

TOWARDS A DATA-DRIVEN MORPHOLOGICAL ANALYSIS OF KAZAKH LANGUAGE

Olzhas Makhambetov, Aibek Makazhanov, Zhandos Yessenbayev,
Islam Sabyrgaliyev, and Anuar Sharafudinov

Nazarbayev University Research and Innovation System,
53 Kabanbay batyr ave., Astana, Kazakhstan

{omakhambetov, aibek.makazhanov, zhyessenbayev,
islam.sabyrgaliyev, anuar.sharaphudinov}@nu.edu.kz

ABSTRACT

We propose a method for complete morphological analysis of Kazakh language that accounts for both inflectional and derivational morphology. Our method is data-driven and does not require manually generated rules, which makes it convenient for analyzing agglutinative languages. The intuition behind our approach is to label morphemes with so called transition labels, i.e. labels that encode grammatical functions of morphemes as transitions between corresponding POS, and use transitivity to ease the analysis. We evaluate our method on a fair-sized sample of real data and report encouraging results.

1 Introduction

Morphological analysis (MA) is one of the crucial steps in automated processing of any language, and in the case of agglutinative languages (ALs) it is hard to overestimate its importance. Agglutination causes words to acquire complex meanings, effectively transforming them into whole phrases. Consider a Kazakh word [*bolmaghandyqtan*] which translates into English phrases [*because something/someone is/was absent*] or [*because something does/did not go certain way*]. MA of the word reveals the underlying phrase: [bol-ma-ghan-dyq-tan] → [exist-NEG-PTCP-NOM-ADV] → [exist - (exist not) - (non

existing) - (nonexistence) - (due to nonexistence)]. Obviously parsing and translating ALs require MA to deal with such cases. Even POS-tagging in ALs benefits from leveraging morphological information [1, 2]. Traditionally the MA problem has been approached by building finite state transducers (FST) [3–5] based on a formal description of the morphology of a language. FST-based approaches require a set of morphological and phonological rules to generate analyses that are both grammatically and orthographically correct. Although there are open source tools that effectively implement transducers [6, 7], certain language-specific morphotactics still need to be implemented. Whilst acknowledging the efficiency and descriptive power of transducers, in the present exploratory study, we focus on a pure data-driven approach, that can be later used as a baseline method or, indeed, as a lightweight morphological analyzer. In this respect, it should be noted that in our approach we do not consider certain language-specific issues. Namely, as we will show later, our method does not account for compound words and certain phonological rules. We plan to address these issues in the future.

We divide the MA problem into the problems of (i) morphological segmentation and (ii) ranking. The respective challenges are: (i)

pruning potentially huge lists of candidate segmentations, while trying to keep the correct ones (precision-recall trade-off) and (ii) employing an effective ranking strategy. To address the first challenge we label every morpheme with its respective POS-to-POS transition label (inflectional morphemes (e.g., plural endings, etc.) are converted to a transition of POS to itself). As we will show later this allowed us to achieve a data coverage of around 97% (i.e. a correct analysis was found for 97% of test words) and maintain a decent precision-recall trade-off. To rank candidate segmentations we use a Hidden Markov Model (HMM) and a Markov chain model, assuming mutual independence of roots and paradigms, and dependence of consequent morphemes within paradigms. Evaluating the models in terms of precision- and recall-at- k , we show that, simple as it is, our approach achieves encouraging performance.

The remainder of the paper is organized as follows. In the next section we review some of the existing work on morphological analysis of morphologically-rich languages in general, and Kazakh language in particular. In Section 3 we thoroughly describe the underlying methodology of our approach. Section 4 presents our experiments and discusses the results. In Section 5 we conclude the study and discuss the future work.

2 Related Work

Statistical approaches to the MA problem have been successfully applied in the past. In a work presented by Hakkani-Tur et al. [8] the distribution of morphological analyses for Turkish is modeled using n -gram models that formulate certain morphosyntactic features, which differ by morphotactical relation of inflectional groups (IGs) within the word and the final IGs of previous words. For Czech language, Hajič et al. [9] combined a rule-based system with a statistical model based on HMM, using these approaches sequentially. Chrupała et al. [10] cast the problem into a classification task, training two maximum entropy classifiers that provide probability

distributions over analyses and word-lemma pairs. The authors use a language independent set of features, and show that their system performs well, achieving respective accuracies of 97%, 94%, and 82% for morphologically-rich languages, such as Romanian, Spanish, and Polish.

