса между требованиями к качеству и вычислительной мощности. В данном исследовании мы использовали самую большую доступную модель, `ru_core_news_lg`, которая должна обеспечивать наилучшее качество. Часть модели NER была обучена с использованием размеченного корпуса текстов Nerus[1] и реализована с использованием рекуррентной нейронной сети.

Stanford NLP Group предоставляет пакет Stanza NLP, который включает инструменты для анализа текста и предварительно обученные модели для более чем 70 языков [90]. Подсистема инструментария NER включает в себя модели для 17 языков. Модель распознавания именованных сущностей использует слои BiLSTM с представлениями на уровне символов и слов, а также слой декодирования CRF для получения прогнозов NER. Для русского языка Stanza NLP предоставляет единственную модель NER, которая была обучена на базе данных WikiNER [91]. Инструмент также предоставляет средства для обучения новой модели.

Последним инструментом, использованным в экспериментах, является диалоговый фреймворк DeepPavlov [92]. Фреймворк включает в себя теггер NER и набор предварительно обученных моделей. Модели могут использовать различные конструкции нейронных сетей, включая рекуррентные нейронные сети, BERT или гибридные. DeepPavlov предоставляет три предварительно подготовленные модели для русского языка. Все они были обучены с использованием набора данных Collection3 [93]. Для нашего исследования мы решили использовать модель `ner_rus_bert_torch`, поскольку авторы инструмента считают ее лучшей моделью, а также используют BERT, который не используется в других инструментах.

Использование описанных инструментов и моделей должно позволить проверить характеристики различных конструкций нейронных сетей в задаче NER для русского языка.

6.2. Корпус текстов

Чтобы оценить качество инструментов, мы не могли задействовать общедоступные наборы данных, поскольку они использовались для обучения моделей NER и дали бы значительное преимущество для соответствующего инструмента. Поэтому для использования в экспериментах мы сформировали собственный корпус текстов. Этот корпус включает тексты из различных источников: статьи из Википедии, новостные статьи, биографии исполнителей и тексты песен. Таким образом, эксперимент покажет качество NER для произвольного авторского текста из русского Интернета.

Корпус текстов, который мы собрали, содержит 50 текстов из различных источников. В каждом тексте мы вручную разметили именованные объекты для категорий PER, LOC, ORG и MISC. Количество слов и NE для каждого текста показано на рис. 1. В целом корпус содержит 616 LOC, 680 PER, 698 ORG и 966 объектов MISC.

Количество категорий было ограничено теми, которые предоставляют обученные модели. Модель Stanza может распознавать сущности во всех четырех категориях, но модели spaCy и DeepPavlov выделяют только категории PER, LOC и ORG.

6.3. Архитектура испытательного стенда

Оценка инструментов была полностью автоматизирована с использованием скриптов Python[2]. Эти скрипты применяют выбранные инструменты и анализируют результат распознавания именованных сущностей. Поток данных внутри испытательного стенда показан на рис. 2.

Каждый выбранный инструмент является библиотекой Python и, следовательно, должен быть встроен в приложение для выполнения. Для каждого из инструментов мы разработали

