Named-entity recognition for Russian using BiLSTM

https://github.com/lilaspourpre/ner_svm

Irina Nikishina

31 March 2018

MCL171

Table of contents

Introduction

Issue relevance

The Language itself

- issue is mostly solved for English
- no open and efficient tools for Russian

Dialog Evaluation

- FactRuEval 2016

Subtask of other NLP tasks

- Information Extraction
- Co-reference resolution
- ...

FactRuEval

- devset (122 texts, ≈ 30940 tokens)
- testset(132 texts, ≈ 59382 tokens)
- files structure:

Сегментация на токены и предложения (*.tokens)

Каждая строка - один токен. Предложения разделены пустой строкой.

Описание одного токена состоит из следующих полей:

- id токена
- позиция начала токена (от начала текста)
- длина токена
- текст токена

Разделитель полей - пробел. В токене пробела быть не может.

Спаны (*.spans)

Каждая строка - один спан. Разделитель полей - пробел.

Поля:

- id спана
- тип спана
- позиция первого символа спана от начала текста
- длина спана в символах
- первый токен спана
- длина спана в токенах

Справочно (после решётки):

- все id входящих токенов
- все тексты входящих в спан токенов

Упоминания объектов (*.objects)

Каждая строка - одно упоминание объекта. Разедлитель полей - пробел.

Поля:

- id упоминания
- тип упоминания
- список идентификаторов входящих в упоминание спанов

Справочно (после решётки):

- текст всех входящих в упоминание объекта спанов

BILOU tagging

B	I	L	O	U
'beginning'	'inside'	'last'	'outside'	'unit'

Дональд	*Трамп*	*приехал*	*в*	*Москву*	*с*	*визитом*
BPer	LPer	O	O	ULoc	O	O

в	*ПАО*	«	*Газпром*	»	.
O	Borg	O	LOrg	O O	

Output format

- Three types of entities: Person, Location, Org

start_index	length	type
0	12	Person
44	7	Location
69	20	Org
...	...	...

Methodology

Feature creation

Word Level

word case

word length

special characters

letters type

part of speech

morphological case

affixes

Text level

position

case concordance

document frequency

lowercase in context

case (window=2)

part-of-speech (window=2)

morphological case (window=2)

punctuation (window=2)

letters type (window=2)

