KazNERD: Kazakh Named Entity Recognition Dataset

Rustem Yeshpanov, Yerbolat Khassanov, Huseyin Atakan Varol

Institute of Smart Systems and Artificial Intelligence,
Nazarbayev University, Nur-Sultan, Kazakhstan
{rustem.yeshpanov, yerbolat.khassanov, ahvarol}@nu.edu.kz

Abstract

We present the development of a dataset for Kazakh named entity recognition. The dataset was built as there is a clear need for publicly available annotated corpora in Kazakh, as well as annotation guidelines containing straightforward—but rigorous—rules and examples. The dataset annotation, based on the IOB2 scheme, was carried out on television news text by two native Kazakh speakers under the supervision of the first author. The resulting dataset contains 112,702 sentences and 136,333 annotations for 25 entity classes. State-of-the-art machine learning models to automatise Kazakh named entity recognition were also built, with the best-performing model achieving an exact match F_1-score of 97.22% on the test set. The annotated dataset, guidelines, and codes used to train the models are freely available for download under the CC BY 4.0 licence from https://github.com/IS2AI/KazNERD.

Keywords: Named Entity Recognition, NER, Kazakh, Dataset, Annotation Guidelines, CRF, BiLSTM, BERT

1. Introduction

Named entity recognition (NER) refers to a subtask of information extraction aimed at identifying named entities (NEs) in semi- or unstructured text and classifying them into pre-specified types (Nadeau and Sekine, 2007). NEs, in turn, generally refer to (proper) names of persons, organisations, and geographical locations (Sang and Meulder, 2003), as well as numerical and temporal expressions, including quantities, monetary units, percentages, dates, or durations (Chinchor, 1998). Widely used in natural language processing applications, including automatic text understanding (Cheng and Erk, 2020), machine translation (Babych and Hartley, 2003), question answering (Aliod et al., 2006), and knowledge base development (Etzioni et al., 2005) to name a few, NER has been of interest not only to scientific research, but also to business (Schön et al., 2019) and defence (Han et al., 2020) ever since 1995, when the term was coined (Grishman and Sundheim, 1996).

By virtue of most of the early works in information extraction being launched as part of United States Government initiatives (e.g., ACE, MUC, TIPSTER) (Maynard et al., 2003), a great deal of research in NER concerns English. Nonetheless, an equally large proportion of NER research has been dedicated to different well-resourced languages, such as Spanish, French, German, Japanese, Chinese, Russian (see Nadeau and Sekine (2007), for a detailed overview), as well as to less resourced ones, such as Sindhi (Ali et al., 2020), Romanian (Dumitrescu and Avram, 2020), and Icelandic (Ingólfsdóttir et al., 2019).

Likewise low-resourced, the language of interest of this paper—Kazakh—has only latterly appeared on the radar of NER researchers. Underrepresented and lexically underdeveloped because overshadowed by Russian, which was promoted as a lingua franca during the Soviet era (Dave, 2007), the earliest NER research in this agglutinative Turkic language dates back as recently as 2016. Although there is evidence for annotated corpus construction as part of Kazakh NER research (Akhmed-Zaki et al., 2020; Tolegen et al., 2016), to our knowledge, neither of the corpora is publicly available. In addition, none of the studies into Kazakh NER appears to have developed annotation guidelines—or at least adapted those existing in other languages—to take into account cases characteristic of the Kazakh language.

Given this relatively nascent stage of Kazakh NER accompanied by the digital underrepresentation of the language and the lack of freely accessible annotated corpora, it is hoped that our research will fill the existing gaps in the field and thus contribute to its further development. Particularly, we built a dataset consisting of 112,702 sentences from television news, of which 86,246 are unique sentences and 26,456 are their various representations. All sentences in the dataset were manually annotated by two native Kazakh-speaking linguists, supervised by the first author. This resulted in the largest Kazakh NE annotated corpus. To assist the annotators in making the right choices when presented with expressions potentially matching NEs, annotation guidelines in Kazakh were developed. The guidelines contain rules for annotating 25 NE types, as well as relatable examples of Kazakh NEs. Finally, we built four state-of-the-art machine learning models to automatise Kazakh NER, with the highest exact match F_1-score reaching 97.22% on the test set.

The remainder of the paper proceeds as follows: Section 2 reviews existing research on Kazakh NER. Section 3 discusses data collection and preparation, the development of the guidelines and dataset. Section 4 provides the annotated dataset specifications, including the description of NEs, as well as the dataset structure and statistics. Section 5 offers the details of the implemented NER models, the experimental setup, and the evaluation criteria and results. Section 6 discusses the results of the experiment. Section 7 concludes the paper.

2. Related Work

As mentioned earlier, Kazakh is a digitally low-resourced language, with a small number of (annotated) corpora freely available. That said, recently, there have been progressive efforts made to address such

underrepresentation. Khassanov et al. (2021) have built a crowdsourced freely accessible Kazakh speech corpus (KSC) containing 332 hours of transcribed audio. In another work, Mussakhojayeva et al. (2021a) have constructed the first publicly available large-scale Kazakh text-to-speech synthesis dataset consisting of approximately 93 hours of transcribed audio recordings spoken by male and female professional narrators.

While Kazakh speech processing research has been gathering momentum, thanks to the recent development of publicly available datasets, Kazakh NER research can hardly boast of commensurable progress, which appears to be chiefly due to a lack of such resources. One of the earliest studies into Kazakh NER was conducted by Sadykova and Ivanov (2016). To build a manually-annotated Kazakh NE corpus, two experts were tasked with labelling 1,000 news articles with a set of seven NEs—namely, (1) person, (2) organisation, (3) location, (4) geopolitical entity (GPE), (5) event, (6) award, and (7) tender—using the brat rapid annotation tool (BRAT) (Stenetorp et al., 2012). Approximately 3,000 NEs are reported to have been tagged, of which 1,084 were persons, 974 locations, and 973 organisations. However, no breakdown of the remaining NEs is provided in the paper, nor is reference made to the metric applied to achieve an inter-annotator agreement (IAA) score of 0.86–0.89 (Artstein and Poesio, 2008). Another criticism is that, while the annotation guidelines are reported to have been developed specifically for the task, there is no mention of how to access them or the resulting annotated corpus.

