KazNERD: Kazakh Named Entity Recognition Dataset

rustem.yeshpanov@nu.edu.kz yerbolat.khassanov@nu.edu.kz ahvarol@nu.edu.kz
Institute of Smart Systems and Artificial Intelligence

LREC 2022 Marseille

Introduction

Named Entities (NEs):
- names of persons
- names of organisations
- names of geographical locations
- numerical expressions
- temporal expressions etc.

Named Entity Recognition (NER):
- NE identification in semi-/unstructured texts
- NE classification into pre-specified types

Kazakh:
- Turkic language
- earliest NER research in 2016

Annotated Dataset:
- largest annotated dataset for Kazakh
- 112,702 sentences from TV news
- 136,333 annotations
- publicly available and free for use

Annotation Guidelines:
- first annotation guidelines for and in Kazakh
- rules for and examples of 25 NE types
- publicly available and free for use

NER Models:
- 4 models
- XLM-RoBERTa: F_1-score = 97.22%
- publicly available and free for use

Related Work

Kazakh NER research (2016–2020):
- unclear breakdown of annotated NEs
- occasional mention of annotation guidelines
- lack of IAA[1] assessment
- no free access to annotated datasets

KazNERD Construction

Source Data:
- TV news
- 86,246 sentences

Sentence Representations:
- 6 (AID, BID, CID, DID, EID, FID)
- 112,702 sentences

Annotation Scheme:
- IOB2 = BIO[2]

Annotation Guidelines:
- user-friendly
- last updated: 25 April, 2022

Annotation Workflow:
- 2 native Kazakh linguists + 1 supervisor
- 2-week training period
- Webanno annotation tool
- 1,500 sentences per day for 6 months
- IAA = 0.95–0.97 Fleiss' kappa

KazNERD Specifications

Annotated NEs:
- 25 NE types

ADAGE | ART | CARDINAL | CONTACT | DATE | DISEASE | EVENT | FACILITY | GPE | LANGUAGE | LAW | LOCATION | MISCELLANEOUS | MONEY | NON_HUMAN | NORP | ORDINAL | ORGANISATION | PERCENTAGE | PERSON | POSITION | PRODUCT | PROJECT | QUANTITY | TIME

- 136,333 annotations
- CARDINAL, DATE, GPE ⇧
- CONTACT, ADAGE, NON_HUMAN ⇩

Structure:
- training set: 80% (33,177 unique NEs)
- validation set: 10% (6,547 unique NEs)
- test set: 10% (6,742 unique NEs)
- CONLL-2002 files

Experiment & Results

Evaluation Criteria:
- precision & recall
- F_1-score

Models (F_1):
- CRF[3]: 92.41%
- BiLSTM[4]-CNN[5]-CRF: 93.51%
- BERT[6]: 96.24%
- XLM-RoBERTa: 97.22%

NEs (⇧ & ⇩):
- MONEY: 99.89%
- PERSON: 99.36%
- ADAGE: 64.52%
- NON_HUMAN: 0%

NEs (F_1):
- 14 NEs: $F_1 > 95\%$
- 8 NEs: $85\% < F_1 < 95\%$
- 3 NEs: $F_1 < 85\%$

Challenges

- lower-cased NEs
- coordinated NEs
- nested NEs
- metonymy
- NE class ambiguity

Future Work

- fine-grained models
- domain-independent models
- various domains & genres

Freely downloadable from

CC BY

https://github.com/IS2AI/KazNERD

issai.nu.edu.kz
issai@nu.edu.kz

[1] Inter-Annotator Agreement
[2] Beginning, Inside, Outside
[3] Conditional Random Field
[4] Bidirectional Long Short-Term Memory
[5] Concurrent Neural Network
[6] Bidirectional Encoder Representations from Transformers