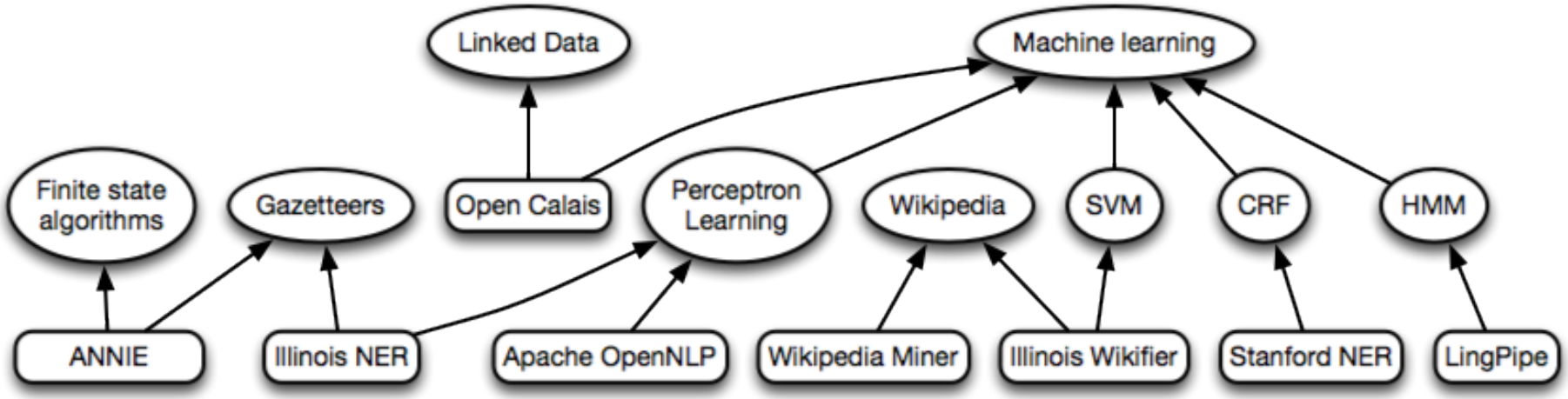
MSM 2013 Challenge: Annotowatch

Stefan Dlugolinsky, Peter Krammer, Marek Ciglan, Michal Laclavik

Institute of Informatics,
Slovak Academy of Sciences

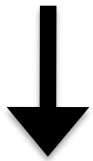
Approach

- Not to create a new NER method
- Combine various existing NER taggers, which are based on diverse methods through ML
- Candidate NER taggers and their methods:



Candidate NER taggers evaluation results

- worse performance on micropost text than on news texts for which the taggers were intended to be used, but:
  - high recall when unified (low precision drawback)
  - together can discover different named entities



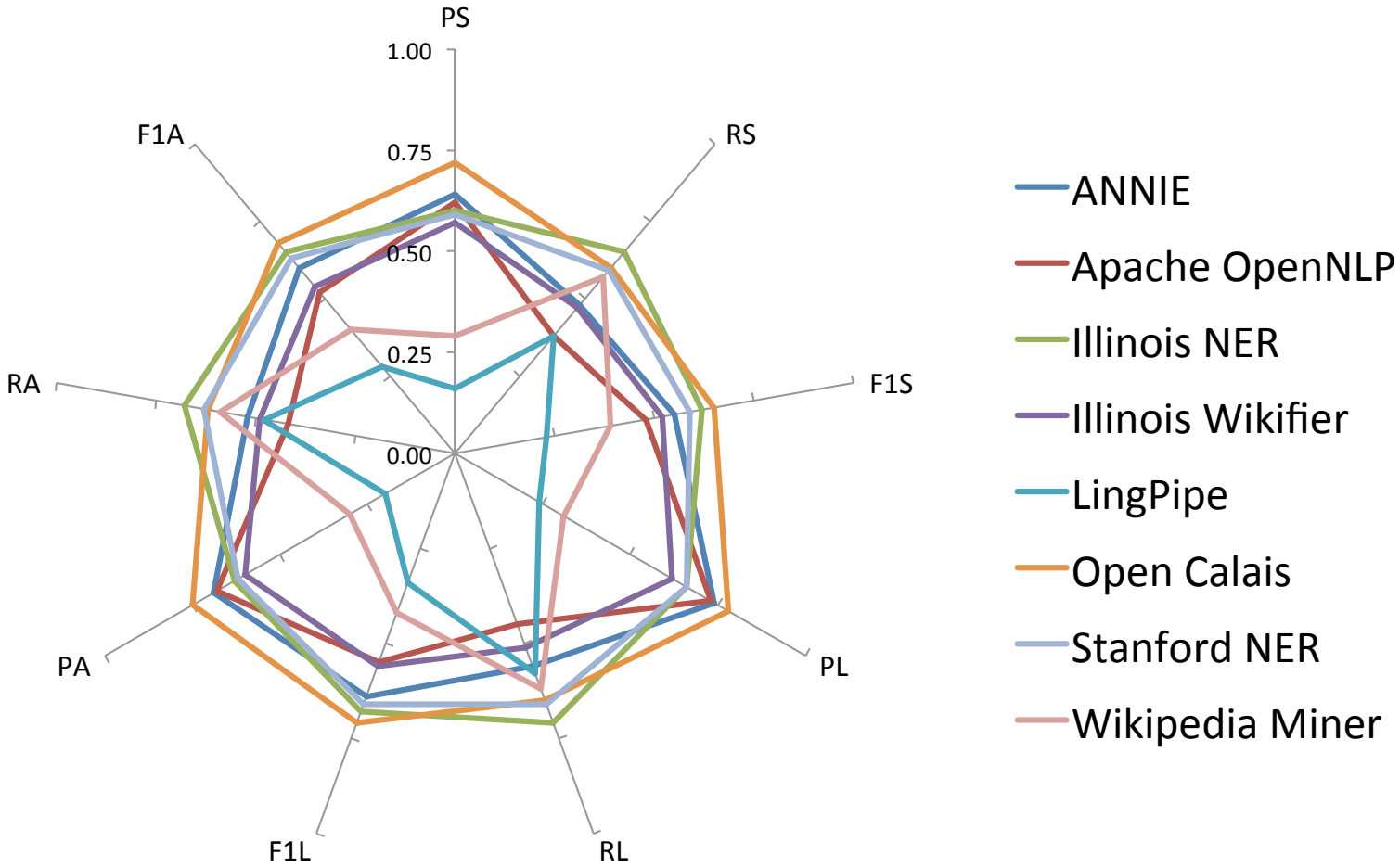
We felt that there can be a superior performance achieved by combining the taggers