Tolegen et al. (2016) created a Kazakh NE corpus, annotated according to the IOB (Inside, Outside, Beginning) scheme, from 2,500 general news media articles. The corpus is reported to consist of 18,054 sentences and 270,306 words. Annotation was performed using a self-developed web-based tool, with two native Kazakh speakers using the MUC-7 NE task definition (Chinchor, 1998) as a guide. More than 14,000 NEs were labelled in three categories: 4,292 persons, 7,391 locations, and 2,560 organisations. The IAA measured with Fleiss' kappa ranged from 0.93 to 0.98 (Fleiss, 1971). Furthermore, the scholars conducted an extensive analysis of Kazakh morphological and word type features and were the first to apply a statistical model to Kazakh NER based on conditional random fields (CRFs) (Lafferty et al., 2001), achieving an F_1-score of 89.81%.

The same model was used as a baseline in Tolegen et al (2020), where the researchers approached the Kazakh NER task by comparing (1) a bidirectional long short-term memory (BiLSTM) model (Hochreiter and Schmidhuber, 1997), (2) BiLSTM with CRF (BiLSTM-CRF), and (3) a tensor layer-based deep neural network (DNN) model. While the performance of the BiLSTM model yielded a result significantly lower than that of the baseline model (78.76%), the performance of the BiLSTM-CRF model varied depending on whether or not character embedding was used, 86.45% and 80.28%, respectively. The DNN model outperformed the other models, producing an F_1-score of 90.49%. Although the three models were trained on the annotated corpus built in Tolegen et al. (2016), neither of the studies provides information on access to it.

In Kozhirbayev and Yessenbayev (2020), an annotated NE corpus comprising 29,629 sentences was constructed in the IOB format, with the names of persons, organisations, and locations tagged along with Other, a category for NEs of interest that presumably fall outside the three said categories. Four methods to address the Kazakh NER task were applied—specifically, (1) the random forest classifier (Ho, 1995), (2) the Naïve Bayes classifier (Friedman et al., 1997), (3) CRFs, and (4) a hybrid method of BiLSTM and CRF. The results show that, while the first two methods achieved an F_1-score in the range of 81% to 89%, the hybrid method was notably outperformed by the CRFs, 88% versus 99%, in turn. However, the study included no information on what guidelines were followed to build the corpus, the quantities of NEs in the corpus, and how, if any, annotation accuracy checks were performed.

Kuralbayev et al. (2020) compared four NER models—(1) CRFs, (2) LSTM with character embedding, (3) LSTM-CRF, and (4) bidirectional encoder representations from transformers (BERT) (Devlin et al., 2019)—to anonymise 40,000 court decisions in Kazakh and Russian. The names of persons, organisations, locations and addresses were tagged using a self-built annotation tool. The scholars note that the BERT model, which was run without fine-tuning, reached an F_1-score of 87%, with the results of the other models peaking at 82%. Nevertheless, some notes of caution are warranted here, because, although the model is reported to have achieved high accuracy for both Kazakh and Russian, it was trained exclusively on Russian data. Furthermore, surprisingly, no mention is made of the guidelines used or the IAA assessment, considering that the annotation was carried out by over 150 local university students recruited for the task. Nor is it stated how many NEs were anonymised as a result.

The last study on Kazakh NER we discuss in this paper is by Akhmed-Zaki et al. (2020), who applied the BiLSTM, CRF, and BERT methods to a dataset collected from Kazakh online news portals. The dataset was manually annotated using the IOB scheme with four NEs—(1) persons, (2) locations, (3) organisations, and (4) other. In this study, too, the BERT model performed the best with an F_1-score of 97.99%, followed by CRF (94.27%) and BiLSTM (85.31%). While the study provides clear information on the parameters of the BERT model and formulae for the precision, recall, and F_1-scores computed, it is still limited by the lack of clarity on the volume of the data. Although the dataset built is claimed to consist of 7,153 sentences, the scholars explicitly state that it was split into 6,507, 2,531, and 3,015 sentences for training, validation, and test sets, respectively, which is 12,053 sentences in the aggregate. It is also unclear whether the category Other was used for NEs that were not names of persons, locations, and organisations, but were still of interest (see, e.g., Kozhirbayev and Yessenbayev (2020)), or whether it simply referred to a category of words that are not annotated as NEs and are labelled as O in the IOB scheme. Much like in the previous studies, no reference is made to the annotation guidelines adhered, the annotators and their backgrounds, the measurement of IAA, and the means of accessing the annotated dataset.

Representation designations	Example sentence		Count
AID	«Доу Джонс» **ORG**	бес бүтін жүзден сексен алты проценткe **PERC** құнсызданды.	86,246
BID	«Доу Джонс» **ORG**	5,86 проценткe **PERC** құнсызданды.	23,969
CID	«Доу Джонс» **ORG**	5,86%-ке **PERC** құнсызданды.	1,340
DID	Dow Jones **ORG**	бес бүтін жүзден сексен алты проценткe **PERC** құнсызданды.	809
EID	Dow Jones **ORG**	5,86 проценткe **PERC** құнсызданды.	326
FID	Dow Jones **ORG**	5,86%-ке **PERC** құнсызданды.	12
Total number of sentences			**112,702**

Table 1: Details of sentence representations, including designations, sentence count of each representation variant, and an example sentence translated as 'Dow Jones has depreciated by 5.86%.'

