# Text Normalization and Spelling Correction in Kazakh Language

Gaukhar Slamova[1] and Meruyert Mukhanova[1]

[1] Suleyman Demirel University, Engineering and Natural Sciences, Information Systems, 040900, Kaskelen, Almaty, Kazakhstan
{150103077, 150103018}@stu.sdu.edu.kz

**Abstract.** Text normalization is significant step in preprocessing of informal, social media and short texts in the Natural Language Processing (NLP) tasks. Researches in the field are mostly on English, but not on the agglutinative languages such as Kazakh, Korean, Japanese, which are determined as morphologically rich languages, and complex compared to English. In this paper, we present text normalization and auto correction of words for Kazakh language, we convert informal text into grammatically correct form. To do the auto correction task, firstly we countered keyboard error while typing words, then choose the best match from them. Additionally, we categorized words to several groups and separated text into modules of words. The exact match score of the overall system on the provided datasets are 85.40 per cent.

## 1 Introduction

Text normalization is the task of transforming informal writing into its standard form in the language. It is an important processing step for a wide range of Natural Language Processing (NLP) tasks such as text-to-speech synthesis, speech recognition, information extraction, parsing, and machine translation. (Richard Sproat, Alan W. Black, Stanley F. Chen, Shankar Kumar, Mari Ostendorf, Christopher Richards, 2001) Text normalization involves merging different written forms of token into a canonical normalized form; for example, a document may contain the equivalent tokens “Mr.”, “Mr”, “mister”, and “Mister” that would all be normalized to a single form (Nitin Indurkhya, Fred J. Damerau, 2010).

Normalization poses multiple challenges, as we know it is a task of mapping all out-of-vocabulary non-standard word tokens to in-vocabulary standard forms, to deal with it we should convert raw text into grammatically correct sentence by modifying punctuation and capitalization, and adding, removing, or reordering words. Also, we gave specific values to some types as date, phone, currency, URL, etc. On informal texts as usual a lot of mistakes, it is useful to correct them. To spelling correction task, we consider keyboard typing mistakes, character repetition and other tools. In this paper, we propose spelling correction and text preprocessing by mentioned above techniques, it gives higher precision accuracy than other methodologies.

The rest of this paper is organized as follows. In Section 2 we discuss previous approaches to the normalization problem. Section 3 presents our normalization framework, including the actual normalization and learning procedures. In Section 4 we

introduce evaluation metric, and present experimental results of our model with respect to several categories. Finally, we conclude in Section 5.

## 2 Related Work

Early studies of text normalization include machine learning approach in text-to-speech and social media, and with usage of neural network in it. In this paper, we use similar method as in works which investigated text normalization in social media, because of recent rise heavily informal writing in messaging applications, text normalization is a huge problem of every language.

Previous works handled text normalization process by producing noisy text where normalized text go through a noisy channel; this approach called noisy channel model. (Moore, Eric Brill and Robert C., 2000) presented a method for modelling the spelling correction as a noisy channel model based on string to string edits; this model gives significant improvements compared to early studies. (Kristina Toutanova and Robert C. Moore, 2002) enhanced the string to string edits model by modelling pronunciation similarities between words achieved a substantial performance improvement over the previous best performing models for spelling correction. (Monojit Choudhury, Rahul Saraf, Vijit Jain, Animesh Mukherjee, Sudeshna Sarkar, and Anupam Basu, 2007) introduced a supervised HMM channel model which adopted the spellchecking metaphor based on character-level edit which has been extended by (Paul Cook and Suzanne Stevenson, 2009) who used unsupervised noisy channel model using probabilistic models for common abbreviation and various spelling errors types. (Kobus Catherine, François Yvon, and Géraldine, 2008) presented French SMS messages normalization process by normalizing the orthography with combination of Statistical Machine Translation and automatic speech recognition approaches. (Bo Han and Timothy Baldwin, 2011) presented model for identifying and normalizing ill-formed words, generating correction candidates based on morphophonemic similarity over SMS corpus and Twitter. (Joseph Kaufmann and Jugal Kalita, 2010) used a machine translation approach with a pre-processor for syntactic normalization rather than lexical. (Liu, Deana Pennell and Yang, 2011) presented two-phase method for expanding abbreviations using a machine translation system trained at the character level during the first phase and in the second phase utilizing an in-domain language model, in the context of neighbouring words. (Fei Liu, Fuliang Weng, and Xiao Jiang, 2012) proposed a cognitively-driven normalization system that integrates different human perspectives in normalizing the nonstandard tokens, including the enhanced letter transformation, visual priming, and string/phonetic similarity.