Along with supervised methods several unsupervised approaches were proposed [11, 12]. In a work by Creutz and Lagus [11] words are initially segmented using a baseline algorithm, which is based on a recursive minimum description length (MDL) model. Then, initial segmentations are reanalyzed by more advanced models formulated in a maximum a posteriori probability, a maximum likelihood or an MDL framework. The authors refer to this collection of models as the Morfessor. A slightly modified version of the Morfessor was presented by Kohonen et al. [12], who implemented a semi-supervised extension to the baseline algorithm.

Recently there have been attempts to develop formal methods for morphological analysis of Kazakh. While Sharipbayev et al. [13] addressed the problem of Kazakh word forms generation for all inflectional parts of speech, employing semantic neural networks¹ [14], a number of works [15–18] resorted to finite state approaches. Kairakbay et al. [18] present a formal description of the Kazakh nominal paradigm, and Zafer et al. [16] provide a rather vague description of a two-level Kazakh morphology. Both works, however, do not report any significant results. Kessikbayeva et al. [15] also resort to a finite state morphology, and provide a thorough description of the nominal and verb paradigms, and formalize some of the derivational rules. Using the Xerox finite state toolkit [19] the authors conduct experiments on a set of 2000 randomly chosen analyses and report an overall data coverage of 96% (precision was not reported). Finally, Makazhanov et al. [20] address the problem in a context of spelling correction. The authors

¹ Unfortunately the authors do not provide any information on the results of their experiments.

formalize nominal and verb paradigms and develop an error tolerant FSA, reporting 83% general accuracy on a dataset of 1700 word-error pairs.

Our work differs from the aforementioned works on Kazakh morphology in that (i) it considers both inflections and derivatives; (ii) it needs no manual rule generation; (iii) it was evaluated on the largest data set available for Kazakh.

3 Methodology

We divided the task of morphological analysis into two major components: (i) segmenting input words into morpheme sequences; (ii) finding the most probable sequence of morpheme-tag pairs (analysis).

In the absence of a transducer segmenting an input word becomes challenging. A naive approach is to try labeling all possible letter sequences in a given word. This is, however, computationally prohibitive and we want to do better than that. The first thing that comes to mind is to use a morpheme dictionary acquired from a labeled data, and search for matches in a given word. However, simple matching does not account for a natural morpheme order that exists in the language. One could parse all the morpheme sequences and infer this order, eventually ending up building a sort of a state machine. This approach, however, has a potential of missing correct analyses where a certain morpheme sequence occur that had not been seen in a training set.

To account for such omissions, we convert all morpheme labels into POS transitions, i.e. for a given analysis [bol-ma-ghan-dar] \rightarrow [exist-NEG-PTCP-NOM.PL], we construct the following representation: [bol_R_VB-ma_VB_VB-ghan_VB_PTCP-dar_PTCP_PTCP]².

Now, suppose, that in a training set we have seen both morphemes ghan-PTCP and dar-NOM.PL, but we have not observed them in a

sequence, i.e. a pattern [ghandar] never occurred. Suppose, also, that we have seen a sequence of respective allomorphs [gen-NOM-der-NOM.PL], or in a transitive notation: [gen_VB_PTCP-der_PTCP_PTCP]. A method that works with conventional morpheme labels fails to segment this pattern, because a [ghan-PTCP-dar-NOM.PL] sequence had not occurred. However, due to the fact that we have seen transitional labels VB_PTCP and PTCP_PTCP, the transition-based method constructs a link, and successfully segments the pattern.

The segmentation module is developed using recursive function that tries to segment a word, from left to right, into substrings, which are elements of dictionary of morpheme transitions. The process stops when a left substring (prefix) matches a known root or its character length is equal to one. Once we acquire all segmentations we convert transitions back to conventional morpheme tags used in a given language.