3. Annotated Corpus Construction

3.1. Source Data

The source data were obtained from the television news of the Khabar Agency, a major broadcasting network in Kazakhstan. With the agency's permission, the Kazakh transcribed text accompanying the original news posted on their official website[1] was collected over the second half of 2020. The news included reports on events in local and international politics, economy, sports, religion, and education that did not necessarily occur during the data collection period, as some news items were also extracted from the agency's archives. The extracted text[2] was not screened for inappropriate content on the assumption that this must have been prudently done by the agency's content policy department. The text was split sentence-wise—with an identifier assigned to each sentence—and inspected for grammatical and spelling errors (cf. Tolegen et al. (2016)) and homoglyphs. Duplicate sentences and those containing only Russian utterances were removed; sentences with both Kazakh and Russian utterances were retained, as Kazakh-Russian codeswitching is normal practice in Kazakhstan (Pavlenko, 2008; Mussakhojayeva et al., 2021b). Ultimately, the total number of sentences was 86,246.

3.2. Sentence Representation

To enable the developed NER models (see Section 5) recognise instances of the same NE regardless of their typographic characteristics (e.g., numerals written in words and digits), the following six sentence representation variants were adopted:

1) **AID** — All sentence elements were recorded in the Cyrillic script[3]. Arabic and Roman numerals (e.g., 9 → *тоғыз*, IV → *төрт*, etc.), names of organisations, applications, events, and so on, spelt in Latin characters

(e.g., *Bank of America* → *Банк оф Америка*, *Telegram* → *Телеграм*, etc.), terms conventionally spelt in Latin characters (e.g., *PhD* → *ПиЭйчДи*, etc.), and special symbols (e.g., *%* → *процент* or *пайыз*) were recorded in Cyrillic words.

2) **BID** — Sentences of the AID representation with numerals recorded in digits.

3) **CID** — Sentences of the BID representation with percentages recorded using the % symbol.

4) **DID** — Sentences of the AID representation with words conventionally spelt in the Latin script recorded in that script.

5) **EID** — Sentences of the DID representation with numerals recorded in digits.

6) **FID** — Sentences of the EID representation with percentages recorded using the % symbol.

The assigned representation designations, as well as example sentences with the resulting quantity of each variant in the dataset are summarised in Table 1.

3.3. Annotation Scheme

The IOB2 scheme—also referred to as BIO—was selected for annotation (Sang and Veenstra, 1999). Under this scheme, each token in text receives one of three tags—namely, **B**, **I** or **O**, indicating whether a token is at the **B**eginning, **I**nside or **O**utside of an annotated extent. It is similar to the IOB scheme except that a **B** tag is used at the beginning of every NE extent (see Table 2).

Tokens	IOB2 tags
Dow	B-ORGANISATION
Jones	I-ORGANISATION
5	B-PERCENTAGE
,	I-PERCENTAGE
86	I-PERCENTAGE
%-ке	I-PERCENTAGE
құнсызданды	O
.	O

Table 2: Example of the IOB2 annotation scheme

---

[1] www.khabar.kz

[2] Tokens denoting speech disfluencies and hesitations (parenthesised) and background noise [bracketed] were retained in the transcribed text.

[3] At the time of writing, the Kazakh language is undergoing a gradual transition from the Cyrillic to the Latin script, with the full transition scheduled to take place between 2023 and 2031.

3.4. Annotation Guidelines

Considering that none of the studies on Kazakh NER provided Kazakh annotation guidelines that our study could rely on to embark on the task, we decided to create such a set of rules. First, we studied some of the most referenced annotation guidelines for NER—particularly, Chinchor (1998), Brunstein (2002), Raytheon BBN Technologies (2004), Linguistic Data Consortium (2008), and Weischedel et al. (2012). Next, the first author experimentally annotated a random sample of 2,000 sentences to see what NEs could actually be extracted from the data on hand. Twenty-two NEs described in the guidelines studied were found in the sample. The first draft of the annotation guidelines containing the definition of an NE, information on the valid boundaries of NEs, rules for NE classification, and related examples was prepared in Kazakh.

Later, as a result of the annotator training task, it was decided to tag three more NEs whose examples were found in the news reports annotated. The NEs under consideration were NON_HUMAN, MISCELLANEOUS, and ADAGE. While the first two had been previously mentioned in the existing annotation guidelines for NER, the decision to tag ADAGE rested upon the relatively frequent use of Kazakh proverbs and sayings in the training sentences. Due adjustments were made to the guidelines, with some rules clarified and supported by comprehensible examples.

It is also worth mentioning at this point that the guidelines were iteratively amended as annotation proceeded. This was partly due to subsequent encounters with cases unconsidered while drafting the guidelines and partly as a result of daily discussions of questions posed by the annotators hired for the task. For a complete list of the 25 NEs and their brief descriptions, see Table 3. The final annotation guidelines (in Kazakh) are available for download from our GitHub repository[4].

3.5. Annotation Workflow

Two native Kazakh-speaking linguists received training in NER for two weeks under the supervision of the first author. As part of training, 3,500 sentences from the Khabar agency's official website were annotated, by following the developed guidelines. The annotation was carried out using the Webanno web-based tool (Yimam et al., 2013) (see (Neves and Seva, 2021) for an extensive review of various tools for annotation). The annotators worked independently on the same version of a text file, which was subsequently reviewed by the first author for annotation divergences and inconsistencies. The final version of the file contained text with annotations approved or modified as appropriate by the first author. During the training period, the IAA score, computed by Webanno, reached a Fleiss' kappa of 0.94.

The annotation process proceeded for six months, with the annotators labelling 1,500 sentences per day and the first author inspecting these once they were marked *complete* on Webanno. During the period, the IAA score was in the range of 0.95 to 0.97 Fleiss' kappa. Table 3 provides the statistics for annotated NEs.

---

[4]https://github.com/IS2AI/KazNERD

4. KazNERD Specifications

4.1. Named Entity Descriptions

The resulting annotated Kazakh NER dataset (hereafter KazNERD) contains 136,333 NEs. As can be seen from Table 3, the top three NEs in KazNERD are CARDINAL, DATE, and GPE. None of the previous Kazakh NER studies has labelled the first two classes. The latter class, embracing names of geopolitical entities, has often been conflated with names of geographical locations under the class LOCATION.

