Gulshan Dovudov

DTEDI seminar
Laboratory of Natural Language Processing

April 3, 2012

Tajik Automatic Morphological Analyzer

- No proper tools and resources for NLP of Tajik language
- Tajik automatic morphological analyzer
  - divides words into morphemes(prefixes, roots, postfixes)
  - definition part of speech(roots, lemma and word)
  - formation of all word forms of a given word
- works on the base of fixed database of prefixes, roots and postfixes

Word	English	Pref	Root	POS (R)	Post	POS(W)
Китобҳо	books	-	китоб	01/sod/igw/iip/ijn/ijk/ikt/ibj/ ijb/imn/iwt/iww/	ҳо	01/sox/igw/iip/ijn/ijk/ikt/ ibj/ijb/imn/iwj/iww/
сурхтар	More red		сурх	02/sod/sas/sir/sdo/	тар	02/sod/sas/sir/sol/
нагуфтам	I didn't say	на	гуфт	05/sod/fts/azg/wse/wtn/fms/ fgi/fgz/fbv/fnm/sfx/gnz/	ам	05/sox/fts/azg/wyk/wtn/fms/ fik/fgz/fbv/fnm/sfx/gnz/
панчум	Fifth	-	панч	03/sod/wmi/wma/	ум	03/sox/wtr/

Short Characteristics of Tajik Language

- closely related to Persian languages
  - Farsi(Iran)
  - Dari(Afganistan)
  - belongs to Iranian branch of Indo-Iranian languages
- inflectionally belongs to analytical type of languages
- noun are characterized by the number and defined/undefined categories
  - дарахт[daracht]-tree, дарахтхо[darakhto]-trees
  - дарахте[darachte]-a tree, дарахтхое[darakhtoe]-some trees
- verb is characterized by person, time, voice and mood categories
  - нагуфтам[naguftam] - I did not say
- no grammatical categories of gender and case
  - китоб[kitob] - book, мард[mard] - man

Short Characteristics of Tajik Language cont...

- Izafet - main means of word connection in the sentence
  - шахри бузург[shahri buzurg] – a big city
  - китоби ман[kitobi man] – my book
- Китоби сурхи ман[kitobi surkhi man] - my red book
- Ман ба мактаб меравам[man ba maktab meravam] – I am going to school
- Ту ба мактаб нарафти[tu ba maktab narafti'] – You did not go to school
- Шахри бузург[shahri buzurg] – a big city, китоби ман[kitobi man] – my book

Alphabet

- Arabic alphabet
  - additional 4 letters peculiar to the Persian language
- Latin alphabet
  - additional 5 letters peculiar to the Tajik language
- Russian cyrilics
  - additional 6 letters peculiar to the Tajik language
- Tajik is nominative language.
  - main core vocabulary consist of the original old Persian
  - many borrowings from Arabic, Turkish, Russian...

Database of Prefixes of Tajik Language

- prefixes are divided in two groups
  - simple
  - compound
    - double
    - triple
- compound is combination of various not repeated simple prefixes
- by using combinatorial-statistical methods(assuming complete list of 19 common prefixes) identified
  - 19 common prefixes
  - 44 real double
  - 9 real triple

level	count	%	Example
0	1	94,2352889	Кӯча (street), мох (moon),
1	19	5,44170827	ҳам-хона (roommate)
2	44	0,31679127	на-ме-гуфт (He Didn't say),
3	9	0,00621159	на-ме-фар-овард he didn't lower
Sum	73	100	

