3-бөлім Раздел 3 Section 3

Информатика Информатика Computer science

UDC 004.8:004.9

Tukeyev U.A.*, Rakhimova D.R., Zhumanov Zh.M., Kartbayev A.Zh.

Al-Farabi Kazakh National University, Republic of Kazakhstan, Almaty
*E-mail: ualsher.tukeyev@gmail.com

Single state transducer model for Kazakh and Russian morphology

This paper provides a broad overview of issues related to the construction of finite state transducers with one state for the two-level morphology of inflectional languages, particularly, the direct transformation of word endings to the grammatical characteristics. This problem has been studied on the base of the Kazakh and Russian languages, which are usually named the inflectional languages. The solution of this problem is the trivial Mealy automaton with one state, i.e. a single state transducer, and a multi-valued mapping method is used as well. We study the problem of completeness of the finite state transducers input for the analyzed languages. The determination of transducer input completeness for morphological analysis gives a guarantee that all the words of the analyzed language will be accepted. The problem of determining the completeness of the set of possible endings for agglutinative languages is a complex issue. In this article, we define the completeness of a set of endings in Kazakh language. The proposed technology is implemented for the Russian-Kazakh machine translation, a translation quality assessment performed by the method of BLEU.
Key words: machine translation, finite transducer, two-level morphology, inflectional languages, multi-valued mapping.

Тукеев У.А., Рахимова Д.Р., Жуманов Ж.М., Картбаев А.Ж.

Орыс және қазақ тілдері морфологиясының бір күйлі түрлендіргіш моделі

Бұл жұмыста сөз құрылымы күрделі тілдердің екі деңгейлі морфологиясын зерттеу мақсатында бір санаттағы ақырлы автоматты құруды зерттеу, яғни сөздердің жалғауларын грамматикалық характеристикаларға тікелей өзгерту қаралған. Бұл мәселе агглютинативті және флективті жолмен өзгеретін қазақ және орыс тілдерінің негізінде зерттелген. Мели тривиалды автоматы мен көпмағыналы байланыстыруды қолдану осы мәселенің шешілуіне әкелді. Біз аталған тілдердегі мәтіндер үшін ақырлы автоматтың кірісінің толықтығын анықтадық. Осы нәтиже морфологиялық анализ жасалған барлық сөздердің қабылдануына кепілдеме береді. Агглютинативті тілдер үшін қиын мәселе ауқымында қазақ тілінің сөздері үшін толықтық ерекшеліктері зерттелген. Аталған технология орысша-қазақша машинамен аударуы үшін жасалған, аударманың дұрыстығы BLEU үлгісімен тексерілген.
Түйінді сөздер: екі деңгейлі морфология, машинамен аудару, ақырлы автоматтар, флективтік тілдер, көпмәнді байланыстыру.

Тукеев У.А., Рахимова Д.Р., Жуманов Ж.М., Картбаев А.Ж.

Модель преобразователя с одним состоянием для морфологии казахского и русского языков

В статье представлено исследование по построению конечных автоматов с одним состоянием для анализа двухуровневой морфологии языка со сложным строением слов, а именно, прямое преобразование их окончании в грамматические характеристики. Это проблема изучена на основе казахских и русских языков, которые являются флективными и агглютинативными по своей природе. Для решения этой проблемы применены тривиальный автомат Мели с одним состоянием и многозначное отображение. Мы изучили полноту ввода конечного автомата для анализируемых языков. Определение этой полноты для морфологического анализа гарантирует принятие конечным автоматом всех слов анализируемого языка. Для агглютинативных языков проблема определения полноты множества возможных окончаний является сложной задачей. Казахский язык является агглютинативным языком со сложной морфологией для многоуровневого машинного изучения. Далее в статье мы определим полноту множества окончаний казахского языка. Предложенная технология реализована для русско-казахского машинного перевода, оценка качество перевода выполнена метрикой BLEU.

Ключевые слова: двухуровневая морфология, машинный перевод, конечные автоматы, флективные языки, многозначные отображения.

