pioNER: Datasets and Baseline for Armenian Named Entity Recognition

speaker: Tsolak Ghukasyan[1] tsggukasyan@ispras.ru
Garnik Davtyan[2] garnik.davtyan@ispras.ru
Karen Avetisyan[3] karavet@ispras.ru
Ivan Andrianov[4] ivan.andrianov@ispras.ru

Ivannikov Laboratory for System Programming at Russian-Armenian University[1,2,3], Yerevan
Ivannikov Institute for System Programming of the Russian Academy of Sciences[4], Moscow

Motivation

- Create named entity corpora for the Armenian language
- Establish baseline results for named entity recognition

Main contributions

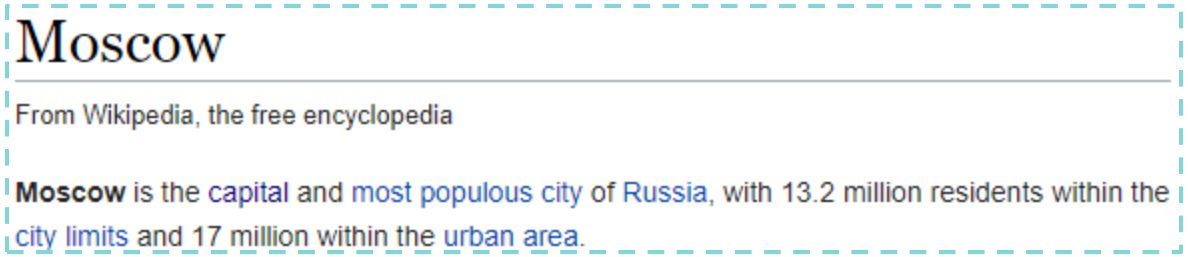
- Silver-standard named entity corpus
- Gold-standard named entity corpus
- GloVe word embeddings
- Baseline results

Plan

- Datasets
  - Automatic train set generation
  - Annotation of test set
- Word embeddings training
- Evaluation of NER models

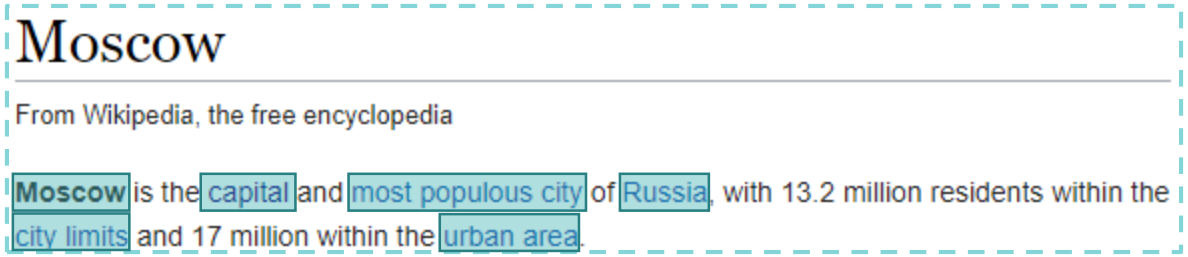
Train set generation

- Automatic generation using Wikipedia (Nothman et al. 2013, Sysoev and Andrianov 2016)
- Steps: *Extract article fragments with outgoing links*



Train set generation

- Automatic generation using Wikipedia (Nothman et al. 2013, Sysoev and Andrianov 2016)
- Steps: *Select outgoing links in the article fragment*



Train set generation

- Automatic generation using Wikipedia (Nothman et al. 2013, Sysoev and Andrianov 2016)
- Steps: *Retrieve the links' target articles*

Train set generation

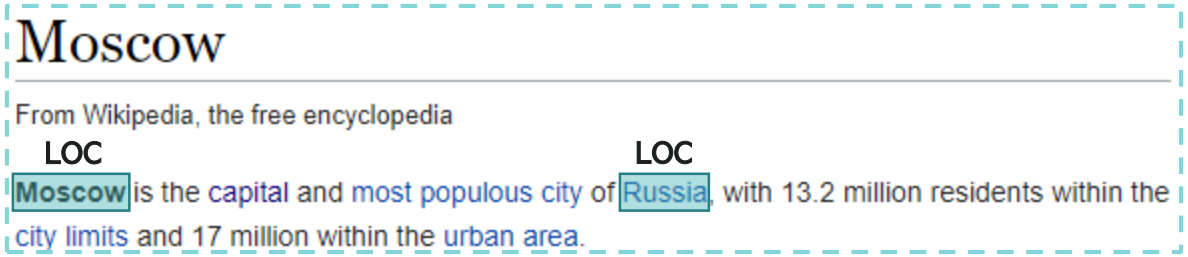
- Automatic generation using Wikipedia (Nothman et al. 2013, Sysoev and Andrianov 2016)
- Steps: *Extract articles' "instance of", "subclass of" categories from their Wikidata element*

Train set generation

- Automatic generation using Wikipedia (Nothman et al. 2013, Sysoev and Andrianov 2016)
- Steps: *Classify articles into named entity types using "instance of", "subclass of" values*