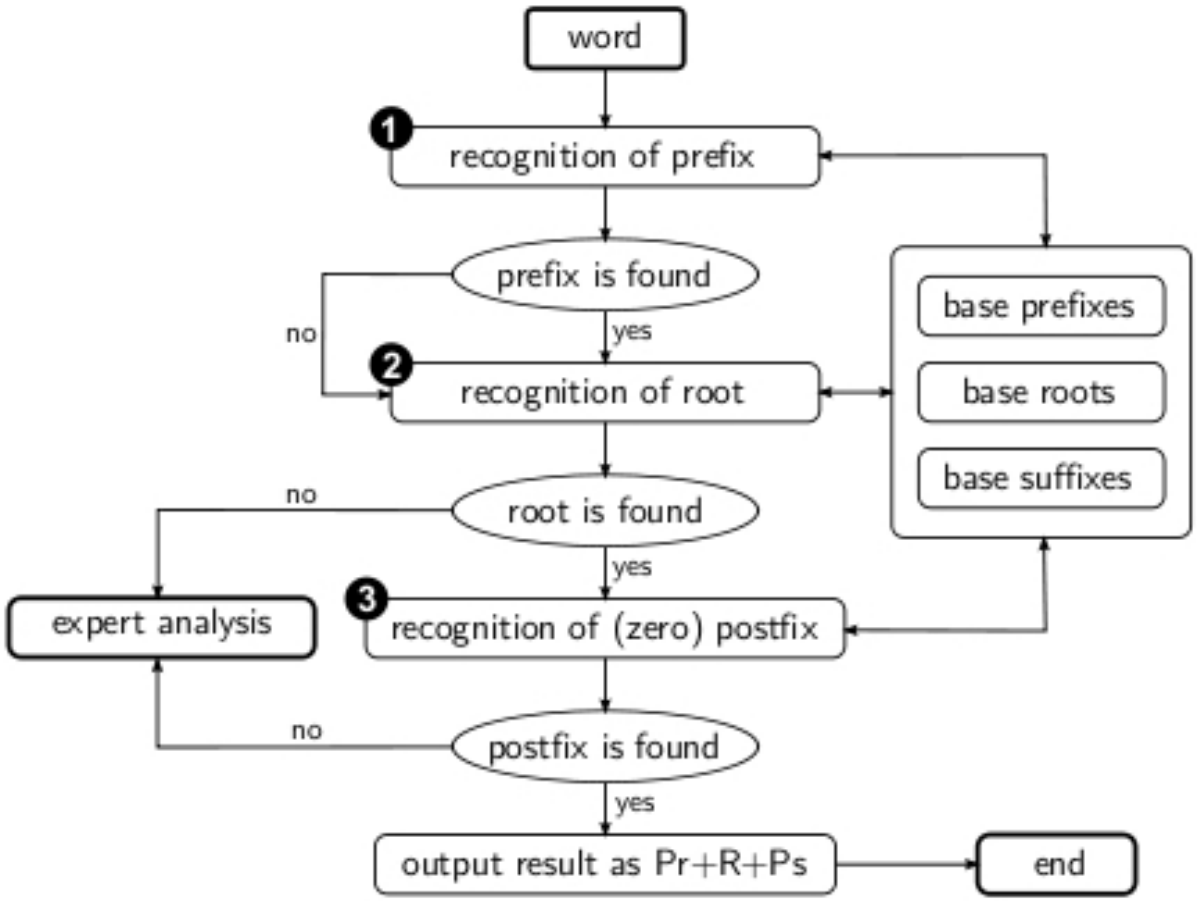
Database of Postfixes In Tajik Language

- postfixes are divided in two groups
  - simple(113 known simple postfixes)
  - compound(up to 8 levels)
- on the base of iterative procedures which applied to representative texts
  - 2533 suffixes with their frequency of occurrence were found

Level	Count	%	Example
0		46.89649933	
1	113	39.25152871	Кор-гар
2	755	11.1242134	Кор-гар-он
3	1017	2.359061804	Кор-гар-ак-он
4	540	0.355709897	Кор-гар-ча-ҳо-ям
5	86	0.01142843	Кор-гар-ча-ҳо-е-ро
6	17	0.001298685	Кор-гар-ча-ҳо-ятон-ро-ву
7	3	0.000194803	Хирад-манд-тар-ин-ҳо-яшон-ро-ву
8	2	6.49343E-05	Сӯз-он-ид-а-ги-ҳо-яшон-ро-ву

Morphological Analysis of Tajik Words

- assume that every word which will be analyzed is correct(no errors in spelling)
- morphological analysis is based on priori specified fixed base morphemes(roots, prefixes, postfixes)



The Parts of Speech of Tajik Language

- there are 6 primary and 5 auxiliary parts of speech
  - verb is divided into 4 subparts - Verb, Infinitive, Gerund and Participle
  - we add to the list Sound imitating words (onomatopoia) and Quantifiers
  - program uses 11 main parts of speech and 5
- for each part of speech group of features is identified
- program uses 207 features which are grouped on the basis of part of speech