Since news reports are normally preceded by (or at least contain) the day or time when a particular event occurred, the frequent use of dates in the dataset was expectable—a total of 25,446 DATE NEs. What is indeed remarkable is the use of numbers in KazNERD. The two classes denoting numbers, CARDINAL (29,260) and ORDINAL (3,870), comprise practically a quarter of the total quantity of NEs in the dataset, a hefty 24.3%.

Interestingly, in KazNERD, the triad of NEs most commonly labelled in Kazakh NER research—locations (2,175), persons (13,577), and organisations (7,587)—ranks only third through fifth, even when GPE (17,543) is combined with LOCATION. Also worthy of note is the class ADAGE. The class deriving purely from our observations of Kazakh news and hardly fitting the conventional profile of an NE per se (but rather labelled out of scholarly interest) numbered 196 entities in total. This is higher or comparable in size to the classes MISCELLANEOUS, CONTACT, and NON_HUMAN, previously described as relatively frequent in the NER literature.

There are only eight instances of the class NON_HUMAN, which includes names of creatures other than humans. Such a scarcity of the NEs in the dataset was expected, given that the source data came from television news, which generally reports real-life events. Nevertheless, it was decided to label the NEs as a separate class for consistency with the existing annotation guidelines for NER.

As regards MISCELLANEOUS, the class embraces names of school and university subjects, types of computer networks and technologies, livestock breeds, and other entities that we had difficulty in categorising or deemed superfluous to label as separate classes.

The remaining NEs in KazNERD have been commonly annotated in the existing NER literature and guidelines. The relatively high ranking of the POSITION class (seventh overall, with 6,141 NEs) can be attributed to the domain of television news, which frequently reports on resolutions and activities of individuals holding official titles and occupational positions. The same applies to news reports on the economy, finance, trade, legal frameworks, business and political objectives, and technology in the country in particular and in the world in general, resulting in NEs annotated for the classes MONEY, PERCENTAGE, QUANTITY, PROJECT, PRODUCT, and LAW, accounting for a total of 11.29% of all NEs found in KazNERD.

The names of national and international cultural and political events, as well as the times and venues at which these were held; geographical, ethnic, and religious origins of persons participating in the events among other things;

No.	Entity classes	Definition	Size #	Size %
1	(ADA)GE	Well-known Kazakh proverbs and sayings	196	0.14
2	ART	Titles of books, songs, television programmes, etc.	2,407	1.77
3	(CAR)DINAL	Cardinal numbers, including whole numbers, fractions, and decimals	29,260	21.46
4	(CON)TACT	Addresses, emails, phone numbers, URLs	198	0.15
5	DATE	Dates or periods of 24 hours or more	25,446	18.66
6	(DIS)EASE	Diseases or medical conditions	1,272	0.93
7	(EVE)NT	Named events and phenomena	1,658	1.22
8	(FAC)ILITY	Names of man-made structures	2,145	1.57
9	GPE	Names of geopolitical entities	17,543	12.87
10	(LAN)GUAGE	Named languages	443	0.32
11	LAW	Named legal documents	533	0.39
12	(LOC)ATION	Names of geographical locations other than GPEs	2,175	1.60
13	(MIS)CELLANEOUS	Entities of interest but hard to assign a proper tag to	244	0.18
14	(MON)EY	Monetary values	4,560	3.34
15	(NON)_HUMAN	Names of pets, animals or non-human creatures	8	0.01
16	NORP	Adjectival forms of GPE and LOCATION; named religions, etc.	3,714	2.72
17	(ORD)INAL	Ordinal numbers, including adverbials	3,870	2.84
18	(ORG)ANISATION	Names of companies, government agencies, etc.	7,587	5.57
19	(PERC)ENTAGE	Percentages	4,283	3.14
20	(PER)SON	Names of persons	13,577	9.96
21	(POS)ITION	Names of posts and job titles	6,141	4.50
22	(PROD)UCT	Names of products	738	0.54
23	(PROJ)ECT	Names of projects, policies, plans, etc.	2,111	1.55
24	(QUA)NTITY	Length, distance, etc. measurements	3,908	2.87
25	TIME	Times of day and time duration less than 24 hours	2,316	1.70
	Total number of named entities		**136,333**	**100**

Note. The parenthesised NE classes will thus be referenced in the tables hereafter.

Table 3: A list of 25 NEs, their short description and statistics

works of art and the languages in which these were produced, are reflected in labelling the classes EVENTS (1,658), NORP (3,714), TIME (2,316), FACILITY (2,145), ART (2,407), and LANGUAGE (443).

Lastly, the comparatively frequent instances of the class DISEASE (1,272 NEs) in KazNERD may be explained by two interrelated factors. First, at the time of conducting the present study, the coronavirus disease 2019 (COVID-19) pandemic received massive public attention, which led to the source data often reflecting information on the outbreak of the disease across the country and worldwide. Second, the national media regularly discussed symptoms of various diseases similar to those observed in individuals infected with COVID-19, which resulted in the names of the diseases appearing in the source news reports.

4.2. Structure and Statistics

To allow reproducibility of the NER experiment between different research groups, KazNERD was split into three sets: training (80%), validation (10%), and test (10%). Table 4 provides statistical information on the number of tokens, sentences, and NEs in the dataset and per set. An evenly proportional distribution of sentence representations and NEs across the sets was ensured. We also saw to it that a sentence and its representations were only assigned to the same set. More detailed information on the numbers of NEs and sentence representations across the three sets can be found in Tables 5 and 6.

Furthermore, we extracted all unique NEs from KazNERD and computed the intersection between the training, validation, and test sets (see Figure 1). The total numbers of unique NEs in the training, validation, and test sets are 33,177, 6,547, and 6,742, respectively. We found that 42% of the unique NEs in the test set do not appear in the training and validation sets, which confirms its suitability for evaluating the generalisation capability of the NER models.

