Grammatical Dictionary Generation Using Machine Learning Methods

June 4, 2016

Marina Mazurova

Moscow, Russia

sleepofnodreaming12@gmail.com

Grammatical Dictionaries

- Grammatical Dictionary of Russian by A. Zaliznyak
- Polish Grammatical Dictionary
- Bulgarian Grammatical Dictionary
- Bashkir Grammatical Dictionary
- etc.

What's the Purpose?

- Generally, I aim to develop the automatic dictionary generation system
- In this work, I am investigating a possible approach to the task

Input: Looking for Compromises

- Relatively large (at least about several hundred thousand tokens) collection of unannotated texts
- Machine-readable grammar — a description of language's inflectional types (paradigms)
- Some additional data (details will follow)

Idea: Two-Stage Algorithm

- Draft generation: compilation of a collection of all lexeme hypotheses possible (so-called pseudolexemes) according to a lexeme formation algorithm

- Draft filtration: removal of pseudolexemes supposed to be false

The Data: Corpora

Language	Corpus size, word usages	Frequency distribution size, tokens
Greek	37,554,322	672,806
Albanian	19,543,008	346,635
Udmurt	6,368,427	283,070
Kazakh	904,561	107,715
Katharevousa	359,805	48,391

Algorithm: Draft Generation

- UniParser Grammar ➙ List of affixes
- List of affixes ➙ Finite-state automaton
- For each word form, all possible stem-affix divisions are found
- All word forms' parsing options are transformed into a draft by joining all word forms having the same stem combined with affixes attributed to the same paradigm into a pseudolexeme

More about Pseudolexemes

As a noun

πληγ.

.ὰς	pl,acc
.ὴ/.ή	sg,nom
…	
.ῶν	gen
.άς	pl,acc

As an adjective

πληγ.

.άς	pos,f,pl,acc
.ή/.ὴ	pos,f,sg,nom
…	
.ῶν	pos,pl,gen
.ὰς	pos,f,pl,acc

Algorithm: Performance

Depending on the length of the input word list

FD size, word forms	Time, sec
26,926	1,855
53,852	3,523
80,778	6,797
107,704	7,005

Depending on the number of inflections

Paradigms	Number of inflections	Time, sec
N-soft	766	2,01
All nominal	3,104	4,29
All verbal	82,260	5,85

Houston, We Have a Problem: Misparsing

сұлуысың ‘beauty’ (Kazakh, noun)

сұлуы.

.	imper,2,sg / indic,prs,3
.мыз	indic,prs,1,pl
.сың	indic,prs,2,sg

The Data: Dictionaries

Language	Dictionary size, lex
Katharevousa	403 (adjectives only)
Udmurt	21656
Albanian	45861
Kazakh	22024 / 14527 (there were two overlapping dictionaries)

Draft Filtration: Data Sets

Language	Lexemes formed, number of		Valid lexemes, number of		Valid lexemes, %	
	Full	Cut	Full	Cut	Full	Cut
Albanian	2,047,093	799,004	5,635	4,378	0,27	0,54
Kazakh	62,704	23,354	7,479	5,155	11,9	22,1
Katharevousa	3,959	—	370	—	9,3	—
Udmurt	278,036	101,577	7,728	5,789	2,7	5,7

Draft Filtration: Features

- Purely distributional features: variance- and entropy-based. Word form frequency distribution inside a paradigm's used
- Grammatical category distribution features
- Features interpreting a proportion of word forms postulated and found

Draft Filtration: Evaluation

- Standard precision, recall and F measure
- Weighted measures

$$P_f = \frac{\sum_{f \in \pi} \sum_{w \in tp} \sum_{s \in w} I(f,s)}{\sum_{f \in \pi} \sum_{w \in t} \sum_{s \in w} I(f,s)} \qquad R_f = \frac{\sum_{f \in \pi} \sum_{w \in tp} \sum_{s \in w} I(f,s)}{\sum_{f \in \pi} \sum_{w \in p} \sum_{s \in w} I(f,s)}$$

$$F_f = \frac{2 \cdot P_f \cdot R_f}{P_f + R_f}$$

I(f, s) — a function defining a number of tokens may be represented as a combination of an inflection f and a stem s in a corpus.

