SÜNİ İNTELLEKT ALƏTLƏRİNİN PRAKTİKADA TƏTBİQİ NÜMUNƏLƏRİ

UOT 004.8
UOT 811

DOI: 10.25045/ASUCAAI.2025.01

ВЫЧИСЛИТЕЛЬНЫЕ МОДЕЛИ МОРФОЛОГИИ И СИНГАРМОНИЗМА КАЗАХСКОГО ЯЗЫКА И ИХ ИСПОЛЬЗОВАНИЕ ДЛЯ НЕЙРОННЫХ МОДЕЛЕЙ

Уалшер Тукеев Ануарбекулы о,
Д.т.н., профессор
Казахский Национальный Университет им. Аль-Фараби
Алматы, Казахстан
ualsher.tukeyev@gmail.com

Ualşer Tukeyev

Qazax dilinin morfologiyasi və sinharmonizminin hesablama modelləri və neyron modelleri üçün istifadəsi

Xülasə

Məqalədə tam sonluqlar dəstinə (CSE – Complete Set of Endings) əsaslanan morfologiya hesablama modeli, qazaq dilinin sinxarmonizm hesablama modeli və onların qazaq dilində söz formalarının generasiyasında istifadəsi təqdim olunur. Qazaq dilinin söz formaları siyahıları dataset şəklində təbii dilin emalı ilə bağlı müxtəlif tapşırıqların neyron modellərinin öyrədilməsi üçün istifadə oluna bilər. Qazax dilində sinxarmonizm hesablama modeli və qazax dilində söz formalarının generasiyası üçün proqram nümunələri təqdim olunmuşdur.

Ualsher Tukeyev

Computer Models of Morphology and Synharmonism of the Kazakh Language and Their Use for Neuron Models

Abstract

The study presents a computational model of morphology based on the Complete Set of Endings (CSE), a computational model of synharmonism for the Kazakh language, and their use in generating word forms of the Kazakh language. The lists of word forms in Kazakh, presented as datasets, can be used to train neural models for various natural language processing tasks. Samples of programs for the computational model of synharmonism in the Kazakh language and the generation of word forms in the Kazakh language are provided.

Açar sözlər: Qazax dili, morfologiya, sinharmonizm, neyron modelləri, CSE-model.
Keywords: Kazakh Language, Morphology, Synharmonism, Neuron Models, CSE Model.
Ключевые слова: Казахский язык, морфология, сингармонизм, нейронные модели, CSE-модель.

1. Введение

Тюркские языки составляют семью, включающую более 35 языков [Дыбо, А., 2013], на которых говорят более 180 миллионов человек [Gutman, A. and Avanzati, B., 2013] в нескольких странах. В тюркскую группу языков входят такие государственные языки, как азербайджанский, казахский, киргизский, узбекский, турецкий, туркменский. Языками субъектов государств являются алтайский, балкарский, башкирский, каракалпакский, крымскотатарский, кумыкский, ногайский, татарский, тувинский, уйгурский, хакасский, шорский, якутский.

В области обработки естественных языков (ОЕЯ) существуют две группы вычислительных моделей и методов для обработки языков: преобразователи с конечным числом состояний (FST – finite state transducers) и методы машинного обучения. Группа методов FST требует использования ориентированного на пользователя языка программирования для описания исходных данных для новых языков. Вторая группа методов, группа машинного обучения, требует большого объема электронных исходных данных для машинного обучения, чего нет для многих языков с ограниченными ресурсами.

В данной работе описывается вычислительная модель морфологии, основанная на полных наборах окончаний (CSE – Complete Set of Endings) для тюркских языков на примере казахского языка. Предлагаемый подход позволяет пользователю (лингвисту) использовать универсальные (управляемые данными) алгоритмы и программы для ряда задач ОЕЯ, таких как определение основ слов (стемминг), морфологический анализ текста и сегментация текста. Этот подход особенно важен для большого количества языков с ограниченными ресурсами, для которых исходных данных (корпусов) для методов машинного обучения все еще недостаточно.

Вместе с тем в данной работе рассматривается вычислительная модель сингармонизма казахского языка и использование его совместно с CSE-моделью для генерации датасета словоформ, которое может быть использовано для обучения нейронных моделей. По мнению автора, это является актуальным для малоресурсных языков.