To select the most probable segmentation we have conducted ranking experiments using two models. The first approach is based on Markov chains, where the probability of a sequence of morphemes is computed on morpheme bigrams using a chain rule (i.e. the current morpheme depends only on the previous one):

$$P(W) \propto P(r_t) \prod_{i=1}^n P(m_{t_i}|m_{t_{i-1}})$$

where r_t is a POS-tag-labeled root of a given word and m_t is a grammatically-labeled morpheme. We estimate morpheme bigram probabilities using Maximum Likelihood Estimation (MLE). To account for a data sparseness problem we assign a portion of the probability mass to unseen cases by employing the Laplace smoothing:

$$P(m_{t_i}|m_{t_{i-1}}) \approx \frac{N(m_{t_i}, m_{t_{i-1}}) + \alpha}{N(m_{t_{i-1}}) + \alpha|V|}$$

where $N(m_{t_i}, m_{t_{i-1}})$ is the count of a given morpheme bigram, $|V|$ denotes the cardinality of a set of unique morphemes, and smoothing parameter $\alpha = 0.1$ (estimated empirically). We have to note that while computing the

² Notice that a plural ending [dar-NOM.PL] is converted to a [PTCP_PTCP] transition, i.e. inflectional morphemes are replaced by transitions of POS to themselves.

probability of the first morpheme that immediately follows the root, we assume that it depends on the POS of the root. The probability of a root is estimated in the following manner:

$$P(r_t) \approx \frac{N(r_t) + \alpha}{N + \alpha|W|}$$

where $N(r_t)$ is the count of a given POS-tag-labeled root and N is the total number of all words in the training set. In order to prioritize segmentations with vocabulary roots we heavily penalize segmentations containing OOV roots. As in the previous case, parameter α is estimated empirically to be equal to 0.1. The described model will be referred to as a simple Markov chain (SMC).

In the second approach we model a distribution of segmentations using HMM, and try to maximize the posterior probability, $P(T|W)$:

$$P(T|W) \approx \frac{P(T)P(W|T)}{P(W)}$$

where $P(T)$ denotes a probability of a morpho-tag sequence, and $P(W|T)$ denotes a probability of a word given a tag sequence T . The denominator $P(W)$ remains constant for all segmentations, and thus can be dropped.

We compute $P(T)$ using a chain rule:

$$P(T) = \prod_{i=1}^n P(t_i|t_{i-1})$$

the probabilities of morpho-tag bigrams are estimated in the following manner:

$$P(t_i|t_{i-1}) \approx \frac{N(t_i, t_{i-1}) + \beta}{N(t_{i-1}) + \beta|V|}$$

where $N(t_i, t_{i-1})$ denotes the count of a given bigram, $\beta = 0.9$ (estimated empirically) and $|V|$ is the cardinality of a set of all unique morpho-tags. We compute $P(W|T)$ as follows:

$$P(W|T) \approx \frac{N(m_i, t_i) + \beta}{N(t_i) + \beta|W|}$$

where $N(m_i, t_i)$ is the count of a given tagged morpheme (not just a morpho-tag, but also a surface form), and $|W|$ denotes the cardinality of a set of such all unique tagged morphemes.

	train, Δ -per fold	overall
# roots	22 980	24 255
# mrphs, unigr-s	1 623	1 655
# mrphs, innfl.	332	338
# mrphs, deriv.	1 291	1 317

Table 1. Per fold and overall characteristics of the data set

As it can be seen, unlike the previous model, this one is more abstract, and operates mostly with morpheme tags (except for calculation of $P(W|T)$), leaving out the actual surface forms of morphemes. Hereinafter this model will be referred to as HMM.

As we mentioned in the introduction, in the present study certain language-specific aspects of the MA were not addressed. First, we do not perform analysis of compounds, i.e. in multiple-root words we do not locate every single root and analyze them in isolation. Instead we collapse all roots and possible intermediate paradigms into a single root and consider a paradigm attached to the last root only. For instance, for a word [*ulkendi-kishili*] our method provides the following analysis: [*ulkendi-kishi-li*]→[big-small-ADJ], while the correct analysis is [*ulken-di-kishi-li*]→[big-ADJ-small-ADJ]. Second, in our analyses we do not recover roots or morphemes distorted due to the phonetics of the language. For instance, while a correct analysis for a word [*zhughystyghy*] is [*zhuq-ystyq- y*]→[infect-NOM-NOM-POSS.3SG], our method returns [*zhugh-ys-tygh-y*]→[infect-NOM-NOM-POSS.3SG], i.e. the root [*zhuq*] and a morpheme [*tyq*] remain distorted as [*zhugh*] and [*tygh*] respectively.