There are fewer studies done on the agglutinative language comparing to English, (Gülşen Eryiğit, Dilara Torunoğlu-Selamet, 2017) introduced social media text normalization for Turkish by analyzing Web 2.0 Turkish texts, categorizing them into seven types and providing candidate spelling correction words. (Mohammad Saloot, Norisma Idris, Rohana Mahmud, 2014) propose an approach to normalize the Malay Twitter messages based on corpus-driven analysis. (Panchapagesan Krishnamurthy, P.P. Talukdar, N Sridhar, A.G. Ramakrishnan, 2004) introduced a novel approach to text normalization, wherein tokenization and initial token classification are combined into one stage followed by a second level of token sense disambiguation, is described.

(O. De Clercq, B. Desmet, S. Schulz, E. Lefever, V. Hoste, 2013) used multimodule approach which rely on Machine Translation and transliteration-based system for social media messages in the Dutch language. Agglutinative languages tend to have longer words than fusional ones (Steffen Eger et al., 2016) and spelling correction model would be complex, because of the morphology.

To our knowledge, the work presented here is the first which observed normalization in Kazakh language with the usage of auto correction methodology and value categorization.

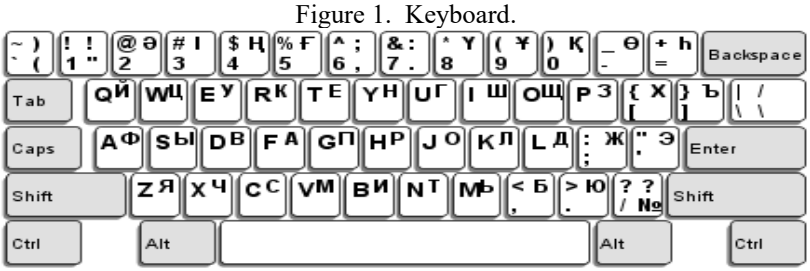
## 3 Evaluation

In this section we introduce our normalization framework, which consider both spelling correction and text preprocessing processes. Morphologically rich languages such as Kazakh, Korean, Finnish, Arabic, Turkish, etc. are considered as highly inflectional; their characteristic is that one stem in these languages may have hundreds of possible forms.

### 3.1 Spelling Correction

Spelling errors are categorized into two classes: typographic and cognitive. Cognitive errors phonetic or orthographic similarity of words; person does not know how to spell a word. Typographic errors are related to the keyboard and hand/finger movement where spelling errors happen because of two letters keys' closeness on the keyboard. (Kukich, 1992)

Figure 1. Keyboard.




In the Figure 1, on the upper-right corner are shown Kazakh language letters. On Kazakh alphabet there are 42 letters, where 9 vowels.

**Table 1.** Vowels and Consonants in Kazakh Language.

| Form | English |
|---|---|
| Vowels | а, ә, е, о, ө, ұ, ү, ы, і |
| Consonants | б, г, ғ, д, ж, з, й, к, қ, л, м, н, ң, п, р, с, т, х, h, ш |

To spelling correction Spelling errors have been classified into four types: Deletion, Insertion, Substitution and Transposition. (Damerau, 1964) Deletion errors where characters are repeated, as in қаты→қатты, is observed significantly more

frequently than in a non-repeating context showing that visually conspicuous errors tend to be corrected. Substitution errors of visually similar characters (e.g., аға→ағa) are in fact very common. (Yukino Baba, Hisami Suzuki, 2012)

We make correction within four parts:

- Selection Mechanism – choose candidate with the highest probability
- Candidate model – gives candidate for the given word.
- Language model – probability of the candidates acquireness on the text
- Error model – probability that another word was typed when author mean exact word.

When we trying to find most likely correct candidate (x) to word out of all possible candidates that has maximum probability to intended correction to given word, w:

$$x = argmax_{x \in candidates} \mathrm{P(x|w)}$$

By Bayes' Theorem it is equivalent to:

$$argmax_{x \in candidates} \frac{P(w|x)P(x)}{P(w)}$$

Since P(*w*) is the same for every possible candidate *c*, we can factor it out, giving:

$$argmax_{x \in candidates} \mathrm{P(w|x)P(x)}$$

Consider the misspelled word "сенің" and the two candidates "сенім" and "сенің". Correction candidate "сенің" seems good because words look similar and only change is "ң" to "н", it is an accusative case of noun. On the other hand, "сенім" is a very common word and a noun, this is the correct spelling of word. The point is that to estimate P(*x*|*w*) we consider both the probability of candidate and the probability of the change from *x* to *w*.

