Automatic detection of morphological paradigms using corpora information

Alexey Sorokin[1,2,3], Irina Khomchenkova[1]

[1]Moscow State University, [2]Moscow Institute of Science and Technology,
[3]General Internet Corpus of Russian

Dialogue
22nd International Conference on Computational Linguistics
Moscow, RSUH, 4th June, 2016

Problem formulation

- Goal: detect an inflection paradigm for a base form.

Problem formulation

- Goal: detect an inflection paradigm for a base form.
- Practical application:
  - Machine translation (word form synthesis, *корова+N+Pl+Instr* $\mapsto$ *коровами*),

Alexey Sorokin, Irina Khomchenkova
[1]Moscow State University, [2]Moscow Institute of Science and Technology, [3]General Internet Corpus of Russian

Problem formulation

- Goal: detect an inflection paradigm for a base form.
- Practical application:
  - Machine translation (word form synthesis, *корова+N+Pl+Instr* $\mapsto$ *коровами*),
  - Extension of morphological resources (`wiktionary.org`),

Alexey Sorokin, Irina Khomchenkova
[1]Moscow State University, [2]Moscow Institute of Science and Technology, [3]General Internet Corpus of Russian

Problem formulation

- Goal: detect an inflection paradigm for a base form.
- Practical application:
  - Machine translation (word form synthesis, *корова+N+Pl+Instr* $\mapsto$ *коровами*),
  - Extension of morphological resources (`wiktionary.org`),
  - Lemmatization, morphological analysis

Alexey Sorokin, Irina Khomchenkova
[1]Moscow State University, [2]Moscow Institute of Science and Technology, [3]General Internet Corpus of Russian

Problem formulation

- Goal: detect an inflection paradigm for a base form.
- Practical application:
  - Machine translation (word form synthesis, *корова+N+Pl+Instr* $\mapsto$ *коровами*),
  - Extension of morphological resources (`wiktionary.org`),
  - Lemmatization, morphological analysis
- Principal problems:
  - How to encode a paradigm?

Alexey Sorokin, Irina Khomchenkova
[1]Moscow State University, [2]Moscow Institute of Science and Technology, [3]General Internet Corpus of Russian

Problem formulation

- Goal: detect an inflection paradigm for a base form.
- Practical application:
  - Machine translation (word form synthesis, *корова+N+Pl+Instr* $\mapsto$ *коровами*),
  - Extension of morphological resources (`wiktionary.org`),
  - Lemmatization, morphological analysis
- Principal problems:
  - How to encode a paradigm?
  - How to detect a paradigm?

Alexey Sorokin, Irina Khomchenkova
[1]Moscow State University, [2]Moscow Institute of Science and Technology, [3]General Internet Corpus of Russian

Abstract paradigms

- Abstract paradigms – a way to encode inflection tables.

Abstract paradigms

- Abstract paradigms – a way to encode inflection tables.
- First, we extract the longest common subsequence (LCS) from each inflection table.

песок	**песок**
песка	**песк**а
песку	**песк**у
песок	**песок**
песком	**песк**ом
песке	**песк**е

Alexey Sorokin, Irina Khomchenkova
[1]Moscow State University, [2]Moscow Institute of Science and Technology, [3]General Internet Corpus of Russian

Abstract paradigms

- Abstract paradigms – a way to encode inflection tables.
- First, we extract the longest common subsequence (LCS) from each inflection table.
- Second, we assign piecewise discontinuous subsequences to variables.

песок	**пес**о**к**	1+о+2
песка	**пес**к**а**	1+2+а
песку	**пес**к**у**	1+2+у
песок	**пес**о**к**	1+о+2
песком	**пес**к**ом**	1+2+ом
песке	**пес**к**е**	1+2+е
LCS = песк	x_1 = пес	x_2 = к

- Abstract paradigm:

  1+о+2#1+2+а#1+2+у#1+о+2#1+2+ом#1+2+е#...

Alexey Sorokin, Irina Khomchenkova
[1]Moscow State University, [2]Moscow Institute of Science and Technology, [3]General Internet Corpus of Russian

Abstract paradigms

- Abstract paradigms – a way to encode inflection tables.
- First, we extract the longest common subsequence (LCS) from each inflection table.
- Second, we assign piecewise discontinuous subsequences to variables.