1 Introduction

An issue of morphological analysis is important in the natural language processing. The attempt to determine a basic concept of finite state approach in the morphological analysis refers to the two-level morphology concept proposed Koskenniemi, 1983[1], and implemented through the use of finite state transducers. In this paper we consider the possibility of using single state transducer(SST) for two-level morphology of inflected languages. SST - is a trivial Mealy finite state transducer (FST), particularly, a FST with one state[2]:

$$y(t) = f_y(x(t)), \tag{1}$$

where $x(t)$ - input of the machine, $y(t)$ - output of the machine, t - current time, f_y - the output function of the machine. The advantage of SST is its high speed. Essentially, SST is a mapping $x(t)$ to $y(t)$.

Mealy machine model is generally represented by the following equations:

$$s(t+1) = f_s(s(t), x(t)), \qquad y(t) = f_y(s(t), x(t)), \tag{2}$$

where $s(t+1)$ - the state of the machine in the next time, $s(t)$ - the current state of the machine, f_s - state transition function of the automaton.

Moore FST model represented by the following equations:

$$s(t+1) = f_s(s(t), x(t+1)), \qquad y(t) = f_y(s(t)). \tag{3}$$

A distinctive feature of the Moore machine is that the output of the machine is determined only by its state.

There are many publications on the branch of using two level morphology and FST technology for different languages [3, 4, 5, 6].

In this paper the use of the SSTs for morphological analysis of the Kazakh and Russian languages is described.

2 Description of the method

Let's consider the steps of machine translation, using SSTs in a scheme of translation. Input of this scheme is a sentence of source natural language.

1. Mark out of words in a sentence.
2. Finding the characteristic part of speech for words.
3. Split words on the stem and ending.
4. Morphological analysis of words with SSTs: "ending "grammatical characteristics."
5. Translation the stem from the source language into the target language.
6. Transfer the grammatical characteristics of a source language word in grammatical characteristics of a target language word.
7. Morphological generation the endings of a target language words by the grammatical characteristics source language words using SSTs.
8. A compound words stems of the target language with the endings.
9. Implementation of structural transfers of the source language sentence to the target language sentence.

Output of this scheme is a sentence of target natural language.

Below the mapping of SSTs of steps 4 and 7 for the Kazakh and Russian languages is a more detailed study.

These mappings allow getting the corresponding word ending in the target language for each word in the source language. Joining the stem and the ending in the target language produces the required output word. After that, phrases and sentences of target natural language are produced by joining words into a sequence.

Formally, a multivalued mapping is defined as follows.

Let X, Y be discrete spaces, $P(Y)$ is a set of all subsets of Y. Then a multivalued mapping F from X onto Y is a correspondence that for each point $x \in X$ assigns an empty subset $F(x) \in Y$, called the image of point x, i.e. $F : X \to P(Y)$. We shall call this mapping m-mapping (from X onto Y).

Let Γ_F be a subset of set $X \times Y$; then $\Gamma_F = \{(x, y) \mid x \in X, y \in F(x)\}$ is called a graph of m-mapping F. A graph of m-mapping F is a tabular representation of m-mapping F, which is very important and convenient for computer representation of multivalued mappings.

Consider conversion of multivalued mapping into single-valued mapping. For these we add to set X an additional set of parameters T:

$$F : X \times T \to Y, \tag{4}$$

Then, the multivalued mapping F can be transformed into a series of single-valued mappings $\{f_i X \to Y\}, f_i(x) \in F(x)$.