[1] https://github.com/natasha/nerus

[2] https://github.com/yarfruct/ner-russian-language-experiments

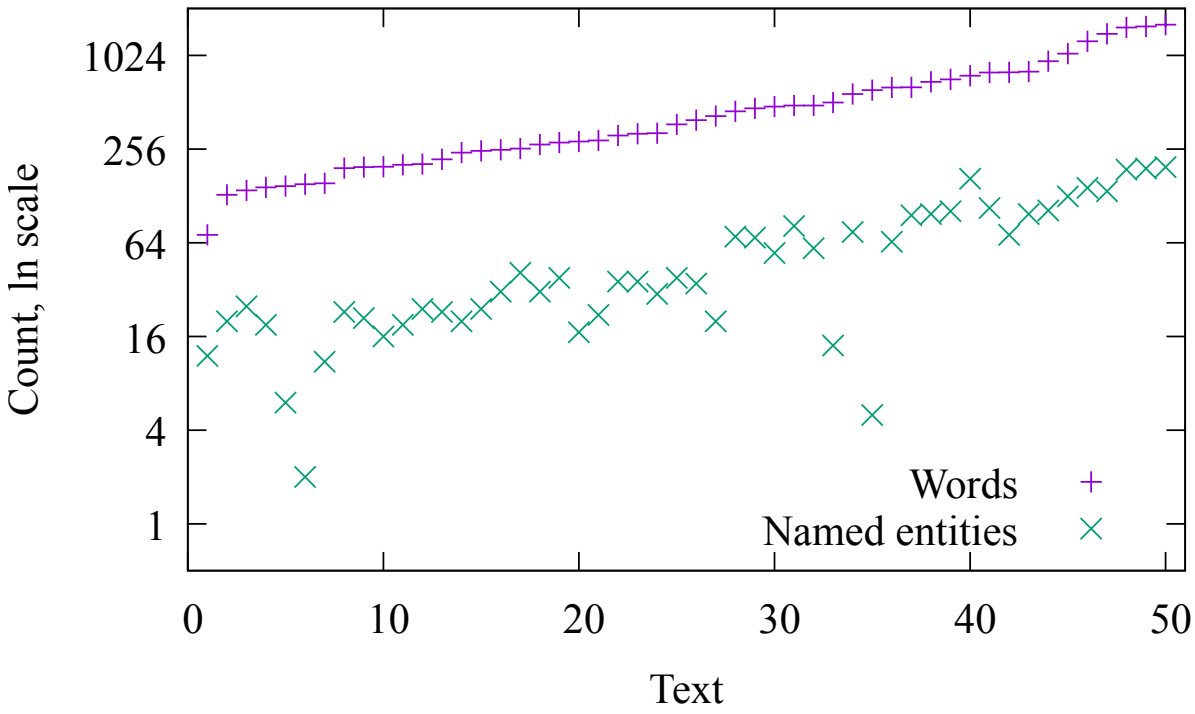

Fig. 1. Words and NE count for texts in the corpus

Рис. 1. Структура корпуса текстов

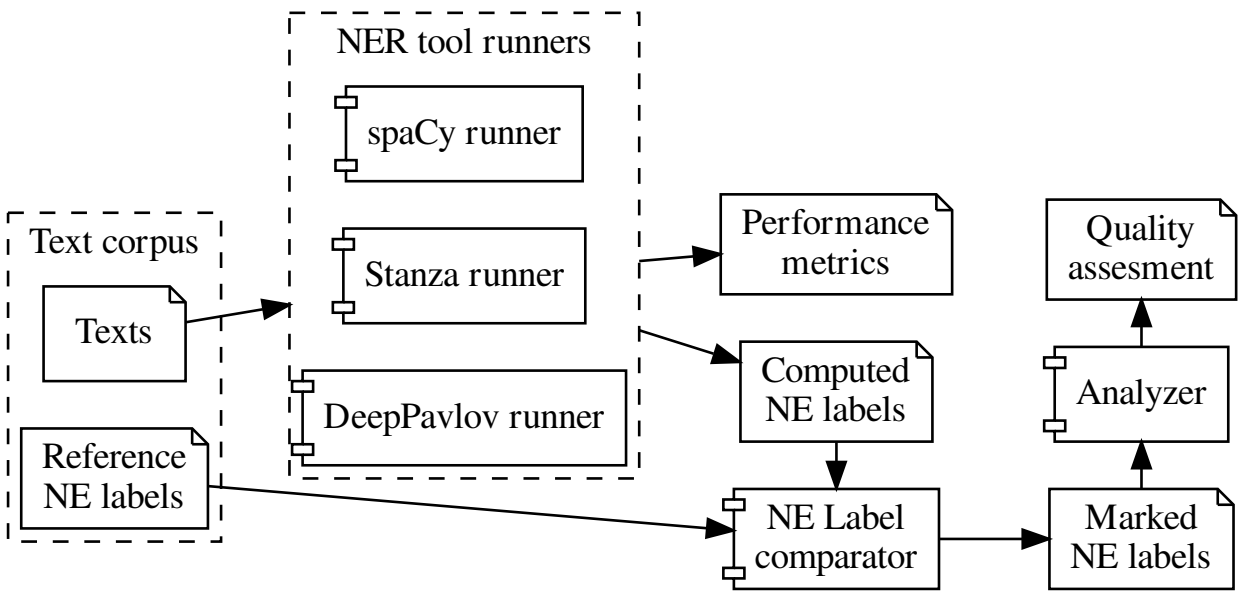

Fig. 2. Test testbed components and data they use for operation

Рис. 2. Схема процесса тестирования

соответствующее приложение, которое обеспечивает унифицированный интерфейс: принимает

текст в качестве входных данных и предоставляет распознанные именованные объекты в общем формате.

