COMPUTATIONAL MODELS OF TURKIC LANGUAGES MORPHOLOGY ON COMPLETE SETS OF ENDINGS

Ualsher Tukeyev,

Professor, doctor of technical science,

Honorary member of National Academy of Science of the Republic of Kazakhstan

Al-Farabi Kazakh National University, Almaty, Kazakhstan

ualsher.tukeyev@gmail.com

Introduction

- In the modern globalization world languages and communication are posing crucial challenges to the society at large. One of the actual problem is low resource languages.
- Artificial intelligence is now a key enabling factor for circulating, sharing and accessing knowledge across languages and cultures.
- One of the cutting edge areas of artificial intelligence is natural language processing (NLP).
- NLP includes: word stemming, morphological analysis, text segmentation, syntactic tagging (POS-tagging), machine translation, language understanding, summarization, information extraction.
- Morphology - main part of linguistics of agglutinative languages.

Introduction

- Turkic languages make up a family including more than 35 languages, which are spoken by more than 160 millions of people across several countries. The Turkic group of languages includes state languages like Azerbaijan, Kazakh, Kyrgyz, Uzbek, Turkish, Turkmen. The languages of the subjects of the states are Altai, Balkar, Bashkir, Karakalpak, Crimean Tatar, Kumyk, Nogai, Tatar, Tuvan, Uyghur, Khakass, Shor, and Yakut.
- This work focuses on the construction of computational models of morphology based on the complete sets of endings (CSE-model) of the Turkic languages.
- The proposed approach allows the user to use universal (data-driven) programs for a number of NLP tasks, such as stemming, morphological analysis and word segmentation.
- One of key features of this approach is that for a new language, only the linguistic resource of that language must be prepared in the form of a relational data model. Then a universal program is used for the corresponding task, driven by the developed data.

Related works

❑ Morphology models:

➢ ‘Item and Arrangement’ (IA-model);

The IA-model focuses on the agglutinative character of word forms. Its main modeling tool performs a linear segmentation of word forms into morphemes. Considering morphemes as its minimal units of grammatical description, the IA-model is well suited for describing the morphology of agglutinative languages.

➢ ‘Item and Process’ (IP-model);

The IP-model focuses on the concept of the dynamic nature of allomorphs, introducing one or more levels of word forms representation. Each morpheme of a word form necessarily has a single deep representation, as well as rules for transition to more superficial levels of representation, taking into account the context, at which allomorphic variation of the morpheme’s representation is possible.

➢ ‘Word and Paradigm’ (WP-model).

The WP-model focuses on the concept of inflection by paradigm. In this morphology model, the word is considered as a whole, rather than a combination of a stem and an ending. Inflection in the WP-model is considered by the similarity, and the minimal unit of grammatical description is the word form.

Related works

❑Computational models of morphology:

- Well-known Computational models of morphology is Two-level morphology (TWOL) of Kimmo Koskaniemi. TWOL represent a word in two levels:
  - Lexical (deep) representation;
  - Surface representation.

- This model based on TWOL rules: transformation from word's Lexical (deep) representation to Surface representation depending of context..
- The existing rule-based methods are mainly based on the technology of two-level morphology, which is mainly based on the IP-model for morphology.
- Software tools have been developed for the implementation of this technology, which are used for many languages.
- To use these tools, special user interface languages have been developed for the initial data (the rules of the two-level morphology technology).
- However, mastering and using a custom language for specifying the initial data for rule-based methods based on two-level morphology is a rather laborious process.
- From the point of view of the author this is a major obstacle for the widespread use of rule-based technologies by linguists for stemming, segmentation and morphological analysis, especially for low-resource languages.

Method

- Computational model of morphology on Complete Set of Endings (CSE)
- Common formal approaches :

Two views of the function.

1) Analytical (algorithmically or rule-production):

Y= F(X), F=2.

2) Tabular view: F=2

Y=F(X)	
X	Y
2	4
3	6

- Our approach for basic task of NLP based on tabular representation of functions.
- Tabular approach lays in the base of relational data model, which is universal approach for modeling of real world, as relational databases.

Method

❑Computational model of morphology on Complete Set of Endings (CSE)

- ❖ Possible approach for description of morphology agglutinative languages:

1) Morphology can presented by description of grammar on Finite-State Automation, Finite-State Transducers.
2) Morphology can presented by list of all word forms. (Bulygina, 1977), (Zaliznyak)
3) Morphology can presented by stem + affixes. (Bulygina T., 1977, Prieto L., 1975), Bektayev K, 1991.

- ❖ Our approach based on enumerate all endings of language. Improoving of the approach of Bektayev K.

Method

- **Computational model of morphology on Complete Set of Endings (CSE)**

  - **Combinatorial approach:**
    - Kazakh affixes to nominal stems:
      Plural (K), possessive (T), case (C), personal (J).
    - Placements: $A_{nk} = n!/(n-k)!$.

    - One type: 4 (4) – K, T, C, J. Two types: 6 (12) – KT, TC, CJ, KC, TJ, KJ. Three types: 4 (24) – KTC, KTJ, TCJ, KCJ. Four types: 1 (24) – KTCJ.
  - **Enumeration of endings.**
    - (Example on KT placement)
      KT: (6 affixes K) * (5 affixes T) = 30 endings.
  - **Complete set of endings of**: Kazakh – 4679, Kyrgyz – 4768, Uzbek-747.

Method

- **Computational data models of morphology for segmentation and morphological analysis**

Relational (table) data model:

- Segmentation

Word endings	Word endings as sequences of affixes
gendermenmin …	gen der men min …

- Morphological analysis

Word endings	Morphological analysis
gendermenmin gensyzdar	<VB>*gen<pp>*der<pl>*men<inst>*min<pos><p1> <VB>*gen<pp>*syz<p2>,frm>dar<pl>

- **The universal (data-driven) algorithms and programs for stemming, segmentation, morph analysis are developed as open source (https://github.com/NLP-KazNU)**

Results

- CSE-model developed for Kazakh, Kyrgyz, Uzbek, Old Turk, Turkish.
- Experiments in Kazakh, Kyrgyz and Old Turk were carried out for stemming and segmentation, in Uzbek and Turkish – for stemming.
- Overall, the accuracy measure for stemming and segmentation achieved 80-90%.
- Now our masters of Computational linguistics educational program are carrying the researches on the base of CSE-model for Kypchak, Karakalpak, Tatar for stemming, segmentation, morphological analysis and machine translation tasks.

Conclusion and future works

The advantage of proposed methodology is that it is oriented towards linguists:

- ❑ in order to solve the problems of stemming, segmentation, morph analysis, it only requires:
  - Building a complete set of language endings for stemming;
  - Building an endings segmentation table for segmentation task;
  - Constructing a table of morph analysis of endings for the tasks of morph analysis;
  - Use the appropriate universal program.

- ❑ Future works is planned both in the direction of:

- increasing of effectiveness of the developed algorithms and programs;
- using proposed methodology for of other languages of Turkic group for NLP tasks.

Thanks for attention!

ualsher.tukeyev@gmail.com