The machine translation process of source language into target language based on the assumptions made by a scheme that mentioned before, especially, a multivalued mapping system for stages of morphological analysis and synthesis will be as follows:

$$F_{s^k} : X_{i^k} \to Y_{i^k}, \qquad F_{t^k} : Y_{j^k} \to Z_{j^k}, \qquad F_{st^k} : Y_{i^k} \to Y_{j^k}, \tag{5}$$

where

X_{i^k} is the space of source natural language L_i endings for the k-th part of speech; it is an input space for multivalued mapping F_{s^k};

Y_{i^k} is the space of grammatical features for the source language's k-th part of speech; it is an output space for multivalued mapping F_{s^k} ;

Y_{j^k} is the space of grammatical features for the target language's k-th part of speech; it is an output space for multivalued mapping F_{t^k};

Z_{j^k} is the space of target natural language L_j endings for the k-th part of speech; it is an output space for multivalued mapping F_{st} ;

F_{s^k} is a multivalued mapping of space of endings for the source language's k-th part of speech into space of grammatical features for the source language's k-th part of speech;

F_{t^k} is a multivalued mapping of space of grammatical features for the target language's k-th part of speech into space of endings for the target language's k-th part of speech;

F_{st^k} is a mapping of space of grammatical features for the source language's k-th part of speech into space of grammatical features for the target language.

3 Completeness of the endings of the Kazakh language

The set of endings of the Kazakh language is necessary for the construction of multi-valued maps, using the model presented before:

- $F_s : X_s \to Y_s$ (for the source language),
- $F_t : Y_t \to Z_t$ (for the target language), where X_s - source language endings,
- Y_s - grammatical characteristics of words of the source language,
- Y_t - grammatical characteristics of words of the target language,
- Z_t - the endings of the target language.

In this mapping system to ensure the correctness of transformations any word of language pair in machine translation requires that a set of endings of a target and (or) a source language was complete. Completeness of set endings of the source language is very important for the morphological analysis of the sentences of the source language, as a guarantee that every word will be analyzed in terms of its grammatical (lexical) properties.

In this paper, we consider the completeness of the endings of the Kazakh language.

Since the completeness of the system of endings of one language in a linguistic pair of machine translation indirectly determines the overall completeness of the transformed system on the lexical level from one language to another language, it is an important issue for all machine translation system.

Consider a system of Kazakh word endings: nominal endings (nouns, adjectives, numerals) and verbal endings (verbs, participles, gerunds, mood and voice).

The nominal endings of the Kazakh language have four types:

- Plural endings (denoted by K),
- Possessive endings (denoted by T),
- Case endings (denoted by C),
- Personal endings (denoted by J).

Consider all types of endings placements variants: of one type, of the two types, of the three types, and of the four types. Number of placements determined by the formula:

$$A_{n^k} = \frac{n!}{(n-k)!}. \tag{6}$$

Then, the number of placements will be determined as follows: $A_{4^1} = \frac{4!}{(4-1)!} = 4$, $A_{4^2} = \frac{4!}{(4-2)!} = 12$, $A_{4^3} = \frac{4!}{(4-3)!} = 24$, $A_{4^4} = \frac{4!}{(4-4)!} = 64$. All possible placements number is 64. Consider what placements are semantically valid.

The endings placements of one type are semantically valid. The endings placements for two types are the following: KT, TC, CJ, JK KC, TJ, CT, JT KJ, TK, CK, JC. The analysis of the semantics of the two types of endings placements shows that bold placements are valid, and the remaining placements belongs to unacceptable. For example, TK is unacceptable: after the possessive endings plural endings are not used, CK is unacceptable: after case endings are not accepted to put the plural, JC is unacceptable: after the personal endings are not accepted to put the case endings, CT is unacceptable: after case endings are not put possessive endings, JT - unacceptable: after personal endings are not put possessive endings. Belongs to the unacceptable the type JK - after personal endings plural ending, as this type of placements is covered by the plural personal endings.

In general, the types of endings T and J are the endings defining the dependence of subjects, objects, and actions. In the words with a nominal base the type TJ is possible for the cases of differences of substances (subjects, objects, actions): 'apa-ng-myn' ('apa-ng' it refers (personification) to a other subject then speaker, and the personal ending '-myn' determines the dependence (personification) to the speaker. In the type TJ double definition of a dependency to one substance (twice perconification to one substance) is prohibited, for example: 'apa-m-myn' do not say.

It should be noted that the type of endings CJ has limitations on cases ilik (genitive) and tabys(accusative).