песок	**пес**о**к**	1+о+2
песка	**пес**к**а**	1+2+а
песку	**пес**к**у**	1+2+у
песок	**пес**о**к**	1+о+2
песком	**пес**к**ом**	1+2+ом
песке	**пес**к**е**	1+2+е
LCS = песк	x_1 = пес	x_2 = к

- Abstract paradigm:

  1+о+2#1+2+а#1+2+у#1+о+2#1+2+ом#1+2+е#...
- Words *кусок*, *моток*, ... have the same abstract paradigm.

Alexey Sorokin, Irina Khomchenkova
[1]Moscow State University, [2]Moscow Institute of Science and Technology, [3]General Internet Corpus of Russian

Technical details

- LCS is extracted by means of finite state automata.
- Gap length inside LCS and in its beginning is bounded (max. 2)

Technical details

- LCS is extracted by means of finite state automata.
- Gap length inside LCS and in its beginning is bounded (max. 2)
- LCS allocation can be ambiguous:

grammeme	variant 1	variant 2
им.ед	**песок**	**песок**
...	...	...
род.мн	**песков**	**песков**
...	...	...
твор.ед	**песком**	**песком**
...	...	...

Alexey Sorokin, Irina Khomchenkova
[1]Moscow State University, [2]Moscow Institute of Science and Technology, [3]General Internet Corpus of Russian

Technical details

- LCS is extracted by means of finite state automata.
- Gap length inside LCS and in its beginning is bounded (max. 2)
- LCS allocation can be ambiguous:

grammeme	variant 1	variant 2
им.ед	**песок**	**песок**
...	...	...
род.мн	**песков**	**песков**
...	...	...
твор.ед	**песком**	**песком**
...	...	...

- We minimize the number of variables.

Alexey Sorokin, Irina Khomchenkova
[1]Moscow State University, [2]Moscow Institute of Science and Technology, [3]General Internet Corpus of Russian

Technical details

- LCS is extracted by means of finite state automata.
- Gap length inside LCS and in its beginning is bounded (max. 2)
- LCS allocation can be ambiguous:

grammeme	variant 1	variant 2
им.ед	**песок**	**песок**
...	...	...
род.мн	**песков**	**песков**
...	...	...
твор.ед	**песком**	**песком**
...	...	...

- We minimize the number of variables.
- Then we minimize the total length of gaps.

Technical details

- LCS is extracted by means of finite state automata.
- Gap length inside LCS and in its beginning is bounded (max. 2)
- LCS allocation can be ambiguous:

grammeme	variant 1	variant 2
им.ед	**песок**	**песок**
...	...	...
род.мн	**песков**	**песков**
...	...	...
твор.ед	**песком**	**песком**
...	...	...

- We minimize the number of variables.
- Then we minimize the total length of gaps.
- Finally, we minimize the number of length 0 gaps.

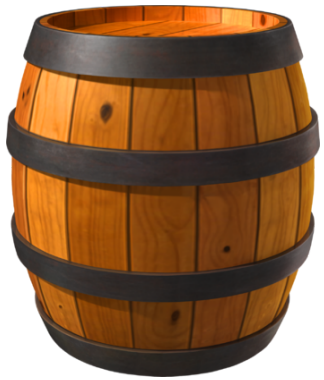
Automatic detection of paradigms

- We encode inflection tables by abstract paradigms.
- How to detect a paradigm by a word form, especially if it's a non-vocabulary one?

Alexey Sorokin, Irina Khomchenkova
[1]Moscow State University, [2]Moscow Institute of Science and Technology, [3]General Internet Corpus of Russian

Automatic detection of paradigms

- We encode inflection tables by abstract paradigms.
- How to detect a paradigm by a word form, especially if it's a non-vocabulary one?
- Rules don't remove the ambiguity.
  - *волчонок* vs *бочонок*

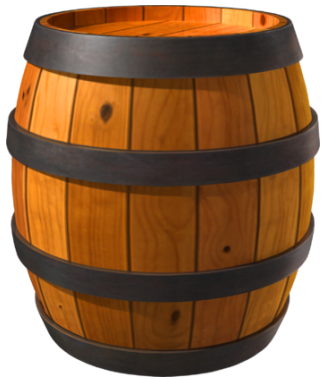


  - *голос* vs *колос* vs *волос*

Alexey Sorokin, Irina Khomchenkova
[1]Moscow State University, [2]Moscow Institute of Science and Technology, [3]General Internet Corpus of Russian

Automatic detection of paradigms

- We encode inflection tables by abstract paradigms.
- How to detect a paradigm by a word form, especially if it's a non-vocabulary one?
- Rules don't remove the ambiguity.
  - *волчонок* vs *бочонок*



  - *голос* vs *колос* vs *волос*
- We will detect paradigms with machine learning.
- Our goal: to evaluate the quality of this method.