2. Модели морфологии естественных языков

Существуют три общепринятые модели морфологии естественных языков [Spencer, A., 1991; Плунгян, В., 2003], а именно: «Item and Arrangement – Элемент и расположение» (IA-модель); «Item and Process – Элемент и процесс» (IP-модель); «Word and Paradigm – Слово и парадигма» (WP-модель).

IA-модель фокусируется на агглютинативном характере словоформ. Его основной инструмент моделирования выполняет линейную сегментацию словоформ на морфемы. Морфема – это минимальная значимая неделимая часть слова. Рассматривая морфемы как минимальные единицы грамматического описания, IA-модель хорошо подходит для описания морфологии агглютинативных языков.

IP-модель фокусируется на концепции динамической природы алломорфов, вводя один или несколько уровней представления словоформ. Каждая морфема словоформы обязательно имеет единственное глубокое представление, а также правила перехода к более

поверхностным уровням представления с учетом контекста, при котором возможны алломорфные вариации представления морфемы морфологии флективных языков.

WP-модель фокусируется на концепции флекции по парадигме. В этой морфологической модели слово рассматривается как единое целое, а не как комбинация основы и окончания. Флексия в WP-модели рассматривается по сходству, а минимальной единицей грамматического описания является словоформа.

На практике при реализации задач обработки естественного языка активно используются модели и методы конечных преобразователей (FST) и методы машинного обучения.

В методах FST современные методы представляют собой двухуровневую (TWOL) морфологическую вычислительную модель [Koskenniemi, K., 1983], которая в основном основана на IP-модели для морфологии. Для реализации этой технологии были разработаны программные средства, которые используются для многих языков. Для использования этих инструментов были разработаны специальные языки пользовательского интерфейса для исходных данных (правила технологии двухуровневой морфологии). Однако освоение и использование пользовательского языка для задания исходных данных для основанных на правилах методов, основанных на двухуровневой морфологии, - довольно трудоемкий процесс. Это является серьезным препятствием для широкого использования лингвистами TWOL технологий, основанных на правилах IP-модели морфологии для стемминга, сегментации и морфологического анализа, особенно для языков с ограниченными ресурсами.

В отличие от предыдущих парадигм, в вычислительной CSE-модели морфологии, используются реляционные модели, минимальными единицами грамматического описания морфологии являются окончания и основы слов. Окончания в CSE-модели рассматриваются как единое целое.

3. Метод

3.1 Вычислительная модель морфологии казахского языка

3.1.1 Вывод допустимых типов окончаний казахского языка

Рассмотрим систему окончаний слов казахского языка двух классов: окончания к именным основам (существительные, прилагательные, числительные) и окончания к глагольным основам (глаголы, причастия, деепричастия, наклонения, залоги) [Tukeyev, U., 2023; Tukeyev, U., Karibayeva, A., Zhumanov, Zh., 2020].

Схема вывода окончаний для каждого класса аффиксов рассматривается отдельно. Однако следующая четырехэтапная процедура одинакова для всех случаев:

- определение комбинации возможных размещений основных типов аффиксов;
- выбор размещений основных типов аффиксов (осуществляется путем проверки их семантической приемлемости в языке);
- перечисление возможных вариантов окончаний для каждого варианта семантически приемлемого размещения основных типов аффиксов;
- объединение окончаний в полный набор окончаний для данного языка.

3.1.2 Система окончаний казахского языка к именным основам

Система окончаний к именным основам слов казахского языка имеет четыре типа аффиксов:

- аффиксы множественного числа (обозначим через K),
- притяжательные аффиксы (обозначим через T),
- падежные аффиксы (обозначим через C),
- личные аффиксы (обозначим через J),
- основу(stem) обозначим через S.

Рассмотрим всевозможные варианты размещений типов аффиксов: из одного типа, из двух типов, из трех типов и из четырех типов. Число размещений определяется формулой:

$A_{nk} = n!/(n-k)!$,

где n – общее количество типов аффиксов, k – количество типов для размещения.

Тогда, количество размещений будет определяться следующим образом:

$A_{41} = 4!/(4-1)!=4$,

$A_{42} = 4!/(4-2)!=12$,

$A_{43} = 4!/(4-3)!=24$,

$A_{44} = 4!/(4-4)!=24$.

Всего возможных размещений 64.

Рассмотрим какие из них семантически допустимы.

Размещения по одному типу аффиксов (**K, T, C, J**) являются все семантически допустимыми по определению.

Размещения по два типа аффиксов могут быть следующие:

KT, TC, CJ, JK

KC, TJ, CT, JT

KJ, TK, CK, JC.