Classification of Kazakh pseudolexemes

Model	Set	P	P_f	R	R_f	F	F_f
Perceptron	full	0,899	0,732	0,011	0,072	0,022	0,131
Perceptron	cut	0,729	0,729	0,045	0,094	0,085	0,167
Linear Regression	full	0,956	0,617	0,001	0,026	0,2	0,05
Linear Regression	cut	0	0	0	0	0	0
SVM	full	0,34	0,592	0,03	0,259	0,055	0,36
SVM	cut	0,316	0,591	0,03	0,259	0,055	0,36
Random Forest	full	0,54	0,669	0,299	0,525	0,385	0,589
Random Forest	cut	0,505	0,701	0,31	0,571	0,384	0,629

The Albanian Problem

Model	P	R	F
Perceptron	0,039	0,21	0,0658
Linear Regression	0	0	0
SVM	0	0	0
Random Forest	0,053	0,003	0,0057

ngarkesë '*load*' (noun)

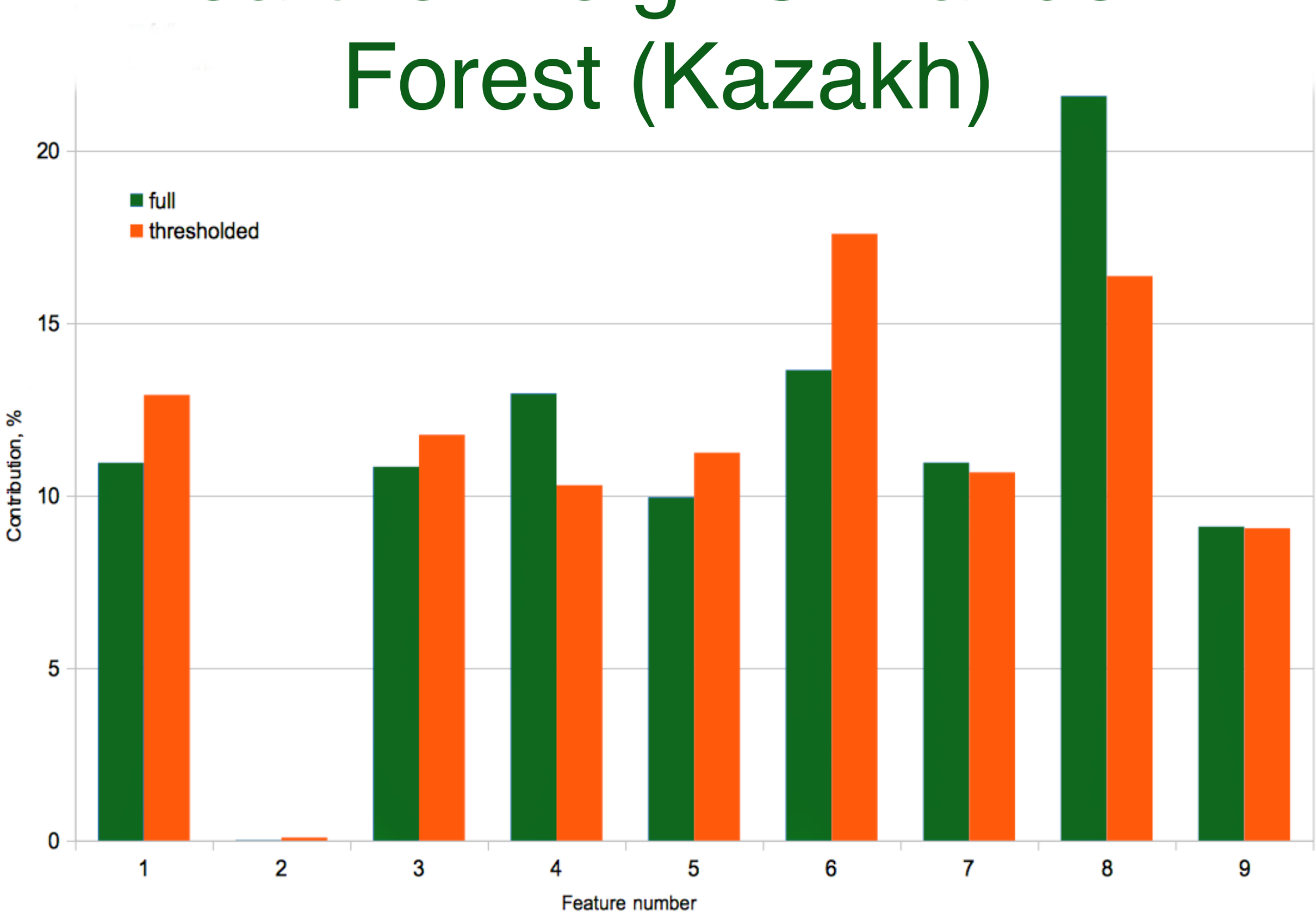
ngarkes. + .e (parsed with adjective paradigms only):

- attributed to 16 paradigms
- each pseudolexeme contains 6 different grammatical forms

Using Another Language's Data as a Training Set

Model	Training Set	P_f		R_f	
		U	K	U	K
SVM	full	0,292	0,065	0,743	0,415
SVM	cut	0,295	0,073	0,744	0,476
Random Forest	full	0,092	0	0,045	0
Random Forest	cut	0,071	0,195	0,039	0,075

Feature Weights: Random Forest (Kazakh)



Thank You!

Project source code:
https://github.com/sleepofnodreaming/gramdicmaker2016
Email me:
sleepofnodreaming12@gmail.com