Alexey Sorokin, Irina Khomchenkova
[1]Moscow State University, [2]Moscow Institute of Science and Technology, [3]General Internet Corpus of Russian

Initial data

Initial data: inflection tables from wiktionary.org

	наст.	прош.	повелит.
Я	желáю	желáл желáла	—
Ты	желáешь	желáл желáла	желáй
Он Она Оно	желáет	желáл желáла желáло	—
Мы	желáем	желáли	—
Вы	желáете	желáли	желáйте
Они	желáют	желáли	—
Пр. действ. наст.	желáющий		
Пр. действ. прош.	желáвший		
Деепр. наст.	желáя		
Деепр. прош.	желáв, желáвши		
Пр. страд. наст.	желáемый		
Пр. страд. прош.	желáнный		
Будущее	буду/будешь… желáть		

Alexey Sorokin, Irina Khomchenkova
[1]Moscow State University, [2]Moscow Institute of Science and Technology, [3]General Internet Corpus of Russian

Initial data

- Initial data: inflection tables from wiktionary.org. We are trying to reconstruct a table by its base form (lemma).
- 5000 most frequent nouns and verbs, not accentuated.

Initial data

- Initial data: inflection tables from wiktionary.org. We are trying to reconstruct a table by its base form (lemma).
- 5000 most frequent nouns and verbs, not accentuated.
- Nouns: 12 forms (6 cases $\times$ 2 numbers).
- Verbs: 13 forms (infinitive, 6 present forms, 4 past forms, 2 imperative forms).

Initial data

- Initial data: inflection tables from wiktionary.org. We are trying to reconstruct a table by its base form (lemma).
- 5000 most frequent nouns and verbs, not accentuated.
- Nouns: 12 forms (6 cases $\times$ 2 numbers).
- Verbs: 13 forms (infinitive, 6 present forms, 4 past forms, 2 imperative forms).
- If a paradigm is defective, we keep unfilled cells empty.

Initial data

- Initial data: inflection tables from wiktionary.org. We are trying to reconstruct a table by its base form (lemma).
- 5000 most frequent nouns and verbs, not accentuated.
- Nouns: 12 forms (6 cases $\times$ 2 numbers).
- Verbs: 13 forms (infinitive, 6 present forms, 4 past forms, 2 imperative forms).
- If a paradigm is defective, we keep unfilled cells empty.
- Nouns and verbs are classified separately.

Features for classification

- Features for classification: suffixes (max. length: 5) and prefixes (only for verbs, max. length: 3).

Features for classification

- Features for classification: suffixes (max. length: 5) and prefixes (only for verbs, max. length: 3).
- Features are encoded with a binary scheme:

	$а	$к	$ка	$ла	$ик	$рка	...
арка	1	0	1	0	0	1	...
школа	1	0	0	1	0	0	...
блик	0	1	0	0	1	0	...
...	...	...	...	...	...	...	...

Features for classification

- Features for classification: suffixes (max. length: 5) and prefixes (only for verbs, max. length: 3).
- Features are encoded with a binary scheme:

	$a	$к	$ка	$ла	$ик	$рка	...
арка	1	0	1	0	0	1	...
школа	1	0	0	1	0	0	...
блик	0	1	0	0	1	0	...
...	...	...	...	...	...	...	...

- Verbs: we calculate the length of suffixes without *-ся*, *-сь*.

Features for classification

- Features for classification: suffixes (max. length: 5) and prefixes (only for verbs, max. length: 3).
- Features are encoded with a binary scheme:

	$а	$к	$ка	$ла	$ик	$рка	...
арка	1	0	1	0	0	1	...
школа	1	0	0	1	0	0	...
блик	0	1	0	0	1	0	...
...	...	...	...	...	...	...	...

- Verbs: we calculate the length of suffixes without *-ся*, *-сь*.
- We select features (10%) by a class probability.

$$\max_i P(c(L) = c_i | f_j(L) = 1)$$

Features for classification

- Features for classification: suffixes (max. length: 5) and prefixes (only for verbs, max. length: 3).
- Features are encoded with a binary scheme:

	$а	$к	$ка	$ла	$ик	$рка	...
арка	1	0	1	0	0	1	...
школа	1	0	0	1	0	0	...
блик	0	1	0	0	1	0	...
...	...	...	...	...	...	...	...