#	POS	#	POS	#	POS	#	POS
1	Noun	5	Verb	9	Adverb	13	Particles
2	Adjective	6	Infinitive	10	Preposition	14	Interjections
3	Numeral	7	Gerund	11	Postposition	15	Sound imitating words
4	Pronoun	8	Participle	12	Conjunction	16	Numerativ

Defining Part of Speech

- identifying parts of speech and features of each root by an expert
- at present from 63,570 roots there is defined POS for 2,379 roots which cover 67 %
- For each prefix and postfix there is also defined description
- while defining the word speech the lemma is also defined

Word	English	Pref	Root	POS (R)	Post	POS(W)	Lemma
Китобҳо	books	-	китоб	01/sod/igw/iip/ ijn/ijk/ikt/ibj/ijb/ imn/iwt/iww/	ҳо	01/sox/igw/iip/ijn /ijk/ikt/ibj/ijb/imn /iwj/iww/	Китоб
коргарон	Workers		коргар	01/sod/igw/imf/ ijn/ijk/iat/ibj/ijb/ imn/iwt/iww/	гарон	01/sox/iws/iip/ijn /ijk/ikt/ijd/i80/im n/iwj/iww/	Коргар

Creating All Word Forms From a Given Word

- for creating all word forms would be taken into consideration
  - the POS of given word
  - form of prefix of postfix (word making or wordforming)
- for one verb there more than 1200 word forms can be created
- for one noun there more than 800 word forms can be created

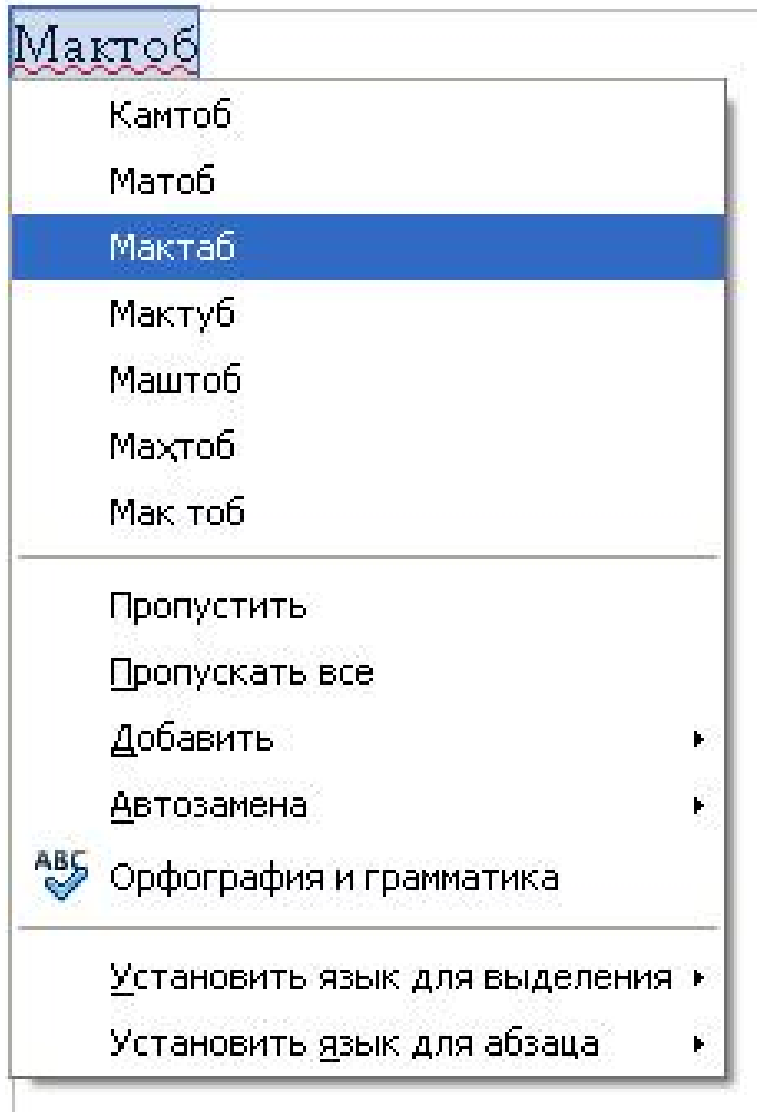
id	word	pref	root	suf	id	word	pref	root	suf
1	кӯбам		кӯб	+ам	11	кӯбеду		кӯб	+ед-у
2	кӯбаму		кӯб	+ам-у	12	кӯбанд		кӯб	+анд
3	кӯбӣ		кӯб	+ӣ	13	кӯбанду		кӯб	+анд-у
4	кӯбиву		кӯб	+ӣ-ву	14	бикӯбам	би	кӯб	+ам
5	кӯбию		кӯб	+ӣ-ю	15	бикӯбаму	би	кӯб	+ам-у
6	кӯбад		кӯб	+ад	16	бикӯбб	би	кӯб	+ӣ
7	кӯбаду		кӯб	+ад-у	17	бикӯбиву	би	кӯб	+ӣ-ву
8	кӯбем		кӯб	+ем	18	бикӯбию	би	кӯб	+ӣ-ю
9	кӯбему		кӯб	+ем-у	19	бикӯбад	би	кӯб	+ад
10	кӯбед		кӯб	+ед	20	бикӯбаду	би	кӯб	+ад-у