Помимо типов NE, методы инструментов предоставляют показатели производительности, которые позволяют отслеживать время выполнения распознавания именованных объектов. Для уменьшения влияния вариабельности отдельного эксперимента, каждый запуск инструмента повторялся 10 раз. В качестве результата взяты среднее арифметическое и медиана времени выполнения.

Следующим важным компонентом испытательного стенда является анализатор распознавания категорий именованных сущностей. Он берет категории NE, вычисленные инструментом, и сравнивает их с эталонными. Результатом работы компонента является список NE, в котором указывается, является ли вычисленная категория NE правильной или нет. Последний шаг — это расчет F-меры. Оценка рассчитывается для всех именованных объектов и для каждой поддерживаемой категории.

Помимо компонентов, показанных на рис. 2, существует сценарий запуска, который способен выполнять все остальные компоненты. Во время одного выполнения пользователь может выбрать инструмент, который он хочет запустить, и указать местоположение исследуемого корпуса текстов. Категория MISC распознаётся только Stanza.

6.4. Результаты экспериментов

Table 7. Results of experiment **Таблица 7.** Результаты экспериментов

	Stanza с MISC	Stanza без MISC	spaCy	DeepPavlov
F-мера, %	82	87.1	79.5	89
F-мера PER, %	94.1	94.1	86.1	93.8
F-мера ORG, %	75.7	75.7	66	79.8
F-мера LOC, %	95.9	95.9	88.2	94
F-мера MISC, %	65.2	–	–	–
Среднее время исполнения, сек.	138.376	138.376	80.604	79.938
Медианное время исполнения, сек.	138.042	138.042	78.81	79.36

Эксперименты проводились на персональном компьютере, оснащенном процессором Intel® Core™ i5-9300H, работающим под управлением операционной системы Ubuntu 20.04. В ходе экспериментов использовались последние выпущенные версии инструментов: Stanza 1.3, spaCy 3.2 и DeepPavlov 3.7.9. Все инструменты используют только процессор для работы нейронной сети.

Результаты экспериментов представлены в таблице 7. Лучшие результаты можно увидеть у DeepPavlov с F-мерой 89 %, Stanza и spaCy следуют с 82 % и 79,5 % соответственно. Если не учитывать MISC, Stanza обеспечивает F-меру 87,1 %, сокращая разрыв между DeepPavlov. Если учитывать F-меру и время выполнения, DeepPavlov является лучшим инструментом для собранного корпуса. Все инструменты дают хорошие результаты для именованных сущностей категорий PER и LOC, но плохие результаты для категории ORG. Категория MISC также плохо распознается Stanza.

7. Заключение

Самый большой прогресс в области NER достигнут с помощью моделей, основанных на глубоком обучении и зависящих от объёмных эталонных наборов данных. Это относится как отдельным типам задач в этой области, так и к обработке текстов на национальных языках. Создание размеченных корпусов данных помогает решить эту проблему, но является очень дорогостоящим процессом, требующим постоянного обновления в связи с изменяющимся состоянием сфер жизни.

Одним из перспективных направлений исследований в области NER является развитие методов на основе обучения без учителя или на основе правил. Возможной базой предобработки текста для таких методов могут служить интенсивно развивающиеся модели языков в существующих инструментах NLP. Такой подход может решить проблему больших вычислительных ресурсов, которые

требуются для автономных систем NER. Кроме того, использование сложных лексических параметров текста и баз правил позволит лучше понять и объяснить процессы и результаты решения поставленных задач.