Анализ семантики размещений двух типов аффиксов показывает, что выделенные жирным шрифтом размещения являются допустимыми (**KT, TC, CJ, KC, TJ, KJ**), а остальные размещения относим к недопустимым.

Итак, количество допустимых (правильных) размещений из двух типов аффиксов будет равно 6.

Размещения аффиксов из трех типов будут следующие:

KTC, **KTJ**, **TCJ**, TCK, CJK, CJT, JKT, JKC

KCJ, KCT, TJK, TJC, CTK, CTJ, JTK, JTC

KJT, KJC, TKC, TKJ, CKT, CKJ, JCK, JCT.

Определение допустимых размещений аффиксов из трех типов сделаем по правилу:

если в размещении из трех типов есть недопустимые размещения из двух типов, то это размещение – недопустимо.

Тогда, допустимых размещений аффиксов из трех типов будет 4 (**KTC**, **KTJ**, **TCJ, KCJ** выделено жирным).

Размещения аффиксов из четырех типов будут следующие:

KTJC, TKJC, CKTJ, JKTC

KTCJ, TKCJ, CKJT, JKCT

KJTC, TJKC, CTKJ, JTKC
KJCT, TJCK, CTJK, JTCK
KCTJ, TCJK, CJKT, JCKT
KCJT, TCKJ, CJTK, JCTK

Определение допустимых размещений аффиксов из четырех типов сделаем по правилу:

если в размещении из четырех типов есть недопустимые размещения из двух типов, то это размещение – недопустимо.

Тогда, допустимых размещений аффиксов из четырех типов будет 1 (**KTCJ** выделено жирным).

Итого, допустимых размещений из одного типа – 4, из двух типов - 6, из трех типов – 4, из четырех типов – 1.

Итак, суммарное число типов допустимых размещений типов аффиксов в словах с именными основами – 15.

Перечисляя возможные варианты окончаний для каждого семантически приемлемого размещения основных типов аффиксов получается общее количество возможных окончаний слов данной группы.

Итого количество окончаний для слов с именными основами казахского языка составляет – 2004.

Система окончаний казахского языка к глагольным основам включает следующие виды:

- система окончаний глаголов;
- система окончаний причастий;
- система окончаний деепричастий;
- система окончаний наклонений;
- система окончаний залогов.

Аналогично методом комбинирования получаются множества окончаний для вышеуказанных видов частей речи. В целом, для казахского языка выведено 4890 окончаний.

На этом этапе существенную роль для формирования окончаний играет правила сингармонизма языка. Ниже представлена вычислительная модель сингармонизма казахского языка.

3.2 Вычислительная модель сингармонизма казахского языка

Правила сингармонизма казахского языка для каждой части речи языка имеет свои особенности (правила). Рассмотрим вычислительные модели сингармонизма для каждой части речи в отдельности. Важной особенностью предлагаемой вычислительной модели сингармонизма является использование вычислительной CSE-модели морфологии, что позволяет рассматривать в правилах всевозможные окончания для частей речи.

3.2.1 Вычислительная модель сингармонизма слов с именными основами

Ниже представлены подмножества букв, играющие важные роли в правилах сингармонизма слов с именными основами.

vsolid = ['а', 'о', 'ұ', 'ы', 'у']; vsoft = ['ә', 'ө', 'ү', 'і', 'е','у','и']; cdeaf = ['п', 'к', 'қ', 'т', 'с', 'ш', 'х', 'h']; cvoiced = ['б', 'в', 'г', 'ғ', 'д', 'ж', 'з']; cv1 = ['б', 'в', 'г', 'ғ', 'д']; cv2 = ['ж', 'з']; cvoiced1 = ['б', 'в', 'г', 'д']; csonor = ['л', 'м', 'н', 'ң','р', 'й', 'у']; cs1 = ['м', 'н', 'ң']; cs2 = ['л', 'р', 'й', 'у']; cs3 = ['л', 'м', 'н', 'ң']; cs4 = ['р', 'й', 'у'].