### 3.1 Replacement rules

Kazakh is morphologically rich language; one stem has a very large number of word forms. It is not efficient to use a lexicon lookup for storing and checking all possible candidates of word forms in the dataset. But morphological analyzer helps to find all possible word forms, lemmas, and inflectional or derivational structures.

Kazakh is generally verb-final, though various permutations on subject–object–verb word order can be used. Inflectional and derivational morphology, both verbal and nominal, in Kazakh, exists almost exclusively in the form of agglutinative suffixes. Kazakh is a nominative-accusative, head-final, left-branching, dependent-marking language. (Mukhamedova, Raikhangul, 2015)

**Table 2.** Declension of Words.

| Case | Possible Forms | шелек "*bucket*" | кеме "*ship*" | бас "*head*" | тұз "*salt*" |
|---|---|---|---|---|---|
| **Nom** | — | шелек | кеме | бас | тұз |
| **Acc** | -ні, -ны, -ді, -ды, -ті, -ты, -н | шелек**ті** | кеме**ні** | бас**ты** | Тұз**ды** |
| **Gen** | -нің, -ның, -дің, -дың, -тің, -тың | шелек**тің** | кеме**нің** | бас**тың** | тұз**дың** |
| **Dat** | -ге, -ға, -ке, -қа, -не, -на | Шелек**ке** | кеме**ге** | бас**қа** | тұз**ға** |
| **Loc** | -де, -да, -те, -та | Шелек**те** | кеме**де** | бас**та** | тұз**да** |
| **Abl** | -ден, -дан, -тен, -тан, -нен, -нан | шелек**тен** | кеме**ден** | бас**тан** | тұз**дан** |
| **Inst** | -мен(ен) -бен(ен) -пен(ен) | шелек**пен** | кеме**мен** | бас**пен** | тұз**бен** |

( Zitouni and R. Sarikaya, 2009) list the below problems related to issue with agglutinative languages:

- Increase in dictionary size;
- Poor language model probability estimation;
- Higher out-of-vocabulary rate;
- Inflection gap for machine translation

**Table 3.** Some form for the Kazakh word ‘Кітап’.

| Word form | English |
|---|---|
| Кітап | Book |
| Кітаптар | Books |
| Кітаптағы | In the book |
| Кітаптың | Of the book |
| Кітапқа | To the book |
| Кітапта | At the book |
| Кітаптан | From the book |
| Кітаппен | With the book |
| Кітап | Book |

We make candidate generation for the nonstandard word forms. In informal texts mostly used slangs, abbreviations, character repetitions, logograms, wrong letter cases, spelling errors related to pronunciation, vowels misspelling errors. To normalize such words, we make following candidate generation layer:

- Letter case transformations;

- Accent normalization
- Spelling correction

Replacement rules considered as a regular expression pattern and used for handling with character repetitions, emails, URLs, etc. Following word types tagged by the specific labels:

- E-mails: labeled as @email[example@gmail.com]
- URLs: labeled as @URL[http://sdu.edu.kz]
- Emoticons: labeled as @emoji[>3]
- Money: labeled as @money[$500]
- Date: labeled as @date[25.02.2018]
- Phone: labeled as @phone[87772349134]

Texts contain different word cases: uppercase, lowercase and mixed case. We converted: uppercase words into first letter upper remaining letters lower, if the word length less than five; lowercase word remains the same; and mixed case word into first letter upper remaining letters lower.

## 4 Evaluation

We performed evaluation for both word spelling correction and replacement rules. For the training dataset we used most popular and valuable novels of Kazakh literature written by Mukhtar Auezov “Abai Zholy” (The path of Abai) which consists of 16893 words.

**Table 4.** Examples of word correction.

| Misspelling | Correct | Guess |
|---|---|---|
| Қыздартың | Қыздардың | Қыздардың |
| Сагыз | Сағыз | Сағыз |
| Атам | Адам | Атам |
| Сенің | Сенім | Сенің |
| Сагыніш | Сағыныш | Сагыныш |
| Жанын | Жаным | Жанын |

**Table 5.** Text Normalization results.

| System | Accuracy, per cent |
|---|---|
| Keyboard correction | 90.7 |
| Replacement | 80.1 |
| Total | 85.40 |

As shown in Table 5, spelling correction with the usage of keyboard model errors gave higher accuracy than word replacement to find normalized form of value. Noisy non-standard words correction not inserts words into the dataset, it generates best fit candidate to the misspelling word. We made testing to 500 words, constructed testing

dataset according to words from “Abai Zholy”. Instead of using lexicon lookup, we propose to use keyboard model for Kazakh language.