References

[1] R. Grishman and B. Sundheim, "Message understanding conference-6: A brief history", in *Proceedings of the 16th International Conference on Computational Linguistics (COLING 96), Copenhagen, August 1996*, 1996, pp. 466–471.

[2] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification", *Lingvisticae Investigationes*, vol. 30, no. 1, pp. 3–26, 2007.

[3] R. Sharnagat, "Named entity recognition: A literature survey", *Center For Indian Language Technology*, pp. 1–27, 2014.

[4] J. Li, A. Sun, J. Han, and C. Li, "A survey on deep learning for named entity recognition", *IEEE Transactions on Knowledge & Data Engineering*, vol. 34, no. 1, pp. 50–70, 2022.

[5] G. Popovski, B. K. Seljak, and T. Eftimov, "A survey of named-entity recognition methods for food information extraction", *IEEE Access*, vol. 8, pp. 31 586–31 594, 2020.

[6] J. Piskorski, B. Babych, Z. Kancheva, O. Kanishcheva, M. Lebedeva, M. Marcińczuk, P. Nakov, P. Osenova, L. Pivovarova, S. Pollak, *et al.*, "Slav-ner: The 3rd cross-lingual challenge on recognition, normalization, classification, and linking of named entities across slavic languages", in *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, 2021, pp. 122–133.

[7] A. Starostin, V. Bocharov, S. Alexeeva, A. Bodrova, A. Chuchunkov, *et al.*, "Factrueval 2016: Evaluation of named entity recognition and fact extraction systems for russian", in *Computational Linguistics and Intellectual Technologies*, 2016, pp. 702–720.

[8] Y. Jiang, C. Hu, T. Xiao, C. Zhang, and J. Zhu, "Improved differentiable architecture search for language modeling and named entity recognition", in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 3585–3590.

[9] R. Speck and A.-C. Ngonga Ngomo, "Ensemble learning for named entity recognition", in *International semantic web conference*, Springer, 2014, pp. 519–534.

[10] A. Ghaddar and P. Langlais, "Robust lexical features for improved neural network named-entity recognition", in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 1896–1907.

[11] A. Akbik, T. Bergmann, and R. Vollgraf, "Pooled contextualized embeddings for named entity recognition", in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 724–728.

[12] M. Riedl and S. Padó, "A named entity recognition shootout for german", in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2018, pp. 120–125.

[13] P. H. L. de Araujo, T. E. de Campos, R. R. de Oliveira, M. Stauffer, S. Couto, and P. Bermejo, "Lener-br: A dataset for named entity recognition in brazilian legal text", in *International Conference on Computational Processing of the Portuguese Language*, Springer, 2018, pp. 313–323.

[14] Y. Luo, F. Xiao, and H. Zhao, "Hierarchical contextualized representation for named entity recognition", in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 8441–8448.

[15] N. Eligüzel, C. Çetinkaya, and T. Dereli, "Application of named entity recognition on tweets during earthquake disaster: A deep learning-based approach", *Soft Computing*, vol. 26, no. 1, pp. 395–421, 2022.

[16] M. Y. Arkhipov, M. S. Burtsev, *et al.*, "Application of a hybrid bi-lstm-crf model to the task of russian named entity recognition", in *Conference on Artificial Intelligence and Natural Language*, Springer, 2017, pp. 91–103.

[17] R. Hvingelby, A. B. Pauli, M. Barrett, C. Rosted, L. M. Lidegaard, and A. Søgaard, "Dane: A named entity resource for danish", in *Proceedings of the 12th Language Resources and Evaluation Conference*, 2020, pp. 4597–4604.

[18] M. Arkhipov, M. Trofimova, Y. Kuratov, and A. Sorokin, "Tuning multilingual transformers for named entity recognition on slavic languages", *BSNLP'2019*, p. 89, 2019.

[19] J. Piskorski, L. Laskova, M. Marcińczuk, L. Pivovarova, P. Přibáň, J. Steinberger, and R. Yangarber, "The second cross-lingual challenge on recognition, normalization, classification, and linking of named entities across slavic languages", in *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, 2019, pp. 63–74.

[20] C. Helwe, G. Dib, M. Shamas, and S. Elbassuoni, "A semi-supervised bert approach for arabic named entity recognition", in *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, 2020, pp. 49–57.