Thus, the number of valid (correct) placements of two types of endings is 6. The endings plasements of the three types are as follows:
KTC, KTJ, TCJ, TCK,
CJK, CJT, JKT, JKC
KCJ, KCT, TJK, TJC,
CTK, CTJ, JTK, JTC
KJT, KJC, TKC, TKJ,
CKT, CKJ, JCK, JCT.

Determination of permissible placements of three types of endings do the following rule: if the placement of the three types have invalid placement of two types, this placement - unacceptable. Then, the permissible endings placements of three types is 4 (in bold).

The endings placements of the four types are as follows:
KTJC, TKJC, CKTJ, JKTC
KTCJ, TKCJ, CKJT, JKCT
KJTC, TJKC, CTKJ, JTKC
KJCT, TJCK, CTJK, JTCK
KCTJ, TCJK, CJKT, JCKT
KCJT, TCKJ, CJTK, JCTK

Determination of permissible placements endings of the four types follows this rule: if the placement of the four types has invalid placement of two types, this placement is unacceptable. Then, the permissible ending's placements of the four types will be 1 (in bold). Total permissible ending's placements of one type are 4, of two types are 6, of three types are 4, four types is one.

So, the total number of valid types of ending's placements in the nominal words is 15. To the type of endings of words with verbal stems are related to: verbs, participles, adverbs, moons, voices. The system of endings of verbs include the following types: tense (8 tense), person (3 types), negation. Then, the number of possible types of verb endings is 25. The system of participle endings include the following types: participle's base endings (denoted R), plural endings (denoted K), possessive endings (T), case endings (denoted C), personal endings (denoted J). Then, possjble variants of endings types (participle's base endings for all variants is the same) will be:
- one type endings:
RK, RT, RC, RJ;
- two type endings:
RKT, RTC, RCJ, RJK
RKC, RTJ, RCT, RJT
RKJ, RTK, RCK, RJC;
- three type endings:
RKTC, RTCJ, RCJK, RJKT
RKTJ, RTCK, RCJT, RJKC
RKCJ, RTJK, RCTK, RJTK
RKCT, RTJC, RCTJ, RJTC
RKJT, RTKC, RCKT, RJCK
RKJC, RTKJ, RCKJ, RJCT;
- four type endings:
RKTJC, RTKJC, RCKTJ, RJKTC
RKTCJ, RTKCJ, RCKJT, RJKCT
RKJTC, RTJKC, RCTKJ, RJTKC
RKJCT, RTJCK, RCTJK, RJTCK
RKCTJ, RTCJK, RCJKT, RJCKT
RKCJT, RTCKJ, RCJTK, RJCTK.
Let's consider a semantic permissibility of variants of the endings.

All variants of participleTs endings on one type of the endings are semantically permissible. The analysis of semantics of placements of two types of the participle's endings shows, that the placements allocated by bold font are permissible, and other placements is carried to unacceptable. Allowable variants of the endings of participles same as in system of the endings with nominal bases, but from them for participles are inadmissible variant RTJ as sequence ending of participle - possessive endings for participles in all cases means personification action with a verbal basis. And personification action cannot second time doing to personal ending. For example, 'bar-ghan-ym' (my arrival, my coming) is substance, but is impossible to tell 'bar-ghan-ym-syng', as action ('bar-ghan-ym') not personificable, namely, action cannot be transferred to subject.

Similarly, endings RTCJ and RKTCJ have no restrictions on two types of the endings, i.e. possible pairs of the endings inside of these types of the endings are allowable, but they break the previous rule action cannot be transferred to subject. For example, for RTCJ: 'bar-ghan-ym-gha-myn', where 'bar-ghan-ym-gha' (to my arrival - to my coming) - declination of action that cannot be presented by the subject. For RKTCJ: 'bar-ghan-dar-yng-nan-byz', where 'bar-ghan-dar-yng-nan' (from your arrivals - from yours coming) - declination of actions that cannot be presented by subjects.