The three sets are stored in separate files, in the CoNLL-2002 format (Tjong Kim Sang, 2002)—that is, all files contain one token and the corresponding NE tag per line, with blank new lines representing sentence boundaries (see Table 2). Tokens and IOB2 tags are separated by a single space. Additionally, we provide variants of the

		Train	Valid	Test	Total
Sentences	#	90,228	11,167	11,307	112,702
	%	80.06	9.91	10.03	100
Tokens	#	1,043,305	129,223	129,824	1,302,352
	%	80.11	9.92	9.97	100
NEs	#	109,342	13,483	13,508	136,333
	%	80.20	9.89	9.91	100

Table 4: The statistics for the training, validation, and test sets of KazNERD

files containing identifiers heading each sentence, to allow for more nuanced studies requiring representation- and sentence-level detail. The sentence identifiers are formed by combining representation designations (i.e., AID, BID, CID, DID, EID, and FID) with a unique six-digit sentence number, for example, AID123456. Sentences with multiple representations have the same six-digit number but different designations, for example, AID111111 and BID111111.
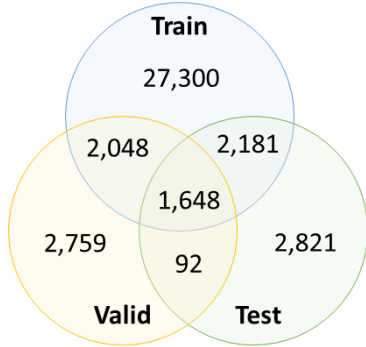

Figure 1: A Venn diagram depicting the intersection of unique NEs between the training, validation, and test sets of KazNERD

Entity classes	Train		Valid		Test	
	#	%	#	%	#	%
ADA	159	0.15	18	0.13	19	0.14
ART	1,953	1.79	225	1.67	229	1.70
CAR	23,550	21.54	2,886	21.40	2,824	20.91
CON	160	0.15	18	0.13	20	0.15
DATE	20,226	18.50	2,609	19.35	2,611	19.33
DIS	1,031	0.94	118	0.88	123	0.91
EVE	1,352	1.24	150	1.11	156	1.15
FAC	1,752	1.60	195	1.45	198	1.47
GPE	14,108	12.90	1,693	12.56	1,742	12.90
LAN	352	0.32	45	0.33	46	0.34
LAW	424	0.39	54	0.40	55	0.41
LOC	1,759	1.61	204	1.51	212	1.57
MIS	194	0.18	24	0.18	26	0.19
MON	3,678	3.36	441	3.27	441	3.26
NON	6	0.01	1	0.01	1	0.01
NORP	2,972	2.72	370	2.74	372	2.75
ORD	3,105	2.84	379	2.81	386	2.86
ORG	6,093	5.57	759	5.63	735	5.44
PERC	3,384	3.09	443	3.29	456	3.38
PER	10,893	9.96	1,352	10.03	1,332	9.86
POS	4,937	4.52	601	4.46	603	4.46
PROD	592	0.54	73	0.54	73	0.54
PROJ	1,694	1.55	206	1.53	211	1.56
QUA	3,094	2.83	407	3.02	407	3.01
TIME	1,874	1.71	212	1.57	230	1.70
Total	**109,342**	**100**	**13,483**	**100**	**13,508**	**100**

Table 5: The distribution of NEs across the training, validation, and test sets of KazNERD

Rep	Train		Valid		Test	
	#	%	#	%	#	%
AID	69,017	76.49	8,549	76.56	8,680	76.77
BID	19,236	21.32	2,368	21.21	2,365	20.92
CID	1,059	1.17	140	1.25	141	1.25
DID	644	0.71	81	0.73	84	0.74
EID	263	0.29	28	0.25	35	0.31
FID	9	0.01	1	0.01	2	0.02
Total	**90,228**	**100**	**11,167**	**100**	**11,307**	**100**

Table 6: The distribution of sentence representations across the training, validation, and test sets of KazNERD

5. NER Experiment

5.1. NER Methods

We applied several state-of-the-art machine learning methods to evaluate the KazNERD corpus. Detailed information on the NER model implementation and feature construction can be found in our GitHub repository[4].

CRF We applied the CRF models implemented by the CRFsuite toolkit (Okazaki, 2007). Specifically, we used the features derived from the surface forms of tokens, including target and context token prefixes, suffixes, and shape features. We note that the CRF models do not incorporate external linguistic resources, such as gazetteers, lookup tables, or word vector features.

BiLSTM-CNN-CRF We used the PyTorch implementation of a BiLSTM-CNN-CRF model (Ma and Hovy, 2016). The model combines the word embeddings with the character-level representations extracted using the CNN and feeds them into the BiLSTM module with the CRF output layer. Word embeddings are usually pre-trained on large unlabelled corpora, but, in the present study, we used randomly initialized embeddings.

BERT A pre-trained BERT model can be readily applied to the NER task, by reinitializing the output layer size to match the NE labels and fine-tuning the model on the NER data. We used the case-sensitive version of the multilingual BERT model within the Hugging Face Transformers framework (Wolf et al., 2020). The model consists of around 110M parameters and was pre-trained on 104 languages with the largest Wikipedia content, which includes the Kazakh language as well.

XLM-RoBERTa We also applied the XLM-RoBERTa model (Conneau et al., 2020), a multilingual version of RoBERTa (Liu et al., 2019), within the Hugging Face Transformers framework. Similar to BERT, it was adapted for the NER task, by reinitializing the output layer and fine-tuning. The rationale behind choosing the model lies in the fact that it has over five times as many parameters as BERT does (560M) and was pre-trained on CommonCrawl data containing 100 languages, Kazakh included.

5.2. Experimental Setup

The four NER models were trained on the training set. The hyperparameters were tuned on the validation set. The final, best-performing, model was evaluated on the test set. The deep learning-based models utilised a single V100 GPU on an NVIDIA DGX-2 machine.

