NAZARBAYEV UNIVERSITY
NATIONAL
LABORATORY ASTANA

Kazakh text normalization using machine translation approaches

Zhanibek Kozhirbayev

Zhandos Yessenbayev

KazNLP: a pipeline for automated processing of texts written in Kazakh language

- The **goal** of the project is to design free, open source programming tools for automated processing of texts written in Kazakh language.
- The following **objectives** are defined in the framework of the project:
    1. developing the initial normalization module;
    2. developing the sentence-word tokenizer;
    3. developing the language identification module;
    4. developing the morphological analyzer;
    5. developing the morphological tagger;
    6. developing the syntactic parser;
    7. developing the spelling checking and correction module;
    8. developing the named entity recognition module;
    9. **developing the secondary normalization module.**

All the modules are implemented in Python.

NOISES in UGC:

Text normalization is the transformation of text into a canonical form and usually useful for further processing.

User generated content (UGC) generally refers to any type of content, i.e. photo, video, audio, text, created by Internet users:

– **spontaneous transliteration**, e.g. Kazakh word “біз” can be spelled in three additional ways: “быз”, “биз”, and “biz”;

– **use of homoglyphs**, e.g. Cyrillic letter “і” (U+0456) can be replaced with Latin homoglyph “i” (U+0069);

– **code switching,** use of Russian words and expressions in Kazakh text and vice versa;

– **word transformations**, e.g. “керемееет”,“ крмт” instead of “керемет” (great), or seg-mentation of words, e.g. “к-е-р-е-м-е-т”;

– **the use of emoji**, e.g. (☺, ☹), and their symbolic counterparts, e.g. [:), : (].

Data collection and annotation

news portals:

NUR.KZ

zakon.kz
сетевое издание

social media (facebook groups):

Kaspi.kz

Qazaqstan online

ҚАЛА & МЕН ДАЛА

STAN.KZ

Data collection and annotation

Total		Stripped of perfect comments		After splitting long comments		Ideal comments	
doc	tok	doc	tok	doc	tok	doc	tok
17181	237092	12896	192853	19799	192853	4285	44239

Table 1. Data set statistics from news portals.

Source	Number of posts	Number of comments
OnlineQazaqstan	17	3287
Newspaper «Қала мен Дала»	18	1490
Kaspi.kz	8	1897
Stan.kz	29	3340
Total	72	10 014

Table 2. Social media dataset statistics.

Parallel comments	Train set	Test set
27005	24 305	2700

Table 3. Final data statistics.

Method description

- statistical machine translation (SMT)
- neural machine translation (NMT)

Pipeline (phrase-based SMT):

- Moses tool
- n-gram language models (3-gram models).
- decoding process was implemented using the beam search stack decoding algorithm.

Pipeline (word-based NMT):

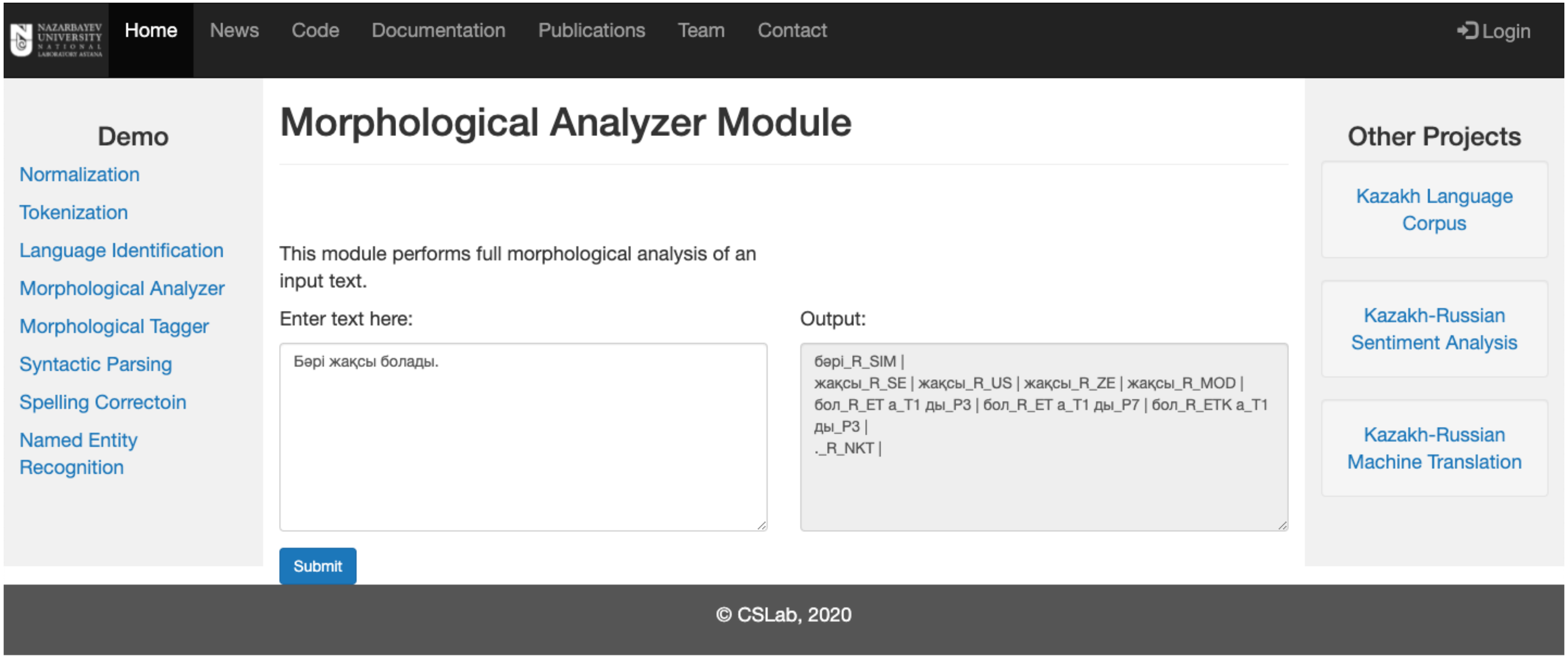
- Seq2Seq model using the Keras library
- 2-layer LSTM encoders and decoders
- trained using the efficient Adam approach to stochastic gradient descent and minimizes the categorical loss function

Experiment results

Model	BLEU score
SMT	21.67
NMT	29.74

Project Repository and Website

- Repository: https://github.com/nlacslab/kaznlp
- Website: https://opendev.kz/kaznlp/



Morphological Analyzer Module

This module performs full morphological analysis of an input text.

Enter text here:

Бәрі жақсы болады.

Output:

```
бәрі_R_SIM |
жақсы_R_SE | жақсы_R_US | жақсы_R_ZE | жақсы_R_MOD |
бол_R_ET а_T1 ды_P3 | бол_R_ET а_T1 ды_P7 | бол_R_ETK а_T1
ды_P3 |
._R_NKT |
```

© CSLab, 2020

Thanks for your attention

Any questions?