Train set generation

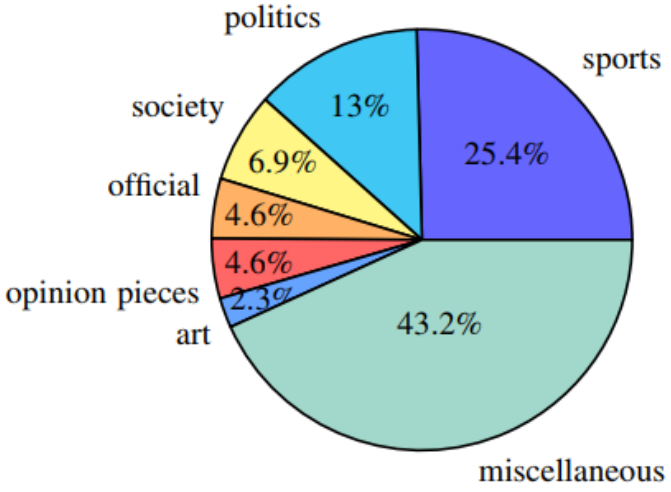
- Automatic generation using Wikipedia (Nothman et al. 2013, Sysoev and Andrianov 2016)
- Steps: *Label links' text according to their target article's type*



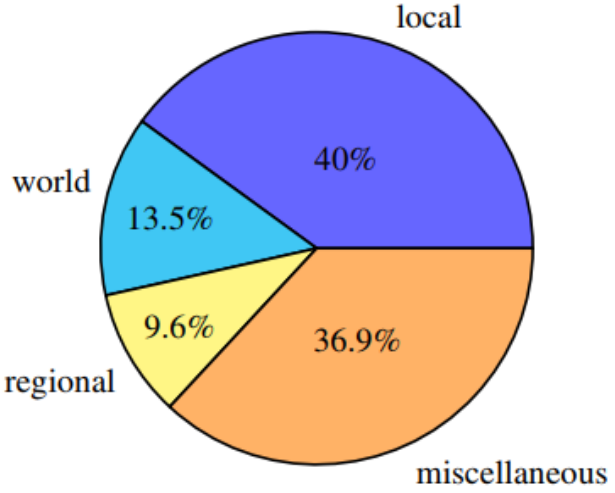
Test set annotation

- Over 250 news articles (sports, politics, art etc)
- Location, organization, person entities
- Annotation tool: BRAT NLP
- BBN Technologies guidelines for TREC 2002 question answering track

Distribution of news articles' topics in the test set



Geographical distribution of news articles' content in the test set



Test set statistics

- Comparable in size with benchmark datasets for other languages

Table 1. Comparison of Armenian, English, German, Spanish and Russian test sets: sentence, token, and named entity counts

Test set	Tokens	LOC	ORG	PER
Armenian	53453	1306	1337	1274
English CoNLL03	46435	1668	1661	1617
German CoNLL03	51943	1035	773	1195
Spanish CoNLL02	51533	1084	1400	735
Russian factRuEval-2016	59382	1239	1595	1353

Word Embeddings

- Data
  - 79 million tokens
  - Source: Armenian Wikipedia, The Armenian Soviet Encyclopedia, a subcorpus of Eastern Armenian National Corpus, news articles, blog posts
- Model
  - GloVe
  - Dimensions: 50, 100, 200, 300
  - Vocabulary size: 400 000

Baselines

- Stanford NER
- SpaCy 2.0 NER
- biLSTM+CRF over GloVe+charBiLSTM features

Evaluation

- Generated data randomly split into 80% as train, 20% as development set
- Manually annotated data as test set
- IOB tagging scheme

Table 2. Evaluation results for named entity recognition algorithms

Algorithm	dev			test		
	Precision	Recall	F1	Precision	Recall	F1
Stanford NER	76.86	70.62	73.61	**78.46**	46.52	58.41
spaCy 2.0	68.19	71.86	69.98	64.83	**55.77**	59.96
Char-biLSTM+biLSTM+CRF	**77.21**	**74.81**	**75.99**	73.27	54.14	**62.23**

Error analysis

On development set:

- Errors caused by incompleteness of generated annotations
- Low precision for organization entities
- Error rate higher on descriptor words

Table 3. Confusion matrix of charBiLSTM+biLSTM+CRF on the development set

		Predicted						
		O	B-PER	B-ORG	B-LOC	I-ORG	I-PER	I-LOC
Actual	O	26707	100	57	249	150	78	129
	B-PER	107	712	6	32	2	4	0
	B-ORG	93	6	259	58	8	0	0
	B-LOC	226	25	32	1535	5	3	20
	I-ORG	67	1	5	3	289	3	19
	I-PER	46	5	0	1	6	660	8
	I-LOC	145	0	1	13	45	11	597
	Precision (%)	97.5	83.86	71.94	81.17	57.23	86.95	77.23

Future work

- Improved approaches for dataset generation (e.g. WiNER)
- Richer annotation
  - more entity types (e.g. Event)
  - more fine-grained types (e.g. City, Country, Region instead of Location)

Thank you!