Правила для формирования слов с множественными окончаниями:

Правила для ДАР/ДЕР	Правила для ТАР/ТЕР

if stem[-1] in (cs3 or cv2):
if stem[-2] in vsolid and ending[1] in vsolid:
if ending[:3] == 'дар':
endingtl = ending
if stem[-2] in vsoft and ending[1] in vsoft:
if ending[:3] == 'дер':
endingtl = ending
	if stem[-1] in (cdeaf or cv1):
	if stem[-2] in vsolid and ending[1] in vsolid:
	if ending[:3] == 'тар':
	endingtl = ending
	if stem[-2] in vsoft and ending[1] in vsoft:
	if ending[:3] == 'тер':
	endingtl = ending

Правила формирования слов с окончаниями принадлежности:

Правила для окончаний принадлежности (стем последний символ - гласный)

if stem[-1] in vsolid and ending[1] in vsolid and stem[-2] in vsolid and stem[-1] == 'у':
if ending == 'м' or ending == 'ң' or ending[:3] in ('ңыз', 'мыз') or (ending[:2] == 'сы'):
endingtl = ending
if stem[-1] in vsolid and ending[1] in vsolid:
if ending == 'м' or ending == 'ң' or ending[:3] in ('ңыз', 'мыз') or (ending[:2] == 'сы'):
endingtl = ending
if (stem[-1] in vsoft and ending[1] in vsoft and stem[-2] in vsoft and stem[-1] == 'у'):
if ending == 'м' or ending == 'ң' or ending[:3] in ('ңіз', 'міз') or (ending[:2] == 'сі'):
endingtl = ending
if (stem[-1] in vsoft and ending[1] in vsoft and stem[-2] in vsoft and stem[-1] == 'у'):
if ending == 'м' or ending == 'ң' or ending[:3] in ('ңіз', 'міз') or (ending[:2] == 'сі'):
endingtl = ending

Аналогичные вычислительные правила сингармонизма составлены для других частей речи казахского языка.

4. Использование вычислительной модели морфологии для формирования датасета словоформ казахского языка

Важной особенностью нейронных моделей является необходимость большого размера исходных данных для обучения нейронных моделей, что не всегда является доступных для многих тюркских языков. Поэтому, с точки зрения автора, вычислительная CSE-модель

морфологии является одной из возможностей автоматического формирования датасета словоформ языка для обучения нейронных моделей. Ниже представлена программа формирования словоформ казахского языка для существительных с окончаниями множественного числа. В основе данной программы лежит использование списка всевозможных окончаний казахского языка, составленного по CSE-модели, и вычислительная модель сингармонизма казахского языка, определяющая правильные сочетания стемов и окончаний при генерации словоформ казахского языка.

Программа формирования словоформ для окончаний множественного числа:

```
def DTL_wordforms(stems_list_noun, ending_list):
    combinations = []
    for stem in stems_list_noun:
        print(stem)
        for ending in endings_list:
            if len(ending) > 1:
                ending_D = D_ending(stem, ending)        # check D T L
                if ending_D != '':
                    combinations.append(stem + ending_D)
                ending_T = T_ending(stem, ending)
                if ending_T != '':
                    combinations.append(stem + ending_T)
                ending_L = L_ending(stem, ending)
                if ending_L != '':
                    combinations.append(stem + ending_L)
    return combinations
```

Аналогично составляются вычислительные модели для формирования словоформ других частей речи. Ниже на Рис. 1 представлен скрин генерации словоформ казахского языка.

5. Заключение

В статье представлена модель вычислительной морфологии на основе полного набора окончаний (CSE), модель вычислительного сингармонизма казахского языка и их использование при генерации словоформ казахского языка. Представлена вычислительная модель синхронизма в казахском языке и примеры программ для генерации словоформ в казахском языке. В будущем планируется: обучение нейронных моделей по датасетам словоформ, по датасетам стемов и окончаний, формирование датасетов параллельных словоформ для различных пар тюркских языков и их использование для различных задач обработки естественных языков.

Результаты работы могут быть полезны для дальнейших исследований в области вычислительной лингвистики и разработки новых систем обработки казахского языка.