- Verbs: we calculate the length of suffixes without *-ся*, *-сь*.
- We select features (10%) by a class probability.

$$\max_i P(c(L) = c_i | f_j(L) = 1)$$

- Classification algorithm — logistic regression.

Results of classification

- 50% of training data, 5-fold random partition.

	Per-table	Per-form
Nouns	77.38	93.50
Verbs	76.30	88.86

Results of classification

- 50% of training data, 5-fold random partition.

	Per-table	Per-form
Nouns	77.38	93.50
Verbs	76.30	88.86

- Size of training data increases – minor accuracy growth (1%).

Results of classification

- 50% of training data, 5-fold random partition.

	Per-table	Per-form
Nouns	77.38	93.50
Verbs	76.30	88.86

- Size of training data increases – minor accuracy growth (1%).

Main errors:

Animacy	*швед, одиночка*
Gender of nouns ending with *-ль*	*артель, дизель*
Forms of imperative mood	*будоражить*(**будоражи*)
Root alternation	*отозвать*(**отозву*)
Stress and e/ё	*свистнуть*(**свистнёшь*)

Alexey Sorokin, Irina Khomchenkova
[1]Moscow State University, [2]Moscow Institute of Science and Technology, [3]General Internet Corpus of Russian

Character n-grams

- Some word forms can be eliminated for phonological reasons (*осуществься*).

Character n-grams

- Some word forms can be eliminated for phonological reasons (*осуществься*).
- Features: logarithmic probabilities according to basic classifier and to character n-gram model.

Character n-grams

- Some word forms can be eliminated for phonological reasons (*осуществься*).
- Features: logarithmic probabilities according to basic classifier and to character n-gram model.
- We use 5-grams, trained on word forms from training set.

Task	No character scores		Character scores as features	
Nouns	77.38	93.50	77.42	93.50
Verbs	76.30	88.86	80.37	90.92

- No improvement for nouns...

Corpora features

- Correct detection of a noun paradigm is problematic without corpora information.

Corpora features

- Correct detection of a noun paradigm is problematic without corpora information.
- Additional features:
  - Total logarithmic frequency of unique word forms:
    $C = \sum_{j=1}^{m} \log C(w_j)$
  - This doesn't help in the case of animacy/inanimacy.

Corpora features

- Correct detection of a noun paradigm is problematic without corpora information.
- Additional features:
  - Total logarithmic frequency of unique word forms:
    $C = \sum_{j=1}^{m} \log C(w_j)$
  - This doesn't help in the case of animacy/inanimacy.
  - Discrepancy between expected frequency of word forms and the observed one.

$$D(\mathcal{N}, \mathcal{P}) = \sum_j q_j \log \frac{q_j}{p_j} \cdot \log N$$

q_j	—	frequency of j-th word form of predicted paradigm (relative, among all of its word forms)
p_j	—	expected relative frequency of j-th word form of predicted paradigm
N	—	total frequency of paradigm word forms

Alexey Sorokin, Irina Khomchenkova
[1]Moscow State University, [2]Moscow Institute of Science and Technology, [3]General Internet Corpus of Russian

Final results

- Frequencies were obtained from Russian National Corpus (`http://www.ruscorpora.ru/corpora-freq.html`)

Final results

- Frequencies were obtained from Russian National Corpus (`http://www.ruscorpora.ru/corpora-freq.html`)
- Classification results:

Method	Nouns		Verbs	
Basic	77.38	93.50	76.30	88.83
+ngrams	77.42	93.50	80.37	90.92
+counts	80.21	95.34	84.30	93.81
+counts+divergencies	82.73	95.67	83.66	93.73
+counts+divergencies+ngrams	**82.80**	**95.71**	**86.51**	**95.66**

Comparison with other systems

SIGMORPHON 2016 Shared Task: Morphological Reinflection

Language	Winner	Our system	Place
Arabic	95.47	83.52	3
Finnish	96.8	87.86	5
Georgian	98.5	94.8	4
German	95.8	94.22	3
Hungarian	99.3	89.82	5
Russian	91.46	89.67	3
Spanish	98.84	98.76	3
Turkish	98.93	92.03	3

Спасибо за внимание!

Thank you for attention!