4 Experiments

We evaluate our models in terms of precision- and recall at- k on an annotated subset of Kazakh Language Corpus [21]. The data set consists of 610 867 word-tokens (78 704 unique). We perform a standard 10-fold cross-validation and report averages and standard deviations per fold.

Table 1 shows the characteristics of the data set as per training fold and overall data.

Morpheme stats counts include allomorphs. As it can be seen, in our data set there are 1 317 derivatives, almost four times as much as inflectional inflections. To the best of our knowledge, for Kazakh language, it is the largest number of derivational morphemes ever considered.

Precision-at- k is calculated as a ratio of correct analyses found at top- k positions to the total number of correctly analyzed tokens in a fold. Recall-at- k is calculated as a ratio of correct analyses found at top- k positions to the total number of all tokens in a fold.

Table 2 contains the results of the performance of the SMC model. As it can be seen 73% of all correct analyses were placed at the first position of the ranking lists, and, in terms of recall, in 71% of the cases correct analyses appeared at the first rank. There is a steady growth with increase in k , and for $k = 5$ the model achieves 90% precision and 87% recall. In general we observe close values for precision and recall for every k . Overall, in 97% of the cases (per fold) a correct analyses was provided.

Table 3 contains the results of the performance of the HMM model. We can see that this model performs slightly lower dragging behind SMC for about 5% (for $k = 1$) in both precision and recall. We believe that this happens because, in contrast to our initial intuition, by ignoring surface forms of morphemes HMM loses some important information rather than resolving ambiguous allomorphic cases. In terms of precision-recall trade-off we observe a trend similar to that of SMC.

When we analyzed the cases where our models failed to put a correct analysis in top-5, we found that a lot of such low ranked cases were due to context related errors. We have performed initial experiments with a context-sensitive model, which utilizes POS information of a preceding root and achieved a top-1 precision of 79% on a 95-to-5% train/test data split. These initial results suggest that incorporating context information may help to boost the accuracy of the method.

k	precision at- k	recall at- k
1	73.2±0.37	70.9±1.06
2	85.3±0.46	83.2±1.05
3	88.8±0.48	86.3±1.20
4	90.0±0.44	87.8±1.21
5	90.6±0.45	87.7±1.13

Table 2. Precision- and recall- k for SMC average± standard deviation per fold

k	precision at- k	recall at- k
1	68.3±0.46	66.2±0.73
2	81.6±0.50	79.0±1.07
3	86.0±0.48	83.7±1.22
4	88.5±0.49	85.7±0.91
5	89.7±0.46	86.8±1.15

Table 3. Precision- and recall- k for HMM average± standard deviation per fold

5 Conclusion and Future Work

We have developed a data-driven method for morphological analysis of Kazakh language that accounts for both inflectional and derivational morphology. The method does not require formalization, in that all the rules are induced directly from labeled data in the form POS-to-POS morpheme transitions. Our experiments suggest that these transition-based morphotactics help in pruning many false patterns while keeping correct analyses as candidate segmentations. We believe that the same technique could be used in the analysis of other agglutinative languages, as all that it requires is labeled data in a given language. We evaluated our method in terms of top- k precision using Kazakh as a reference language. The best of our models achieved 90% performance in terms of precision-at- k . The analysis of generated segmentations revealed that most of errors occurred due to context insensitivity of our method. We have already started experiments on incorporation of context information, and achieved encouraging initial results. Our future work will be directed at development of a context-sensitive extension of the method. In addition, we will make necessary adjustments to the method to

facilitate compound-sensitive and phonetically-correct analyses.

