Springer Link

Inferring the Complete Set of Kazakh Endings as a Language Resource

International Conference on Computational Collective Intelligence

ICCCI 2020: Advances in Computational Collective Intelligence pp 741-751 | Cite as

- Ualsher Tukeyev (1) Email author (ualsher.tukeyev@gmail.com)View author's OrcID profile (View OrcID profile)
- Aidana Karibayeva (1) View author's OrcID profile (View OrcID profile)

1. Al-Farabi Kazakh National University, , Almaty, Kazakhstan

Conference paper
First Online: 19 November 2020

- 241 Downloads

Part of the Communications in Computer and Information Science book series (CCIS, volume 1287)

Abstract

The Kazakh language belongs to low-resource languages. For application of actual modern branches as artificial intelligence, machine translation, summarization, sentiment analysis, etc. to the Kazakh language needs increasing the number of electronic language resources. Although neural machine translation (NMT) has shown impressive results for many world languages, it does not solve the problem of low-resource languages. Therefore, the development of resources and tools perfecting the use of NMT for low-resource languages is relevant. For perfect use of NMT for the Kazakh language needs bilingual parallel corpora, but also needs a perfect method of the segmentation source text. By the opinion of authors, one of the effective ways for source text segmentation is morphological segmentation. The authors propose to use for morphological segmentation of Kazakh text a table of a complete set of Kazakh words' endings. In this paper is described the inferring of the complete set of Kazakh words' endings. Development of the table of the complete set of word' endings of the Kazakh language will allow in one-step (by reference to the table of endings of the language) to perform the segmentation of the word's ending into suffixes. The complete set of endings of the Kazakh language allows guaranteeing the analysis of any word of the Kazakh language, as this is determined by the inferring of the complete system of words' endings of the language.

Keywords

The Kazakh language Morphological segmentation Words' ending Language resource

This is a preview of subscription content, log in to check access.

Notes

Acknowledgements

This work was carried out under grant No. AP05131415 “Development and research of the neural machine translation system of Kazakh language”, funded by the Ministry of Education and Science of the Republic of Kazakhstan for 2018-2020.

References

1. Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, vol. 1, pp. 1715–1725 (2016)
Google Scholar (https://scholar.google.com/scholar?q=Sennrich%2C%20R.%2C%20Haddow%2C%20B.%2C%20Birch%2C%20A.%3A%20Neural%20machine%20translation%20of%20rare%20words%20with%20subword%20units.%20In%3A%20Proceedings%20of%20the%2054th%20Annual%20Meeting%20of%20the%20Association%20for%20Computational%20Linguistics%2C%20vol.%201%2C%20pp.%201715%E2%80%931725%20%282016%29)

2. Tukeyev, U.: Automaton models of the morphology analysis and the completeness of the endings of the Kazakh language. In: Proceedings of the International Conference “Turkic Languages Processing” TURKLANG 2015, Kazan, Tatarstan, Russia, 17–19 September, pp. 91–100 (2015)
Google Scholar (https://scholar.google.com/scholar?q=Tukeyev%2C%20U.%3A%20Automaton%20models%20of%20the%20morphology%20analysis%20and%20the%20completeness%20of%20the%20endings%20of%20the%20Kazakh%20language.%20In%3A%20Proceedings%20of%20the%20International%20Conference%20%E2%80%9CTurkic%20Languages%20Processing%E2%80%9D%20TURKLANG%202015%2C%20Kazan%2C%20Tatarstan%2C%20Russia%2C%2017%E2%80%9319%20September%2C%20pp.%2091%E2%80%93100%20%282015%29)

3. Tacorda, A.J., Ignacio, M.J., Oco, N., Roxas, R.E.: Controlling byte pair encoding for neural machine translation. In: 2017 International Conference on Asian Language Processing, pp. 168–171 (2017)
Google Scholar (https://scholar.google.com/scholar?q=Tacorda%2C%20A.J.%2C%20Ignacio%2C%20M.J.%2C%20Oco%2C%20N.%2C%20Roxas%2C%20R.E.%3A%20Controlling%20byte%20pair%20encoding%20for%20neural%20machine%20translation.%20In%3A%202017%20International%20Conference%20on%20Asian%20Language%20Processing%2C%20pp.%20168%E2%80%93171%20%282017%29)

4. Wu, Y., Zhao, H.: Finding better subword segmentation for neural machine translation. In: Sun, M., Liu, T., Wang, X., Liu, Z., Liu, Y. (eds.) CCL/NLP-NABD -2018. LNCS (LNAI), vol. 11221, pp. 53–64. Springer, Cham (2018).
https://doi.org/10.1007/978-3-030-01716-3_5 (https://doi.org/10.1007/978-3-030-01716-3_5)
CrossRef (https://doi.org/10.1007/978-3-030-01716-3_5)

Google Scholar (http://scholar.google.com/scholar_lookup?title=Finding%20better%20subword%20segmentation%20for%20neural%20machine%20translation&author=Y.%20Wu&author=H.%20Zhao&pages=53-64&publication_year=2018)

