The First Cross-Lingual Challenge on Recognition, Normalization, and Matching of Named Entities in Slavic Languages

Jakub Piskorski, Lidia Pivovarova, Jan Šnajder,
Josef Steinberger, Roman Yangarber

European Commission

UNIVERSITY OF HELSINKI

TakeLab

UNIVERSITY OF WEST BOHEMIA

Outline

- Introduction
- Tasks
- Trial and Test Datasets
- Baseline System: LexiFlexi
- Evaluation Methodology
- Evaluation Results
- Way Forward

Introduction

- **Motivation**
  - foster research on NER, NE lemmatization and their cross-language matching for **Slavic languages**
  - foster development of **“all-rounder”** NER and cross-lingual entity matching solutions not tailored to specific, narrow domains
- **Task**
  - Input: collection of web documents in seven Slavic languages revolving around a certain “focus” entity
  - Output: extract mentions of general-type NEs, compute their base forms and assign them cross-language IDs

Tasks: NE Mention Detection and Classification

- **ORG** (ex. *Citi Handlowy w Poznaniu* - PL)
- **PER** (ex. *Władimir Putin* - PL, *Ukrajinci* - SI)
- **LOC** toponyms, GPEs, facilities, (ex. *Rusko* - CS, *Európska únia* - SK, *Zagrebački Glavni kolodvor* - HR)
- **MISC** (ex. *Motorola Moto X* - PL, *Święta Bożego Narodzenia* - PL)
- no extraction of positional information
- recognition of timex, numex and identifiers and nested NEs **not part of the task**

Tasks: Name Normalization

	Genitive	Nominative (“base”)
hr	*Europske komisije*	*Europska komisija*
cz	*Evropské komise*	*Evropská komise*
pl	*Komisji Europejskiej*	*Komisja Europejska*
ru	Европейской комиссией	Европейская комиссия
sl	*Evropske komisije*	*Evropska komisija*
sk	*Európskej komisie*	*Európska komisia*
ua	Європейської Комісії	Європейська Комісія

Tasks: Entity Matching

	Mention	ID
pl	*Komisja Europejska*	1
pl	*Komisją Europejską*	1
pl	*KE*	1
pl	*Kom Europ*	1
ru	Европейской комиссией	1
sl	*Evropske komisije*	1
sl	*EK*	1

Trial and Test Datasets

- Trial Datasets:
  - **187 docs** related to Beata Szydło, the current prime minister of Poland,
  - **186 docs** related to ISIS
- Test Datasets:
  - **177 docs** related to Donald Trump,
  - **203 docs** related to European Commission
- Languages: Czech, Croatian, Polish, Slovak, Slovenian, Russian and Ukrainian

Test Datasets

	Trump		ECommission	
Language	# docs	# ment	# docs	# ment
Croatian	25	525	25	436
Czech	25	479	25	417
Polish	25	692	24	466
Russian	26	331	24	385
Slovak	24	453	25	374
Slovene	24	474	26	434
Ukrainian	28	337	54	1078
Total	177	3291	203	3588

Table: Quantitative data about the test datasets.

Test Datasets

Entity type	TRUMP	ECOMMISSION
PER	48.4%	11.9%
LOC	26.9%	29.1%
ORG	18.3%	48.4%
MISC	6.4%	9.6%

Table: Breakdown of the annotations according to the entity type.

Inflected forms: in TRUMP dataset min 37.5% (Slovak) and max 57.5% (Croatian)

Test Datasets: Preparation

- pose a search query to Google in each of the target languages
- extract max. 100 links and remove duplicates
- download documents, parse HTML and convert to pure text
- remove documents with obvious HTML parser failure
- select for each language/topic circa 25 documents for annotation (1 person per language)
- 2 persons aligned the cross-language IDs

Baseline system: Lexi Flexi

Basic idea: exploit existing lexico-semantic resources available

1. match names from JRC-Names database (4,05 mln entries) + exploit the cross-lingual entity IDs,
2. match names from a collection of **multi-word named entities** semi-automatically derrived from BABELNET (6,82 mln entries) in unconsumed text,
3. match toponyms from the GEONAMES gazetteer (1,36 mln) in unconsumed part of the texts + exploit cross-lingual IDs,
4. apply language-independent heuristics to match variants of mentions of entities recognised in the previous steps

Evaluation Methodology

Three aspects evaluated:

1. entity recognition
   - relaxed evaluation, partial match
   - relaxed evaluation, exact match
   - strict evaluation
2. entity normalization
   - considers only normalized mentions
3. entity matching
   - the LEA metric (Link-based Entity Aware evaluation)
   - per article / language / across languages

- per NE type / language / topic
- P/R/F figures

LEA for Named Entity Matching

Moosavi, Nafise Sadat and Strube, Michael: Which Coreference Evaluation Metric Do You Trust? A Proposal for a Link-based Entity Aware Metric. ACL 2016.