Models	Valid			Test		
	Precision	Recall	F_1-score	Precision	Recall	F_1-score
CRF	93.62	91.93	92.77	93.20	91.63	92.41
BiLSTM-CNN-CRF	94.51	93.72	94.11	93.84	93.18	93.51
BERT	96.30	96.07	96.19	96.14	96.34	96.24
XLM-RoBERTa	**97.20**	**97.18**	**97.19**	**97.09**	**97.35**	**97.22**

Table 7: Experiment results of four NER models on the validation and test sets of KazNERD

The CRF model was run for 550 iterations using the L-BFGS training algorithm, with the L_1 and L_2 regularisation terms set to 0.1 and 0.01, respectively. The other hyperparameters were left at their default values of the CRFsuite toolkit.

For the BiLSTM-CNN-CRF model, we used a single BiLSTM layer with 256 hidden units and a CNN layer with 30 filters of size 3. The word and character embedding sizes were set to 100 and 30, respectively. We chose an initial learning rate of 0.005 and a batch size of 1,024. To prevent overfitting, a dropout rate of 0.5 was applied. The model was trained for 1,000 epochs using the Adam optimizer and the early stopping criteria based on the validation set, which yielded the highest score on epoch 432.

The BERT model was fine-tuned for 8 epochs, with the initial learning rate set to $5 \cdot 10^{-5}$ and the weight decay rate set to 10^{-4}. We set the batch size to 128 and applied 3,000 warmup steps. Likewise, the XLM-RoBERTa model was fine-tuned for 10 epochs, with the initial learning rate set to 10^{-5} and the weight decay rate set to 10^{-3}. We set the batch size to 64 and applied 800 warmup steps.

5.3. Evaluation Criteria

We evaluate NER performance in terms of exact match using precision, recall and F_1-score (Nadeau and Sekine, 2007) and the standard seqeval script (Nakayama, 2018), requiring that both the type and span of predicted NEs match the gold standard mention.

5.4. Experiment Results

Table 7 presents the performance of the NER models on the validation and test sets of KazNERD, measured by micro-averaging (Yang, 2001). The highest results were achieved by XLM-RoBERTa, followed by BERT, BiLSTM-CNN-CRF and CRF. Specifically, XLM-RoBERTa achieved relative improvements of 1%, 4%, and 5% over BERT, BiLSTM-CNN-CRF, and CRF, respectively. In general, all the NER models performed well, achieving precision, recall, and F_1-scores of above 90%, highlighting the utility of our annotated dataset for the Kazakh NER task. The results of the XLM-RoBERTa model for different NEs are shown in Table 8 and will be discussed in the following section.

6. Discussion

The performance of XML-RoBERTa was above 95% for 14 NE classes and in the range of 85% to 95% for eight classes. Only three classes were predicted with an F_1-score below 85%. The model yielded almost perfect results for MONEY (99.89%) and PERSON (99.36%). This could be explained by the composition of these classes. The extent of MONEY NEs includes a monetary value and an

Entity Classes	Precision	Recall	F_1-score
ADA	83.33	52.63	64.52
ART	97.83	98.25	98.04
CAR	98.48	98.90	98.69
CON	89.47	85.00	87.18
DATE	97.49	98.01	97.75
DIS	90.84	96.75	93.70
EVE	87.27	92.31	89.72
FAC	79.21	80.81	80.00
GPE	98.38	97.59	97.98
LAN	95.74	97.83	96.77
LAW	87.04	85.45	86.24
LOC	91.63	87.74	89.64
MIS	96.15	96.15	96.15
MON	99.77	100.00	99.89
NON	0.00	0.00	0.00
NORP	98.92	98.92	98.92
ORD	97.39	96.63	97.01
ORG	91.84	93.47	92.65
PERC	98.68	98.68	98.68
PER	99.55	99.17	99.36
POS	96.29	99.00	97.63
PROD	88.89	87.67	88.28
PROJ	93.81	93.36	93.59
QUA	97.30	97.30	97.30
TIME	98.69	98.26	98.47
Micro ave.	**97.09**	**97.35**	**97.22**
Macro ave.	**90.16**	**89.20**	**89.53**

Table 8: XLM-RoBERTa performance for different entity classes of the test set

explicit monetary unit (e.g., 50 dollars). This must have made it easier for the model to recognize the class, for monetary units in KazNERD are not very diverse, with "tenge" (the local currency), "dollar", and "euro" making frequent appearances. Likewise, in Kazakh, PERSON NEs often appear as a combination of first and last names, with both capitalised and the latter normally ending in –*ов(а)* "-ov(a)", -*ев(а)* "-ev(a)", -*ин(а)* "-in(a)". These features presumably enabled the model to achieve high prediction accuracy for the class.

The low F_1-scores for NON_HUMAN (0%) and ADAGE (64.52%) on the test set could be due to the apparent insufficiency of instances of the former in the dataset and the form variability of the latter. Increasing the number of NON_HUMAN NEs in the training sample, by expanding

the dataset to embrace domains where the use of names of non-humans is expected (e.g., science fiction, children's stories, or animal fantasies) will likely improve the accuracy of the model. As for ADAGE NEs, they are generally easy to recognise in context thanks to their form fixedness (e.g., *No smoke without fire*). Lexical and grammatical variations of proverbs and sayings are possible (e.g., *There is no smoke without fire* or *Where there is smoke, there is fire*), but still unlikely to preclude humans from continuing to identify these: such phrases bear greater psychological and social significance than do other set expressions (Norrick, 2015). However, this can hardly apply to a machine learning model, which will struggle to decide whether a given expression is a pre-existent variation of a known adage, its nonce restructuring, or not an adage at all, especially if there is inadequacy of data to make inferences from. As mentioned earlier, the class ADAGE was labelled as a result of our scientific curiosity, and further review and investigation as to the worth of this class for the NER task is required.