6 References

- [1] D. Elworthy, "Tagset design and inflected languages," in In EACL SIGDAT workshop iFrom Texts to Tags: Issues in Multilingual Language Analysis, 1995, pp. 1–10.
- [2] J. Hana and A. Feldman, "A positional tag set for Russian," Proceedings of LREC-10. Malta, 2010.
- [3] K. Koskenniemi, "A general computational model for word-form recognition and production," in Proceedings of the 10th international conference on Computational linguistics. ACL, 1984, pp. 178–181.
- [4] K. Oflazer and C. Güzey, "Spelling correction in agglutinative languages." in ANLP, 1994, pp. 194–195.
- [5] H. Sak, T. Güngör, and M. Saraçlar, "A stochastic finite-state morphological parser for Turkish," in Proceedings of the ACL-IJCNLP 2009 Conference. Stroudsburg, PA, USA: ACL, 2009, pp. 273–276.
- [6] M. Hulden, "Foma: a finite-state compiler and library." in EACL (Demos), A. Lascarides, C. Gardent, and J. Nivre, Eds. ACL, 2009, pp. 29–32.
- [7] K. Linden, M. Silfverberg, E. Axelson, S. Hardwick, and T. Pirinen, HFST-Framework for Compiling and Applying Morphologies, ser. Communications in Computer and Information Science, 2011, vol. Vol. 100, pp. 67–85.
- [8] D. Z. Hakkani-Tur, K. Oflazer, and G. Tur, "Statistical morphological disambiguation for agglutinative languages." Computers and the Humanities, vol. 36, no. 4, pp. 381–410, 2002.
- [9] J. Hajič, P. Krbeč, P. Pavel Květoň, K. Oliva, and V. Petkevič, "Serial combination of rules and statistics: A case study in czech tagging," in Proceedings of the 39th Annual Meeting on ACL. Stroudsburg, PA, USA: ACL, 2001, pp. 268–275.
- [10] G. D. Grzegorz Chrupała and J. van Genabith, "Learning morphology with morfette," in Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco: ELRA, may 2008.
- [11] M. Creutz and K. Lagus, "Unsupervised models for morpheme segmentation and morphology learning," ACM Transactions on Speech and Language Processing (TSLP), vol. 4, no. 1, p. 3, 2007.
- [12] O. Kohonen, S. Virpioja, L. Leppänen, and K. Lagus, "Semi-supervised extensions to morfessor baseline," in Proceedings of the Morpho Challenge 2010 Workshop. Espoo, Finland: Aalto University, September 2010.
- [13] A. Sharipbayev, G. Bekmanova, B. Ergesh, A. Buribayeva, and M. K. Karabalayeva, "Intellectual morphological analyzer based on semantic networks," in Proceedings of the OSTIS-2012, 2012, pp. 397–400.
- [14] D. E. Shuklin, "The structure of a semantic neural network extracting the meaning from a text," Cybernetics and Sys. Anal., vol. 37, no. 2, pp. 182–186, Mar. 2001.
- [15] G. Kessikbayeva and I. Cicekli, "Rule based morphological analyzer of Kazakh language," in Proceedings of the 2014 Joint Meeting of SIGMORPHON and SIGFSM. Baltimore, Maryland: ACL, June 2014, pp. 46–54.
- [16] H. R. Zafer, B. Tilki, A. Kurt, and M. Kara, "Two-level description of Kazakh morphology," in Proceedings of the 1st International Conference on Foreign Language Teaching and Applied Linguistics (FLTAL11), Sarajevo, May 2011.
- [17] G. Altenbek and W. Xiao-long, "Kazakh segmentation system of inflectional affixes," in CIPS-SIGHAN, 2010, pp. 183–190.
- [18] B. M. Kairakbay and D. L. Zaurbekov, "Finite state approach to the Kazakh nominal paradigm," in Proceedings of the 11th International Conference on Finite State Methods and Natural Language Processing. St Andrews, Scotland: ACL, July 2013, pp. 108–112.
- [19] A. Ranta, "A multilingual natural-language interface to regular expressions," in Proceedings of the International Workshop on Finite State Methods in Natural Language Processing, ser. FSMNLP '09. Stroudsburg, PA, USA: ACL, 1998, pp. 79–90.
- [20] A. Makazhanov, O. Makhambetov, I. Sabyrgaliyev, and Z. Yessenbayev, "Spelling correction for kazakh," in Proceedings of the 2014 CICLing. Kathmandu, Nepal: Springer Berlin Heidelberg, 2014, pp. 533–541.
- [21] O. Makhambetov, A. Makazhanov, Z. Yessenbayev, B. Matkarimov, I. Sabyrgaliyev, and A. Sharafudinov, "Assembling the kazakh language corpus," in EMNLP. Seattle, Washington, USA: ACL, October 2013, pp. 1022–1031.