Global context level

word embeddings

Previous approaches

Supervised learning

- SVM

Neural networks

- biLSTM

Hypothesis

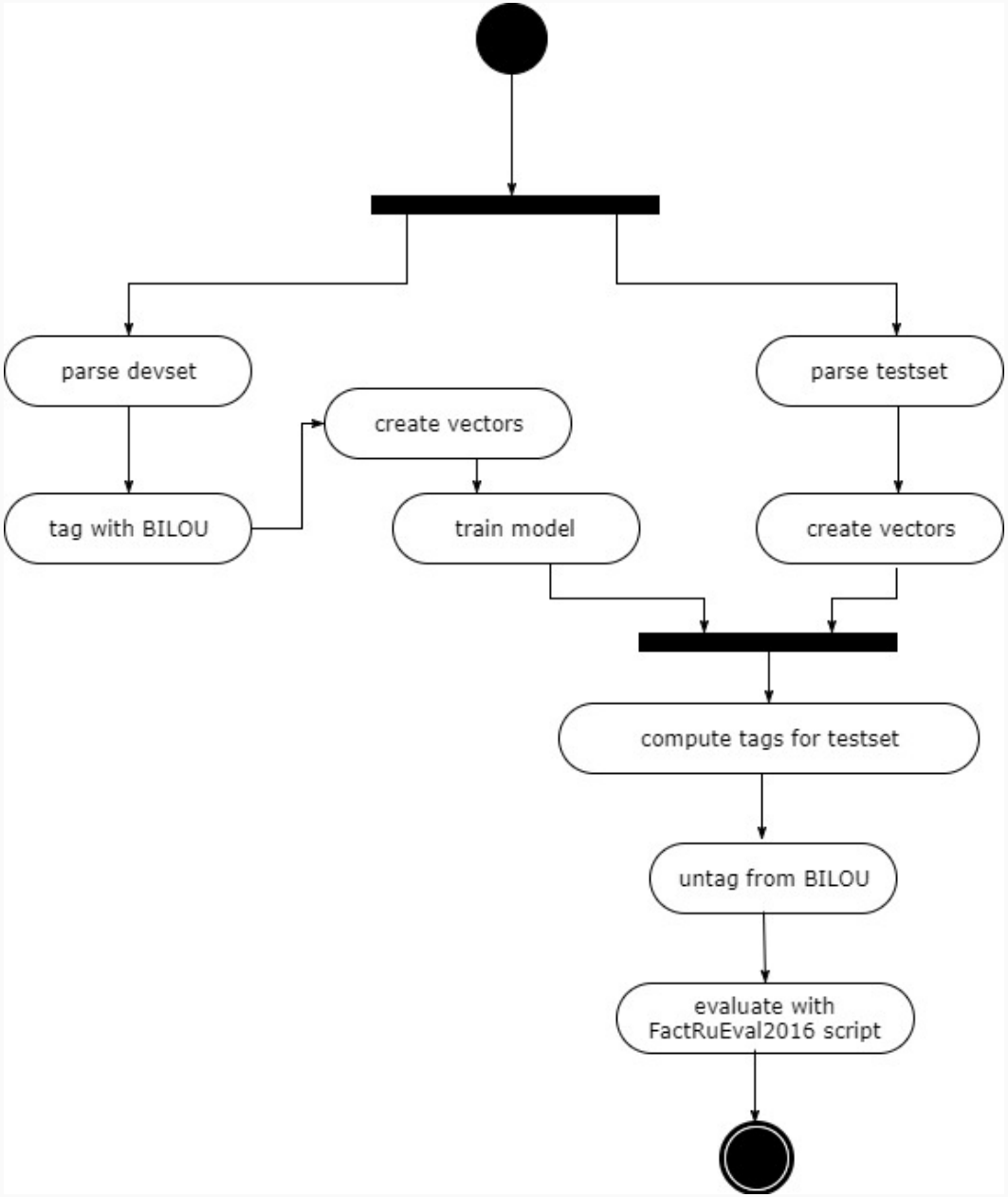
- CNN outperforms biLSTM and SVM in Named Entity recognition task.