[21] C. Liang, Y. Yu, H. Jiang, S. Er, R. Wang, T. Zhao, and C. Zhang, "Bond: Bert-assisted open-domain named entity recognition with distant supervision", in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 1054–1064.

[22] E. Oliveira, G. Dias, J. Lima, and J. Pirovani, "Using named entities for recognizing family relationships", in *Anais do IX Symposium on Knowledge Discovery, Mining and Learning*, SBC, 2021, pp. 24–32.

[23] R. Yeniterzi, G. Tür, and K. Oflazer, "Turkish named-entity recognition", in *Turkish Natural Language Processing*, Springer, 2018, pp. 115–132.

[24] T. Ruokolainen, P. Kauppinen, M. Silfverberg, and K. Lindén, "A finnish news corpus for named entity recognition", *Language Resources and Evaluation*, vol. 54, no. 1, pp. 247–272, 2020.

[25] Y. Fu, N. Lin, Z. Yang, and S. Jiang, "Towards malay named entity recognition: An open-source dataset and a multi-task framework", *Connection Science*, pp. 1–23, 2022.

[26] M. Ju, M. Miwa, and S. Ananiadou, "A neural layered model for nested named entity recognition", in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, pp. 1446–1459.

[27] C. Xia, C. Zhang, T. Yang, Y. Li, N. Du, X. Wu, W. Fan, F. Ma, and P. Yu, "Multi-grained named entity recognition", in *57th Annual Meeting of the Association for Computational Linguistics, ACL 2019*, Association for Computational Linguistics (ACL), 2020, pp. 1430–1440.

[28] J. Wan, D. Ru, W. Zhang, and Y. Yu, "Nested named entity recognition with span-level graphs", in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 892–903.

[29] J. Yu, B. Bohnet, and M. Poesio, "Named entity recognition as dependency parsing", in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 6470–6476.

[30] M. G. Sohrab and M. Miwa, "Deep exhaustive model for nested named entity recognition", in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 2843–2849.

[31] J. Li, D. Ye, and S. Shang, "Adversarial transfer for named entity boundary detection with pointer networks", in *IJCAI*, 2019, pp. 5053–5059.

[32] D. Bareket and R. Tsarfaty, "Neural modeling for named entities and morphology (nemo^2)", *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 909–928, 2021.

[33] S. H. Jeon and S. Cho, "Edge weight updating neural network for named entity normalization", *Neural Processing Letters*, pp. 1–22, 2022.

[34] D. Zhou and T. Liu, "Joint model of biomedical entity recognition and normalization labels based on self-attention", in *Second International Symposium on Computer Technology and Information Science (ISCTIS 2022)*, SPIE, vol. 12474, 2022, pp. 461–466.

[35] A. Fritzler, V. Logacheva, and M. Kretov, "Few-shot classification in named entity recognition task", in *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, 2019, pp. 993–1000.

[36] J. R. Finkel and C. D. Manning, "Nested named entity recognition", in *Proceedings of the 2009 conference on empirical methods in natural language processing*, 2009, pp. 141–150.

[37] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations. arxiv preprint arxiv: 180205365", in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, 2018, pp. 2227–2237.

[38] A. Ghaddar, P. Langlais, A. Rashid, and M. Rezagholizadeh, "Context-aware adversarial training for name regularity bias in named entity recognition", *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 586–604, 2021.

[39] Y. Nie, Y. Tian, Y. Song, X. Ao, and X. Wan, "Improving named entity recognition with attentive ensemble of syntactic information", in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 4231–4245.

[40] G. Aguilar, S. Maharjan, A. P. López-Monroy, and T. Solorio, "A multi-task approach for named entity recognition in social media data", *W-NUT 2017*, p. 148, 2017.

[41] Q. Zhang, J. Fu, X. Liu, and X. Huang, "Adaptive co-attention network for named entity recognition in tweets", in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018, pp. 5674–5681.