$$Recall_{LEA} = \frac{\sum_{k_i \in K}(imp(k_i) \times res(k_i))}{\sum_{k_z \in K} imp(k_z)}$$

$$Precision_{LEA} = \frac{\sum_{r_i \in R}(imp(r_i) \times res(r_i))}{\sum_{r_z \in R} imp(r_z)}$$

- alternative measures in coreference resolution: B^3, CEAF, and BLANC

Participant Systems

JHU

- all languages, NER and Entity Matching subtasks
- statistical tagger SVMLattice,
- NER labels inferred by projecting English tags across bitext,
- the Illinois tagger for English
- a rule-based entity clusterer kripke for Entity Matching

Liner2 (pw)

- Polish only, NER and normalization subtasks
- a generic framework for resolving tasks based on sequence labeling

Evaluation Results

TRUMP		*Language*													
Phase	*Metric*	**cs**		**hr**		**pl**		**ru**		**sk**		**sl**		**ua**	
Recognition	**Relaxed Partial**	lf	47.6	jhu	52.4	pw	66.7	lf	63.6	jhu	46.8	jhu	47.3	lf	54.0
		jhu	46.2	lf	37.0	lf	51.0	jhu	46.3	lf	46.8	lf	46.3	jhu	38.8
						jhu	44.8								
	Relaxed Exact	lf	46.6	jhu	50.8	pw	66.1	lf	62.6	jhu	46.2	jhu	46.0	lf	53.3
		jhu	46.1	lf	35.6	lf	48.8	jhu	43.1	lf	45.2	lf	44.2	jhu	37.3
						jhu	43.4								
	Strict	jhu	46.1	jhu	50.4	pw	66.6	lf	55.6	jhu	47.0	jhu	46.2	lf	50.8
		lf	42.2	lf	37.4	lf	48.0	jhu	41.8	lf	44.8	lf	44.2	jhu	33.2
						jhu	41.0								
Normalization						pw	60.5								
Entity matching	**Document-level**	lf	16.0	lf	31.0	lf	30.0	lf	25.8	lf	26.4	lf	30.1	lf	35.1
		jhu	5.4	jhu	7.3	jhu	6.3	jhu	11.2	jhu	10.1	jhu	9.5	jhu	0.6
	Single-language	jhu	19.3	lf	17.8	lf	24.0	lf	41.7	jhu	22.6	lf	29.4	lf	30.2
		lf	19.0	jhu	17.6	jhu	18.2	jhu	18.9	lf	21.4	jhu	28.7	jhu	10.7
	Cross-lingual	lf	14.3												
		jhu	13.7												

Table: Evaluation results for the Trump corpus.

Evaluation Results

ECommission		*Language*													
Phase	*Metric*	**cs**		**hr**		**pl**		**ru**		**sk**		**sl**		**ua**	
Recognition	**Relaxed Partial**	lf	51.0	jhu	45.9	pw	61.8	lf	62.8	lf	50.3	jhu	47.9	lf	28.4
		jhu	47.6	lf	37.8	jhu	47.3	jhu	46.0	jhu	49.1	lf	43.8	jhu	18.4
						lf	42.8								
	Relaxed Exact	lf	50.0	jhu	43.1	pw	60.9	lf	60.7	lf	49.3	jhu	43.9	lf	28.4
		jhu	44.4	lf	37.2	jhu	42.4	jhu	44.1	jhu	46.4	lf	39.3	jhu	14.7
						lf	41.5								
	Strict	jhu	47.2	jhu	46.2	pw	61.1	lf	53.7	jhu	46.1	jhu	47.8	lf	20.8
		lf	41.2	hr	30.0	jhu	44.8	jhu	46.5	lf	42.5	lf	37.5	jhu	10.8
						lf	34.6								
Normalization						pw	48.3								
Entity Matching	**Document-level**	lf	25.0	jhu	16.0	jhu	13.7	lf	22.7	jhu	13.1	jhu	36.8	lf	1.6
		jhu	3.0	lf	6.7	lf	6.7	jhu	13.7	lf	12.7	lf	25.4	jhu	0.6
	Single-language	jhu	27.3	jhu	22.1	jhu	17.5	lf	45.8	jhu	30.6	jhu	32.2	lf	11.4
		lf	18.0	lf	12.8	lf	13.0	jhu	24.9	lf	23.9	lf	15.2	jhu	4.8
	Cross-lingual	lf	12.0												
		jhu	5.3												

Table: Evaluation results for the European Commission corpus.

Way Forward

- provision of additional test datasets (of similar nature)
- extend the set of the languages covered (inclusion of Baltic languages?)
- refining the NE annotation guidelines
- making the evaluation software publicly available