Project goal

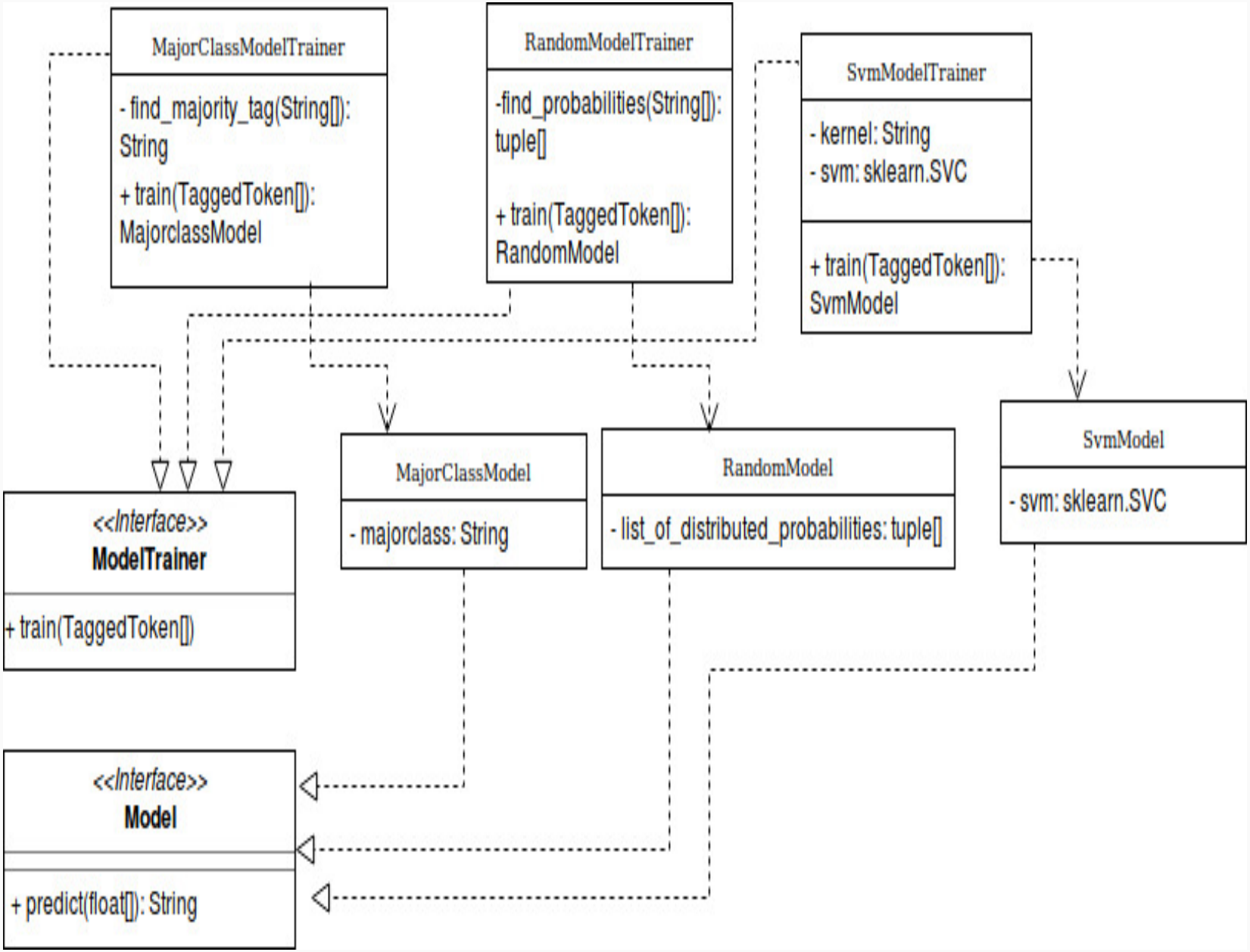
Develop an CNN algorithm using Tensorflow of the automatic NE recognition for the Russian language on FactRuEval2016 data