[42] A. Miranda-Escalada, E. Farré-Maduell, S. Lima-López, L. Gascó, V. Briva-Iglesias, M. Agüero-Torales, and M. Krallinger, "The profner shared task on automatic recognition of occupation mentions in social media: Systems, evaluation, guidelines, embeddings and corpora", in *Proceedings of the Sixth Social Media Mining for Health (# SMM4H) Workshop and Shared Task*, 2021, pp. 13–20.

[43] M. Asgari-Chenaghlu, M. R. Feizi-Derakhshi, L. Farzinvash, M. Balafar, and C. Motamed, "Cwi: A multimodal deep learning approach for named entity recognition from social media using character, word and image features", *Neural Computing and Applications*, pp. 1–18, 2021.

[44] J. Yu, J. Jiang, L. Yang, and R. Xia, "Improving multimodal named entity recognition via entity span detection with unified multimodal transformer", in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 3342–3352.

[45] M. Asgari-Bidhendi, B. Janfada, O. Roshani Talab, and B. Minaei-Bidgoli, "Parsner-social: A corpus for named entity recognition in persian social media texts", *Journal of AI and Data Mining*, vol. 9, no. 2, pp. 181–192, 2021.

[46] R. Priyadharshini, B. R. Chakravarthi, M. Vegupatti, and J. P. McCrae, "Named entity recognition for code-mixed indian corpus using meta embedding", in *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, IEEE, 2020, pp. 68–72.

[47] M. C. Phan and A. Sun, "Collective named entity recognition in user comments via parameterized label propagation", *Journal of the Association for Information Science and Technology*, vol. 71, no. 5, pp. 568–577, 2020.

[48] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need", *Advances in neural information processing systems*, vol. 30, pp. 5998–6008, 2017.

[49] X. Wang, Y. Zhang, X. Ren, Y. Zhang, M. Zitnik, J. Shang, C. Langlotz, and J. Han, "Cross-type biomedical named entity recognition with deep multi-task learning", *Bioinformatics*, vol. 35, no. 10, pp. 1745–1752, 2019.

[50] W. Yoon, C. H. So, J. Lee, and J. Kang, "Collabonet: Collaboration of deep neural networks for biomedical named entity recognition", *BMC bioinformatics*, vol. 20, no. 10, pp. 55–65, 2019.

[51] L. Weber, M. Sänger, J. Münchmeyer, M. Habibi, U. Leser, and A. Akbik, "Hunflair: An easy-to-use tool for state-of-the-art biomedical named entity recognition", *Bioinformatics*, vol. 37, no. 17, pp. 2792–2794, 2021.

[52] D. Kim, J. Lee, C. H. So, H. Jeon, M. Jeong, Y. Choi, W. Yoon, M. Sung, and J. Kang, "A neural named entity recognition and multi-type normalization tool for biomedical text mining", *IEEE Access*, vol. 7, pp. 73 729–73 740, 2019.

[53] Z. Miftahutdinov, I. Alimova, and E. Tutubalina, "On biomedical named entity recognition: Experiments in interlingual transfer for clinical and social media texts", in *European Conference on Information Retrieval*, Springer, 2020, pp. 281–288.

[54] R. Catelli, F. Gargiulo, V. Casola, G. De Pietro, H. Fujita, and M. Esposito, "Crosslingual named entity recognition for clinical de-identification applied to a covid-19 italian data set", *Applied Soft Computing*, vol. 97, p. 106 779, 2020.

[55] L. Luo, Z. Yang, P. Yang, Y. Zhang, L. Wang, H. Lin, and J. Wang, "An attention-based bilstm-crf approach to document-level chemical named entity recognition", *Bioinformatics*, vol. 34, no. 8, pp. 1381–1388, 2018.

[56] W. Hemati and A. Mehler, "Lstmvoter: Chemical named entity recognition using a conglomerate of sequence labeling tools", *Journal of cheminformatics*, vol. 11, no. 1, pp. 1–7, 2019.

[57] Z. Hong, R. Tchoua, K. Chard, and I. Foster, "Sciner: Extracting named entities from scientific literature", in *International Conference on Computational Science*, Springer, 2020, pp. 308–321.

[58] L. Weston, V. Tshitoyan, J. Dagdelen, O. Kononova, A. Trewartha, K. A. Persson, G. Ceder, and A. Jain, "Named entity recognition and normalization applied to large-scale information extraction from the materials science literature", *Journal of chemical information and modeling*, vol. 59, no. 9, pp. 3692–3702, 2019.