## 4 Conclusions

NLP is the recent field of science in the Kazakhstan, there is a lack of tools for preprocessing and spelling correction. In this research, we aimed to explore the necessary components for text normalization of a morphologically rich language, Kazakh, for the further studies related to this field.

In this article, we suggested to use social media and messaging normalization technique for Kazakh language. We hope to have provided a better insight into spelling correction by the keyboard usage in Kazakh alphabet which contains 42 letters 16 characters more than English.

## Acknowledgements

We thank the anonymous reviewers for helpful comments and suggestions. We also thank Kessikbayeva Gulshat for her comments on a preliminary version of this work.

## References

Zitouni and R. Sarikaya. (2009). *Arabic diacritic restoration approach based on maximum entropy models.* London, UK: Computer Speech & Language.

Bo Han and Timothy Baldwin. (2011). *Lexical Normalisation of Short Text Messages: Makn Sens a #twitter.* Portland, Oregon, USA: Proceedings of ACL-HLT.

Damerau, F. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 659-664.

Fei Liu, Fuliang Weng, and Xiao Jiang. (2012). A broad-coverage normalization system for social media language. *ACL*, 1035–1044.

Gülşen Eryiğit, Dilara Torunoğlu-Selamet. (2017). Social media text normalization for Turkish. *Natural Language Engineering*, 835-875.

Joseph Kaufmann and Jugal Kalita. (2010). *Syntactic normalization of Twitter messages.* Kharagpur, India: International Conference on Natural Language Processing.

Kobus Catherine, François Yvon, and Géraldine. (2008). Transcrire les SMS comme on reconnaît la parole. *Actes de la Conférence sur le Traitement Automatique des Langues* (pp. 128–138). Avignon, France: TALN’08.

Kristina Toutanova and Robert C. Moore. (2002). *Pronunciation modeling for improved spelling correction.* Philadelphia, USA: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL.

Kukich, K. (1992). Techniques for automatically correcting. *ACM Computing Surveys*, 24(4).

Liu, Deana Pennell and Yang. (2011). A character-level machine translation approach for normalization of SMS abbreviations. *IJCNLP*, 974–982.

Mohammad Saloot, Norisma Idris, Rohana Mahmud. (2014). An architecture for Malay Tweet normalization. *Information Processing & Management*, 621–633.

Monojit Choudhury, Rahul Saraf, Vijit Jain, Animesh Mukherjee, Sudeshna Sarkar, and Anupam Basu. (2007). Investigation and modeling of the structure of texting language. *International Journal of Document Analysis and Recognition*, 157-174.

Moore, Eric Brill and Robert C. (2000). *An improved error model for noisy channel spelling correction.* Englewood Cliffs, NJ, USA: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics.

Mukhamedova, Raikhangul. (2015). Kazakh: A Comprehensive Grammar. *Routledge*(ISBN 9781317573081).

Nitin Indurkhya, Fred J. Damerau. (2010). *Handbook of Natural Language Processing* (2 ed.). New York, US: Taylor&Francis Group, LLC.

O. De Clercq, B. Desmet, S. Schulz, E. Lefever, V. Hoste. (2013). Normalization of Dutch user-generated content. *Proceedings of the 9th International Conference on Recent Advances in Natural Language Processing* (pp. 179-88). Hissar, Bulgaria: RANLP'13.

Panchapagesan Krishnamurthy, P.P. Talukdar, N Sridhar, A.G. Ramakrishnan. (2004). Hindi Text Normalization. *Conference: Fifth International Conference on Knowledge Based Computer Systems (KBCS)* (p. 10). Hyderabad, India: KBCS.

Paul Cook and Suzanne Stevenson. (2009). *An unsupervised model for text message normalization.* Boulder, USA: CALC 09: Proceedings of the Workshop on Computational Approaches to Linguistic Creativity.

Richard Sproat, Alan W. Black, Stanley F. Chen, Shankar Kumar, Mari Ostendorf, Christopher Richards. (2001). Normalization of non-standard. *Computer Speech & Language, 15*(3), 287-333.

Steffen Eger et al. (2016). A comparison of four character-level string-to-string translation models for (OCR) spelling error correction. *The Prague Bulletin of Mathematical Linguistics*, 77–99.

Yukino Baba, Hisami Suzuki. (2012). How Are Spelling Errors Generated and Corrected? A Study of Corrected and Uncorrected Spelling Errors Using Keystroke Logs. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics* (pp. 373-377). Jeju, Republic of Korea: Association for Computational Linguistics.