Implementation

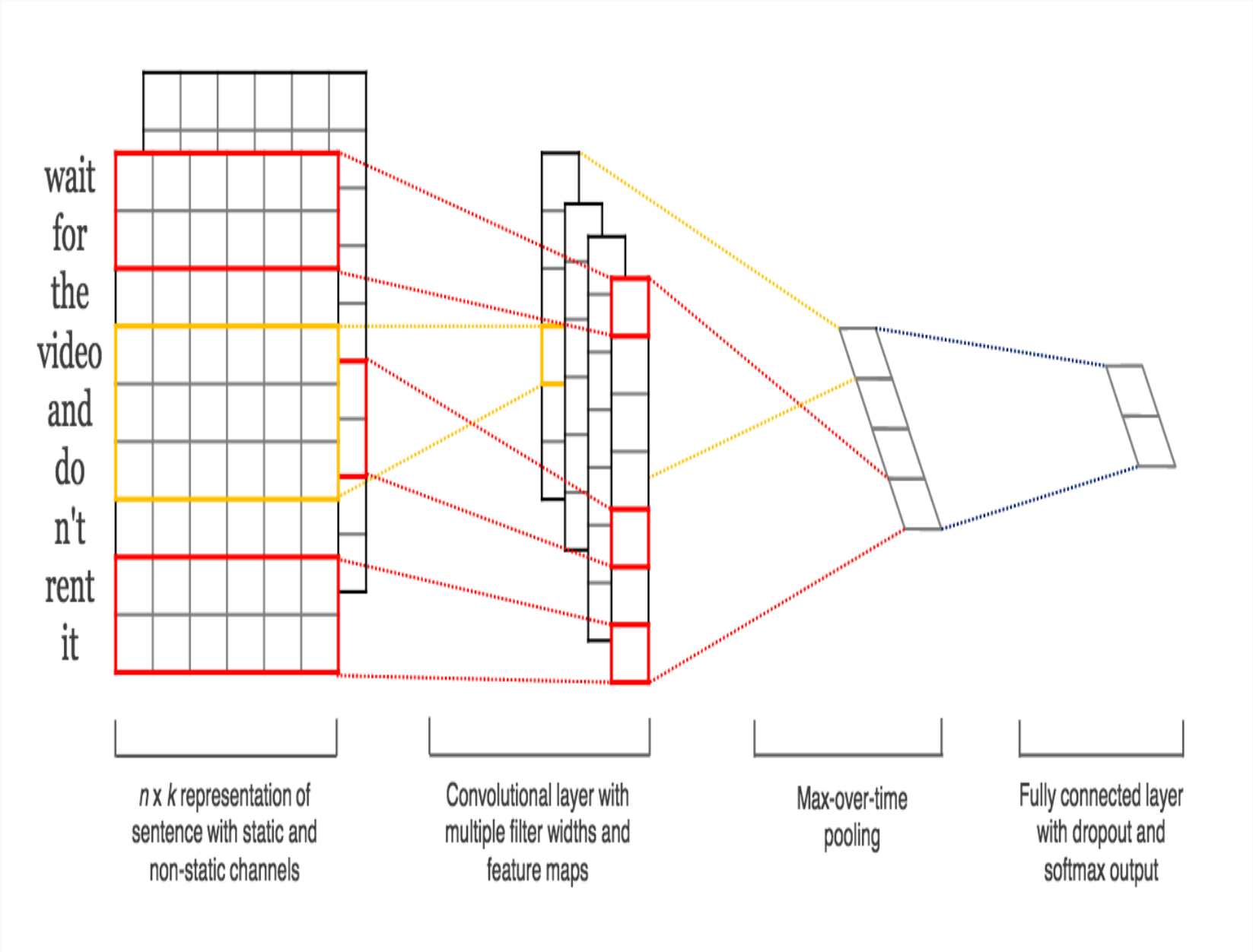
Pipeline


parse devset
create vectors
parse testset
tag with BILOU
train model
create vectors
compute tags for testset
untag from BILOU
evaluate with
FactRuEval2016 script

Model Trainer UML Diagram



Convolutional Neural Network



Convolutional Neural Network

- 100 epochs
- batch_size = 8
- max_len for each batch
- fasttext_size = 100
- output_size = 13
- hidden_size = 512
- conv1
  - filter_size=5
  - num_filters = fasttext_size
  - padding = SAME
- conv2
  - filter_sizes=[3,4,5]
  - num_filters = hidden_size
  - (activation_function = relu) ?
  - padding = SAME
  - max_pool_ksize=[1, filter_size, 1, 1]
- dense layer (activation_fn = tanh)

Evaluation Results

CNN for NERC

type	*precision*	*recall*	*F1*
Person	0.4272	0.6146	0.5040
Location	0.5257	0.2804	0.3657
Organization	0.3933	0.6057	0.4769
Overall	0.4487	0.5121	0.4783

SVM for NERC

type	*precision*	*recall*	*F1*
Person	0.7280	0.7493	0.7385
Location	0.6824	0.7997	0.7346
Organization	0.6231	0.5163	0.5647
Overall	0.6789	0.6756	0.6772

biLSTM for NERC

type	*precision*	*recall*	*F1*
Person	0.7843	0.7215	0.7516
Location	0.6941	0.8268	0.7547
Organization	0.6494	0.6098	0.6289
Overall	0.7050	0.7104	0.7077

Comparison

type	*precision*	*recall*	*F1*
CNN	0.4487	0.5121	0.4783
SVM	0.6789	0.6756	0.6772
BiLSTM	0.7050	0.7104	0.7077

Conclusion

Errors

Марин Ле Пен стала кандидатом на пост главы Франции
Дочь основателя Национального фронта имеет все шансы на победу в выборах.
О французском национальном самосознании читайте в статье Частного корреспондента «Французы по носу и по паспорту»

- Loc = Org
- quotes
- entity overlaping
  - памятник Пушкину -> Loc
  - Пушкину -> Person
- CNN problems

Summary

- BiLSTM outperform CNN
- Feature problems?
- Architecture problems (layers, learning rate, optimizer)?
- > 3 hours of training
- Use LSTM+CRF

Thank you for your attention!

Questions