Since the present study was, to the best of our knowledge, the first to develop a publicly available annotated corpus as well as guidelines in Kazakh for 25 NE classes, it was subject to several challenges. Firstly, although NER generally implies the recognition of proper nouns in text, which are expected to be capitalised given their designation of names of persons, places, organisations and so forth, some Kazakh nouns assigned to certain NE classes in our dataset do not seem to meet this criterion. For example, the NEs *дүйсенбі* "Monday" (DATE), *христиандар* "Christians" (NORP), or *ағылшын тілі* "English" (LANGUAGE) to name a few, are normally lower-cased in Kazakh, unless they appear at the beginning of a sentence. Further studies on Kazakh NER taking such cases into account need to be undertaken.

NE coordination posed another problem. The challenge concerns whether two (or more) coordinated NEs, for example, *Олжас пен Аина Қорғанбек* "Olzhas and Aina Qorganbek" (the names of a husband and a wife followed by their family name; PERSON) or *Байтұрсынов пен Қона-ев көшелерінде* "on Baitursynov and Qonayev Streets" (the names of two local streets followed by the word "streets"; FACILITY) ought to be labelled as a single NE or two separate NEs. Although MUC-6 (Grishman and Sundheim, 1996) originally advocated the separate use of annotations, in KazNERD, it was decided to label coordinated NEs as a single entity in accordance with the recommendations of MUC-7 (Chinchor, 1998), promoting joint annotation.

Another similar issue was related to nested entities: for example, should the expression *Қазақстан Президен-ті* "The President of Kazakhstan" be considered two entities *Қазақстан* (Kazakhstan, GPE) and *Президенті* (President, POSITION) or a single entity *Қазақстан Пре-зиденті* (The President of Kazakhstan, POSITION)? Here again, our decision was guided by MUC-7, encouraging the annotation of such expressions as a single NE. Thus, while developing KazNERD, we chose not to decompose compound entities and not to label subentities. However, future research into Kazakh NER should consider these challenges, with the decision as to which of the approaches is more likely to cover the needs of application areas left to the discretion of those concerned.

As regards challenges related to metonymy (i.e., the use of the name of something to refer to that of something else that is closely associated with it, as in *Downing Street* to refer to the British Prime Minister), consistent with the MUC recommendations, KazNERD generally retains the semantics of common NEs, unless otherwise specified in the developed annotated guidelines. Thus, in *Абайды тану* "cognising Abai" (the name of a great Kazakh poet), the NE *Абайды* is tagged as PERSON, despite the contextual reference to the person's literary works (the NE class ART). This should certainly be borne in mind by enthusiasts willing to make use of KazNERD.

Similarly, challenges presented by the ambiguity between the classes ORGANISATION and FACILITY may presumably account for the comparatively low F_1-score for the latter. In the annotation guidelines, we recommend that, in cases of confusion, preference should be given to ORGANISATION when actions normally characteristic of persons (e.g., *say*, *state*, *report* etc.) are used with names of institutions or if a building houses an institution of the same name, unless explicitly referring to the physical structure alone in a locative manner. Yet, in cases where the distinction is still not clear-cut, such as *Президент ... Ақордада арнайы кеңес өткізді* "President ... held a special meeting in Akorda" (the official workplace of the President of Kazakhstan), we annotated *Ақордада* as ORGANISATION in line with the existing guidelines tagging *White House* or *Kremlin* as ORGANISATION, in spite of the contextual reference to the facility.

7. Conclusion

The present study set out to develop the first publicly available annotated dataset for Kazakh NER. The resulting dataset, KazNERD, contains 112,702 sentences from the television news domain and 136,333 annotations for 25 entity classes. All NEs were labelled using the IOB2 scheme by two native Kazakh speakers under the supervision of the first author, in accordance with the annotation guidelines specially designed in and for the Kazakh language. To automate Kazakh NER, state-of-the-art machine learning models were built, with the best-performing model yielding an exact match F_1-score of 97.22% on the test set. In the future, we aim to focus on developing fine-grained and domain-independent NER models to ensure their external validity. To this end, we intend to train the models on a version of KazNERD supplemented with annotated data from different domains and genres, including transcribed conversations from television and radio shows, podcasts, phone talks, fiction, and senate speeches.

The annotated dataset, guidelines, and codes used in training the models can be freely downloaded under the CC BY 4.0 licence from https://github.com/IS2AI/KazNERD.

8. Acknowledgements

9. Bibliographical References

Akhmed-Zaki, D., Mansurova, M., Barakhnin, V., Kubis, M., Chikibayeva, D., and Kyrgyzbayeva, M. (2020). Development of Kazakh named entity recognition models. In *International Conference on Computational Collective Intelligence (ICCCI)*, volume 12496 of *Lecture Notes in Computer Science*, pages 697–708. Springer.

Ali, W., Lu, J., and Xu, Z. (2020). SiNER: A large dataset for Sindhi named entity recognition. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 2953–2961. European Language Resources Association.

Aliod, D. M., van Zaanen, M., and Smith, D. (2006). Named entity recognition for question answering. In *Proceedings of the Australasian Language Technology Workshop (ALTA)*, pages 51–58. ALTA.

Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Babych, B. and Hartley, A. (2003). Improving machine translation quality with automatic named entity recognition. In *Proceedings of the International EAMT Workshop on MT and Other Language Technology Tools*, pages 1–8.

Brunstein, A. (2002). Annotation guidelines for answer types. *LDC2005T33, Linguistic Data Consortium, Philadelphia*.

Cheng, P. and Erk, K. (2020). Attending to entities for better text understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7554–7561.

Chinchor, N. A. (1998). Overview of MUC-7. In *Proceedings of the Message Understanding Conference (MUC)*. ACL.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 8440–8451. Association for Computational Linguistics.

Dave, B. (2007). *Kazakhstan: Ethnicity, language and power*. Routledge.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186. ACL.