The Way of Controlling Morphological Analyzer

- for each root defined a list of prefixes and suffixes or a combination of prefixes or suffixes that cannot be connected to it
- for each prefix and postfix POS is defined to which they can be connected
- prefixes "но-", "то-", "фар-", suffixes "ок*", "дон*", "аЗо*", "ака*", "ги*", "гб*", "гор*", "дон*", "ин*", "ум*", "ур*" and combination of prefixes and suffixes "бар-+ат*" "бар-е*" "бар-+и*" "би-го3" can not be connected to the root "зан"
- preposition, postposition and interjection can not be connected to any prefixes or postfixes
- sound imitating words(onomatopoia) can be connected only with the suffixes "ас", "ос" and "б"
- postfixes "вар", "гар", "зор", "сор" and "дон" can be connected only to nouns

Application of Results

- which using the result of this work module of checking the spelling of Tajik words was created for Open Office Org



Goals

- In future, the result of this work can be used for the following purposes:
  - Obtaining statistical parameters for the particular POS's
  - Preparation of concordances from the Tajik corpus
  - Automatic generation of frequency dictionaries
  - Corpus tagging (deciding whether дустам[dustam] is my friend or I am friend etc.)
  - Creating a program which will decide which of the possible morphological tags of a word form is correct in a given sentence context.
  - Noun/verbal phrase detection
  - Tajik-OtherLanguage dictionary improvement (dictionary which can translate дустам[dustam], дустхо[dustho] (Eng. Transl.) etc. not only the lemma дуст).
  - Adding Tajik language to a selected search system,
  - Machine translation related to Tajik

Published Articles

- Usmanov, Z. D., Dovudov, G. M. On forming the prefix base to the literary tajik. Reports of the Academy of Sciences of the Republic of Tajikistan, no. 6, 2009, pp 431-436.
- Usmanov, Z. D., Soliev, O. M., Dovudov, G. M. On a set of postfixes of tajik literature language. Reports of the Academy of Sciences of the Republic of Tajikistan, no. 2. 2010, pp 99-103.
- Usmanov, Z. D., Dovudov, G. M. A frequency morphemic dictionary of literary tajik. Reports of the Academy of Sciences of the Republic of Tajikistan, no. 4. 2010, pp 188-191.
- Gulshan Dovudov, Vit Baisa.: Morphological Analysis of Tajik: Notes and Preliminary Results. In: Proceedings of the Fourth Workshop on Recent Advances in Slavonic Natural Language Processing, RUSLAN 2010. Masaryk University, Brno (2010)
- Gulshan Dovudov, Jan Pomikalek, Vit Suchomel, Pavel Smerk.: Building a 50M Corpus of Tajik Language. In: Proceedings of the Fifth Workshop on Recent Advances in Slavonic Natural Language Processing, RUSLAN 2011. Masaryk University, Brno (2011)
- Z.D. Usmanov, Dovudov G.M., Soliev O.M. Tajikcomputermorphological analyzer// License(information resource)reported ZI-03.2.220 TJ, 20.12.2011. National Patent Information Centre.Ministry of Economic Development and Trade ofthe Republic of Tajikistan
- Z.D.Usmanov, Soliev O.M., Dovudov G.M.//Tajiklanguagepackage for System Open Office.Org.//License(information resource)reported ZI-03.2.222 TJ, 11.01.2012. NationalPatentInformation Centre.Ministry of Economic Development and Trade ofthe Republic of Tajikistan