[59] A. Kumar and B. Starly, ""fabner": Information extraction from manufacturing process science domain literature using named entity recognition", *Journal of Intelligent Manufacturing*, vol. 33, no. 8, pp. 2393–2407, 2022.

[60] S. Moon, G. Lee, S. Chi, and H. Oh, "Automated construction specification review with named entity recognition using natural language processing", *Journal of Construction Engineering and Management*, vol. 147, no. 1, p. 04 020 147, 2021.

[61] M. H. Syed and S.-T. Chung, "Menuner: Domain-adapted bert based ner approach for a domain with limited dataset and its application to food menu domain", *Applied Sciences*, vol. 11, no. 13, p. 6007, 2021.

[62] G. Popovski, S. Kochev, B. Korousic-Seljak, and T. Eftimov, "Foodie: A rule-based named-entity recognition method for food information extraction.", in *ICPRAM*, 2019, pp. 915–922.

[63] F.-L. Li, H. Chen, G. Xu, T. Qiu, F. Ji, J. Zhang, and H. Chen, "Alimekg: Domain knowledge graph construction and application in e-commerce", in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020, pp. 2581–2588.

[64] N. Perera, T. T. L. Nguyen, M. Dehmer, and F. Emmert-Streib, "Comparison of text mining models for food and dietary constituent named-entity recognition", *Machine Learning and Knowledge Extraction*, vol. 4, no. 1, pp. 254–275, 2022.

[65] T. Eftimov, B. Koroušić Seljak, and P. Korošec, "A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations", *PloS one*, vol. 12, no. 6, e0179488, 2017.

[66] Z. Liu, M. Luo, H. Yang, and X. Liu, "Named entity recognition for the horticultural domain", in *Journal of Physics: Conference Series*, IOP Publishing, vol. 1631, 2020, p. 012 016.

[67] G. Kim, C. Lee, J. Jo, and H. Lim, "Automatic extraction of named entities of cyber threats using a deep bi-lstm-crf network", *International journal of machine learning and cybernetics*, vol. 11, no. 10, pp. 2341–2355, 2020.

[68] S. Zhou, J. Liu, X. Zhong, and W. Zhao, "Named entity recognition using bert with whole world masking in cybersecurity domain", in *2021 IEEE 6th International Conference on Big Data Analytics (ICBDA)*, IEEE, 2021, pp. 316–320.

[69] M. Tikhomirov, N. Loukachevitch, A. Sirotina, and B. Dobrov, "Using bert and augmentation in named entity recognition for cybersecurity domain", in *International Conference on Applications of Natural Language to Information Systems*, Springer, 2020, pp. 16–24.

[70] T. W. T. Au, I. J. Cox, and V. Lampos, "E-ner–an annotated named entity recognition corpus of legal text", *arXiv preprint arXiv:2212.09306*, 2022.

[71] C. Cetindag, B. Yaziciouglu, and A. Koc, "Named-entity recognition in turkish legal texts", *Natural Language Engineering*, pp. 1–28, 2022.

[72] A. Brandsen, S. Verberne, K. Lambers, and M. Wansleeben, "Can bert dig it? named entity recognition for information retrieval in the archaeology domain", *Journal on Computing and Cultural Heritage (JOCCH)*, vol. 15, no. 3, pp. 1–18, 2022.

[73] E. Kogkitsidou and P. Gambette, "Normalisation of 16th and 17th century texts in french and geographical named entity recognition", in *Proceedings of the 4th ACM SIGSPATIAL Workshop on Geospatial Humanities*, 2020, pp. 28–34.

[74] D. Alexander and A. P. de Vries, "" this research is funded by...": Named entity recognition of financial information in research papers", in *Proceedings of the 11th International Workshop on Bibliometric-enhanced Information Retrieval co-located with 43rd ECIR*, Sl: CEUR, 2021, pp. 102–110.

[75] J. Li, S. Shang, and L. Shao, "Metaner: Named entity recognition with meta-learning", in *Proceedings of The Web Conference 2020*, 2020, pp. 429–440.