Dumitrescu, S. D. and Avram, A. (2020). Introducing RONEC - the Romanian named entity corpus. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 4436–4443. European Language Resources Association.

Etzioni, O., Cafarella, M., Downey, D., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S., and Yates, A. (2005). Unsupervised named-entity extraction from the Web: An experimental study. *Artificial Intelligence*, 165(1):91–134.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Friedman, N., Geiger, D., and Goldszmidt, M. (1997). Bayesian Network Classifiers. *Machine Learning*, 29(2-3):131–163.

Grishman, R. and Sundheim, B. (1996). Message understanding conference-6: A brief history. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 466–471.

Han, X., Ben, K., and Zhang, X. (2020). Research on Named Entity Recognition Technology in Military Software Testing. *Journal of Frontiers of Computer Science and Technology*, 14(5):740–748.

Ho, T. K. (1995). Random decision forests. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 278–282. IEEE.

Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Ingólfsdóttir, S. L., Porsteinsson, S., and Loftsson, H. (2019). Towards high accuracy named entity recognition for Icelandic. In *Proceedings of the Nordic Conference on Computational Linguistics*, pages 363–369. Linköping University Electronic Press.

Khassanov, Y., Mussakhojayeva, S., Mirzakhmetov, A., Adiyev, A., Nurpeiissov, M., and Varol, H. A. (2021). A crowdsourced open-source Kazakh speech corpus and initial speech recognition baseline. In *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 697–706. ACL.

Kozhirbayev, Z. M. and Yessenbayev, Z. A. (2020). Распознавание именованных объектов для казахского языка. *Journal of Mathematics, Mechanics and Computer Science*, 107(3):57–66.

Kuralbayev, A., Mukhsimbayev, B., Bekbaganbetov, A., and Fuad, H. (2020). Named Entity Recognition Algorithms Comparison for Judicial Text Data. In *Proceedings of the IEEE International Conference on Application of Information and Communication Technologies (AICT)*, pages 1–5. IEEE.

Lafferty, J., McCallum, A., and Pereira, F. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers.

Linguistic Data Consortium. (2008). ACE (Automatic Content Extraction) English Annotation Guidelines for Entities Version 6.6 2008.06.13.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ma, X. and Hovy, E. H. (2016). End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the Association for Computational Linguistics (ACL)*. ACL.

Maynard, D., Bontcheva, K., and Cunningham, H. (2003). Towards a Semantic Extraction of Named Entities.

Proceedings of Recent Advances in Natural Language Processing (RANLP), pages 255–261.

Mussakhojayeva, S., Janaliyeva, A., Mirzakhmetov, A., Khassanov, Y., and Varol, H. A. (2021a). KazakhTTS: An Open-Source Kazakh Text-to-Speech Synthesis Dataset. In *Interspeech*, pages 2786–2790.

Mussakhojayeva, S., Khassanov, Y., and Varol, H. A. (2021b). A study of multilingual end-to-end speech recognition for Kazakh, Russian, and English. In *International Conference on Speech and Computer (SPECOM)*, volume 12997 of *Lecture Notes in Computer Science*, pages 448–459. Springer.

Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Nakayama, H. (2018). seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

Neves, M. and Seva, J. (2021). An extensive review of tools for manual annotation of documents. *Briefings in Bioinformatics*, 22(1):146–163.

Norrick, N. R., (2015). *1 Subject Area, Terminology, Proverb Definitions, Proverb Features*, pages 7–27. De Gruyter Open Poland.

Okazaki, N. (2007). CRFsuite: a fast implementation of Conditional Random Fields. *Software Package*.

Pavlenko, A. (2008). Russian in post-Soviet countries. *Russian Linguistics*, 32(1):59–80.

Raytheon BBN Technologies. (2004). OntoNotes Named Entity Guidelines Version 14.0.

Sadykova, Z. N. and Ivanov, V. V. (2016). Формирование корпуса с разметкой сущностей в новостных медиа ресурсах для казахского языка [The formation of a corpus with the markup of entities in news media resources for the Kazakh language]. In *TEL-2016*, pages 137–141.

Sang, E. F. T. K. and Meulder, F. D. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Conference on Natural Language Learning (CoNLL) at HLT-NAACL*, pages 142–147. ACL.

Sang, E. F. T. K. and Veenstra, J. (1999). Representing text chunks. In *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 173–179. ACL.

Schön, S., Mironova, V., Gabryszak, A., and Hennig, L. (2019). A corpus study and annotation schema for named entity recognition and relation extraction of business products. *Proceedings of the Language Resources and Evaluation (LREC)*, pages 4445–4451.

Stenetorp, P., Pyysalo, S., Topic, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). BRAT: A web-based tool for nlp-assisted text annotation. In *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 102–107. ACL.

Tjong Kim Sang, E. F. (2002). Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *Proceedings of the Conference on Natural Language Learning (CoNLL)*.

Tolegen, G., Toleu, A., and Xiaoqing, Z. (2016). Named entity recognition for Kazakh using conditional random fields. In *Proceedings of the International Conference on Computer Processing of Turkic Languages*, pages 1–8.

Tolegen, G., Toleu, A., Mamyrbayev, O., and Mussabayev, R. (2020). Neural Named Entity Recognition for Kazakh. *arXiv preprint arXiv:2007.13626*.

Weischedel, R., Palmer, M., Marcus, M., Hovy, E., Pradhan, S., Ramshaw, L., Xue, N., Taylor, A., Kaufman, J., and Franchini, M. (2012). OntoNotes release 5.0 with OntoNotes DB Tool v0.999 beta.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP)*, pages 38–45. ACL.

Yang, Y. (2001). A study of thresholding strategies for text categorization. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 137–145.

Yimam, S. M., Gurevych, I., de Castilho, R. E., and Biemann, C. (2013). WebAnno: A flexible, web-based and visually supported system for distributed annotations. In *Association for Computational Linguistics (ACL)*, pages 1–6. ACL.