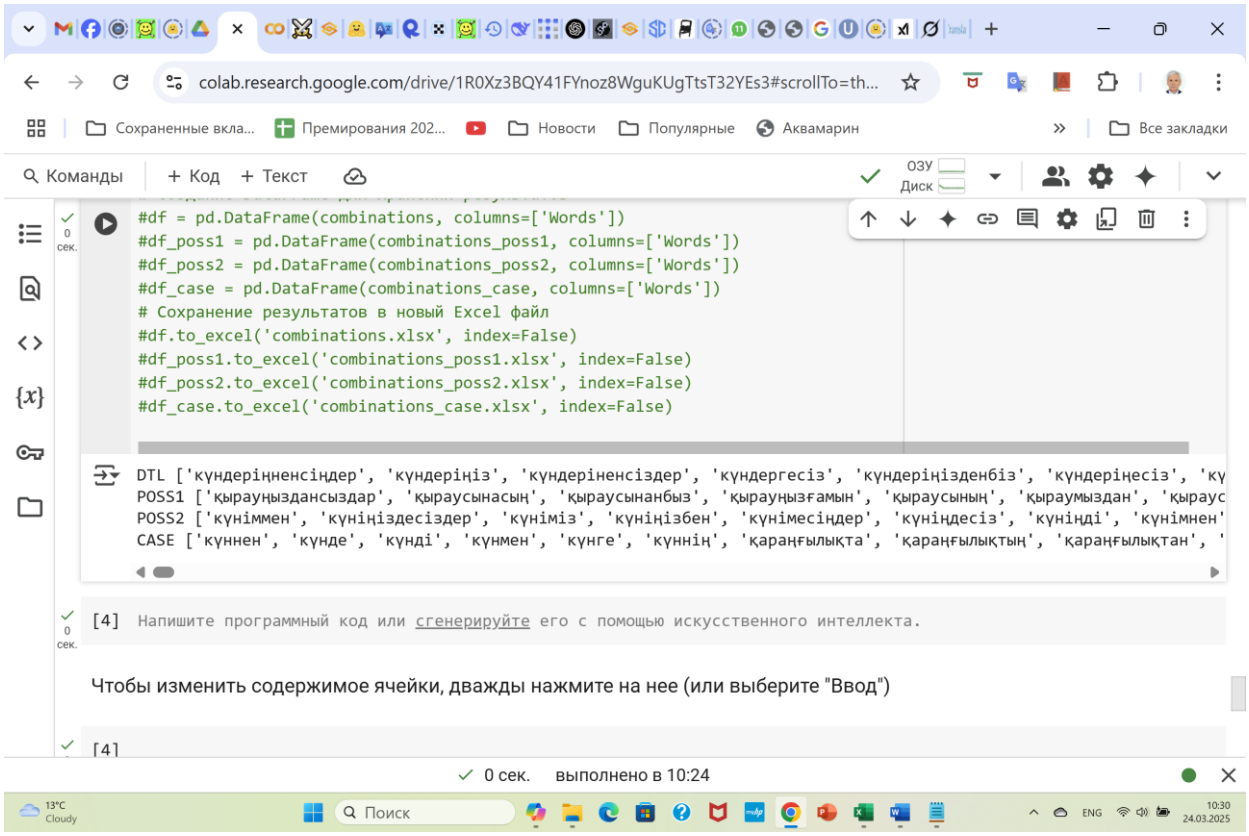

Рис. 1. Скрин генерации словоформ казахского языка (сегмент).

Список литературы:

1. Gutman, A. and Avanzati, B. (2013). The Languages Gulper. Turkic languages. http://www.languagesgulper.com/eng/Turkic.html.

2. Koskenniemi, K. (1983). Two-Level Morphology: A General Computational Model of Word-Form Recognition and Production. Tech. rep. Publication No. 11. Department of General Linguistics. University of Helsinki.

3. Spencer, A. (1991). Morphological theory. An Introduction to Word Structure in Generative Grammar. Blackwell Publishers. pp.512

4. Дыбо, А.В. (2007). Хронология тюркских языков и лингвистические контакты ранних тюрков. http://s155239215.onlinehome.us/turkic/40_Language/Dybo_2007LingivistContactsOfEarlyTurksRu.htm

5. Плунгян, В.А. (2003). Общая морфология: Введение в проблематику: Учебное пособие. Изд. 2-е, исправленное. – М.: Едиториал УРСС, 2003. - 384 с.

6. Tukeyev, U. (2023). A NEW COMPUTATIONAL MODEL FOR TURKIC LANGUAGES MORPHOLOGY AND PROCESSING. Journal of problem in computer science and information technologies, v. 1, n. 1, apr. 2023. doi: https://doi.org/10.26577/JPCSIT.2023.v1.i1.07

7. Tukeyev U., Karibayeva A., Zhumanov Zh. (2020). Morphological Segmentation Method for Turkic Language Neural Machine Translation. Cogent Engineering, Volume 7, 2020 – Issue 1 https://doi.org/10.1080/23311916.2020.1856500