5. Ataman, D., Negri, M., Turchi, M., Federico, M.: Linguistically motivated vocabulary reduction for neural machine translation from Turkish to English. Prague Bull. Math. Linguist. **108**(1), 331–342 (2017)
CrossRef (https://doi.org/10.1515/pralin-2017-0031)
Google Scholar (http://scholar.google.com/scholar_lookup?title=Linguistically%20motivated%20vocabulary%20reduction%20for%20neural%20machine%20translation%20from%20Turkish%20to%20English&author=D.%20Ataman&author=M.%20Negri&author=M.%20Turchi&author=M.%20Federico&journal=Prague%20Bull.%20Math.%20Linguist.&volume=108&issue=1&pages=331-342&publication_year=2017)

6. Creutz, M., Lagus, K.: Unsupervised discovery of morphemes. In: Proceedings of the ACL 2002 Workshop on Morphological and Phonological Learning, vol. 6, pp. 21–30 (2002)
Google Scholar (https://scholar.google.com/scholar?q=Creutz%2C%20M.%2C%20Lagus%2C%20K.%3A%20Unsupervised%20discovery%20of%20morphemes.%20In%3A%20Proceedings%20of%20the%20ACL%202002%20Workshop%20on%20Morphological%20and%20Phonological%20Learning%2C%20vol.%206%2C%20pp.%2021%E2%80%9330%20%282002%29)

7. Koskenniemi, K.: Two-level morphology: a general computational model for word-form recognition and production. Ph.D. thesis, University of Helsinki (1983)
Google Scholar (https://scholar.google.com/scholar?q=Koskenniemi%2C%20K.%3A%20Two-level%20morphology%3A%20a%20general%20computational%20model%20for%20word-form%20recognition%20and%20production.%20Ph.D.%20thesis%2C%20University%20of%20Helsinki%20%281983%29)

8. Oflazer, K.: two-level description of Turkish morphology. Literary Linguist. Comput. **9**(2), 137–148 (1994)
CrossRef (https://doi.org/10.1093/llc/9.2.137)
Google Scholar (http://scholar.google.com/scholar_lookup?title=two-level%20description%20of%20Turkish%20morphology&author=K.%20Oflazer&journal=Literary%20Linguist.%20Comput.&volume=9&issue=2&pages=137-148&publication_year=1994)

9. Beesley, K.R., Karttunen, L.: Finite-State Morphology. CSLI Publications, Stanford University (2003)
Google Scholar (https://scholar.google.com/scholar?q=Beesley%2C%20K.R.%2C%20Karttunen%2C%20L.%3A%20Finite-State%20Morphology.%20CSLI%20Publications%2C%20Stanford%20University%20%282003%29)

10. Kairakbay, B.: A nominal paradigm of the Kazakh language. In: 11th International Conference on Finite State Methods and Natural Language Processing, pp. 108–112 (2013)
Google Scholar (https://scholar.google.com/scholar?q=Kairakbay%2C%20B.%3A%20A%20nominal%20paradigm%20of%20the%20Kazakh%20language.%20In%3A%2011th%20International%20Conference%2

0on%20Finite%20State%20Methods%20and%20Natural%20Language%20Processing%2C%20pp.%20108%E2%80%93112%20%282013%29)

11. Kessikbayeva, G., Cicekli, I.: Rule based morphological analyzer of Kazakh language. In: Proceedings of the 2014 Joint Meeting of SIGMORPHON and SIGFSM, Baltimore, Maryland USA, pp. 46–54 (2014)
Google Scholar (https://scholar.google.com/scholar?q=Kessikbayeva%2C%20G.%2C%20Cicekli%2C%20I.%3A%20Rule%20based%20morphological%20analyzer%20of%20Kazakh%20language.%20In%3A%20Proceedings%20of%20the%202014%20Joint%20Meeting%20of%20SIGMORPHON%20and%20SIGFSM%2C%20Baltimore%2C%20Maryland%20USA%2C%20pp.%2046%E2%80%9354%20%282014%29)

Copyright information

© Springer Nature Switzerland AG 2020

About this paper

Cite this paper as:

Tukeyev U., Karibayeva A. (2020) Inferring the Complete Set of Kazakh Endings as a Language Resource. In: Hernes M., Wojtkiewicz K., Szczerbicki E. (eds) Advances in Computational Collective Intelligence. ICCCI 2020. Communications in Computer and Information Science, vol 1287. Springer, Cham. https://doi.org/10.1007/978-3-030-63119-2_60

- First Online 19 November 2020
- DOI https://doi.org/10.1007/978-3-030-63119-2_60
- Publisher Name Springer, Cham
- Print ISBN 978-3-030-63118-5
- Online ISBN 978-3-030-63119-2
- eBook Packages Computer Science Computer Science (R0)

- Buy this book on publisher's site
- Reprints and Permissions

Personalised recommendations

SPRINGER NATURE

© 2020 Springer Nature Switzerland AG. Part of Springer Nature.

Not logged in Not affiliated 147.30.34.120