[76] B. Nie, C. Li, and H. Wang, "Ka-ner: Knowledge augmented named entity recognition", in *China Conference on Knowledge Graph and Semantic Computing*, Springer, 2021, pp. 60–75.

[77] B.-S. Lin, J.-H. Chen, and T.-H. Chang, "Nerve at rocling 2022 shared task: A comparison of three named entity recognition frameworks based on language model and lexicon approach", in *Proceedings of the 34th Conference on Computational Linguistics and Speech Processing (ROCLING 2022)*, 2022, pp. 343–349.

[78] T. Isazawa and J. M. Cole, "Single model for organic and inorganic chemical named entity recognition in chemdataextractor", *Journal of Chemical Information and Modeling*, vol. 62, no. 5, pp. 1207–1213, 2022.

[79] B. Singh, A. Marathe, A. A. Rizvi, and A. R. Joshi, "Retaining named entities for headline generation", in *Inventive Computation and Information Technologies*, Springer, 2021, pp. 221–234.

[80] Y. Ji, C. Tong, J. Liang, X. Yang, Z. Zhao, and X. Wang, "A deep learning method for named entity recognition in bidding document", in *Journal of Physics: Conference Series*, IOP Publishing, vol. 1168, 2019, p. 032 076.

[81] S. M. Makeev, S. V. SHekshuev, M. G. Petrov, and S. D. Zyuzin, "Modul' obrabotki estestvennogo yazyka na osnove obuchennoj modeli nejronnoj seti s mekhanizmom poiska imenovannyh sushchnostej", *Izvestiya Tul'skogo gosudarstvennogo universiteta. Tekhnicheskie nauki*, no. 9, pp. 56–64, 2022.

[82] K. Krasnashchok and S. Jouili, "Improving topic quality by promoting named entities in topic modeling", in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2018, pp. 247–253.

[83] A. Kumar and B. Starly, ""fabner": Information extraction from manufacturing process science domain literature using named entity recognition", *Journal of Intelligent Manufacturing*, pp. 1–15, 2021.

[84] A. Siekmeier, W. Lee, H. Kwon, and J.-H. Lee, "Tag assisted neural machine translation of film subtitles", in *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, 2021, pp. 255–262.

[85] J. Ding, H. Sun, X. Wang, and X. Liu, "Entity-level sentiment analysis of issue comments", in *Proceedings of the 3rd International Workshop on Emotion Awareness in Software Engineering*, 2018, pp. 7–13.

[86] J. Yu, J. Jiang, and R. Xia, "Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 429–439, 2019.

[87] Z. A. Guven and M. O. Unalir, "Improving the bert model with proposed named entity recognition method for question answering", in *2021 6th International Conference on Computer Science and Engineering (UBMK)*, IEEE, 2021, pp. 204–208.

[88] B. Kleinberg, M. Mozes, A. Arntz, and B. Verschuere, "Using named entities for computer-automated verbal deception detection", *Journal of forensic sciences*, vol. 63, no. 3, pp. 714–723, 2018.

[89] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd, *Spacy: Industrial-strength natural language processing in python, zenodo, 2020.*

[90] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza: A Python natural language processing toolkit for many human languages", in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2020. [Online]. Available: https://nlp.stanford.edu/pubs/qi2020stanza.pdf.

[91] J. Nothman, N. Ringland, W. Radford, T. Murphy, and J. R. Curran, “Learning multilingual named entity recognition from wikipedia”, *Artificial Intelligence*, vol. 194, pp. 151–175, 2013.

[92] M. S. Burtsev, A. V. Seliverstov, R. Airapetyan, M. Arkhipov, D. Baymurzina, N. Bushkov, O. Gureenkova, T. Khakhulin, Y. Kuratov, D. Kuznetsov, *et al.*, “Deeppavlov: Open-source library for dialogue systems.”, in *ACL (4)*, 2018, pp. 122–127.

[93] V. Mozharova and N. Loukachevitch, “Two-stage approach in russian named entity recognition”, in *2016 International FRUCT Conference on Intelligence, Social Media and Web (ISMW FRUCT)*, IEEE, 2016, pp. 1–6.