Thus, the quantity of types of the endings of participles is 11. Let's consider types of the endings of verbal adverbs. They are represented by the endings of transitive future time for which follows personal endings: PJ, where P - the base ending of a verbal adverb, J - personal endings. For the given class we shall allocate only the following base endings: - ghany,-geli, -qaly,-keli. Thus, we count, that quantity of types of the endings of a verbal adverb is 1.

Let's consider the endings of moods, namely, conditional, imperative, desirable. The endings of an indicative mood coincide with the endings of verbs in the present, the past and the future. The type of the endings of declinations is similar adverbs, i.e. the base endings of moods which personal endings follow. Thus, we consider that there are three types of the endings of moods: conditional, imperative, desirable.

Types of the endings of voices, namely, reflexive, passive, joint and compulsory, also are determined under the previous scheme: the base endings of voices for which follow personal endings. And, types of the endings of voices are 4.

So, the total of types of the endings of words with verbal bases is 48. The total of the endings of words with nominal bases plus total of types of the endings of words with verbal bases equal 63.

The following task is on the received types of the endings to determine forms of the endings and their quantity. It to make simply as for each type of a part of speech are available rules. In the given direction authors construct final sets of the endings for all basic parts of speech of the Kazakh language. So, for parts of speech with nominal bases the number of endings equal 862, and the number of endings of parts of speech with verbal bases makes: verbs are 432, participles are 1588, verbal adverbs are 48, moods are 230, voices are 80. Total, '3240' is the number of all Kazakh endings.

Table 1 – Comparative evaluation of the machine translation from Russian into Kazakh for three thematic selections

Type of text	BLEU
T1	30.47
T2	31.90
T3	33.89

4 Practical results

The submitted technology is approved on a set of endings of the Bektayev model for the Kazakh language with the number of endings equal to 753 [7]. To evaluate the quality of machine translation technique we used BLEU. A comparative evaluation of the machine translation from Russian into Kazakh is done for three thematic selections: T1-text of simple sentences with from a general domain, T2-text of simple sentences from a political domain, and T3-text of simple sentences from a technical (Informatics) domain. Low level of estimation for text T3 is explained by poor of dictionary for informatics area.

5 Resume and the future works

The paper presents the application of trivial Mealy transducers with multivalued mappings for a stage of morphological analysis of inflectional languages by the example of the Kazakh and Russian languages. We investigated completeness of trivial Mealy transducers with multivalued mappings on a random input. This is highly important to guarantee a covering of the endings of analyzed language. Also it is important for indirect ensuring of their completeness for a stage of generation in machine translation. Future works include: investigation of completeness of endings for other inflective languages, for example, the Russian language; investigation of possibility to create universal tool for inflectional FST based on endings tables.

References

[1] *Koskenniemi K.* Two-level morphology: A general computational model of word-form recognition and production. // Technical report publication of the University of Helsinki. - 1983. - No.11. - p.115-159.

[2] *Gurenko V.V.* Intoduction to automata theory - M.:MGTU, 2013. - 62 p.

[3] *Oflazer K.* Two-level description of Turkish morphology // Literary and Linguistic Computing. - Stroudsburg. - 1994. - No.2. - p.137-148.

[4] *Washington J. N., Salimzyanov I., Tyers F.M.* Finite-state morphological transducers for three Kypchak languages. // Proceedings of the 9th Conference on Language Resources and Evaluation. - Reykjavik. - 2014. - pp.545-548.

[5] *Kairakbay B.M., Zaurbekov D. L.* Finite State Approach to the Kazakh Nominal Paradigm. // Proceedings of the 11th International Conference on Finite State Methods and Natural Language Processing. - St. Andrews. - 2013. - p.108-112.

[6] *Kessikbayeva G., Cicekli I.* Rule Based Morphological Analyzer of Kazakh Language // Proceedings of the 2014 Joint Meeting of SIGMORPHON and SIGFSM.- Baltimore. - 2014. - p.137-148.

[7] *Bektayev K.* Big Kazakh-Russian and Russian-Kazakh dictionary. - Almaty: Altyn Kazyna, 1999. - 704 p.