Evaluation details 1/4

- evaluated on a modified MSM 2013 Challenge training dataset
- modifications made to original training set:
  - Removed duplicate and overlapping microposts
  - Removed country/nation adjectivals and demonyms (e.g. MISC/English)
- modified dataset of 2752 unique microposts
- no tweaking or configuration made to evaluated taggers prior to evaluation

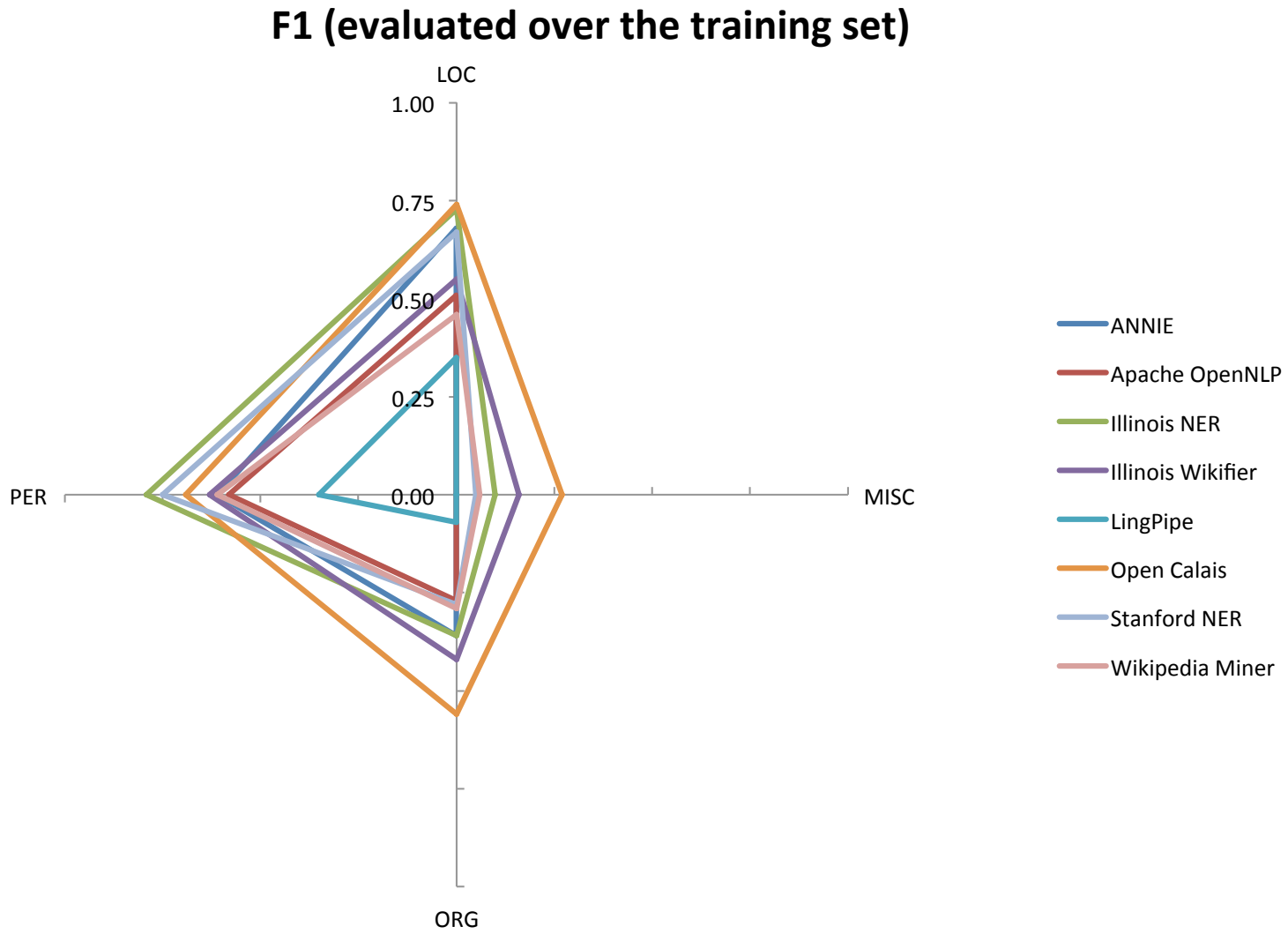
Evaluation details 2/4

Micro summary – P, R, F_1 (evaluated over the training set)



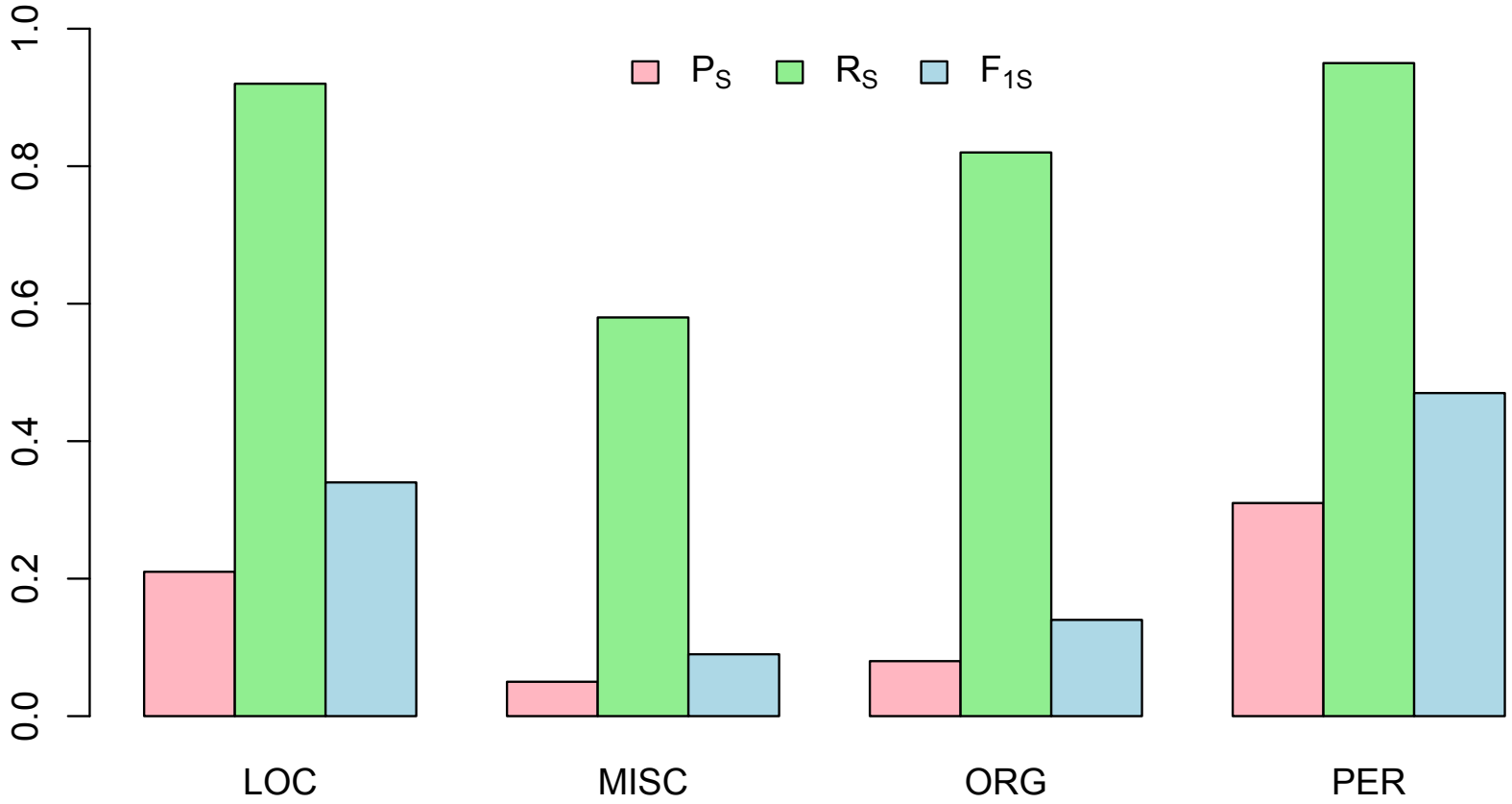
S – strict comparison (exact offests), L – lenient comparison (overlap), A – average of S&L

Evaluation details 3/4



Evaluation details 4/4

Unified NER taggers' results (evaluated over the training set)



Chosen NER taggers

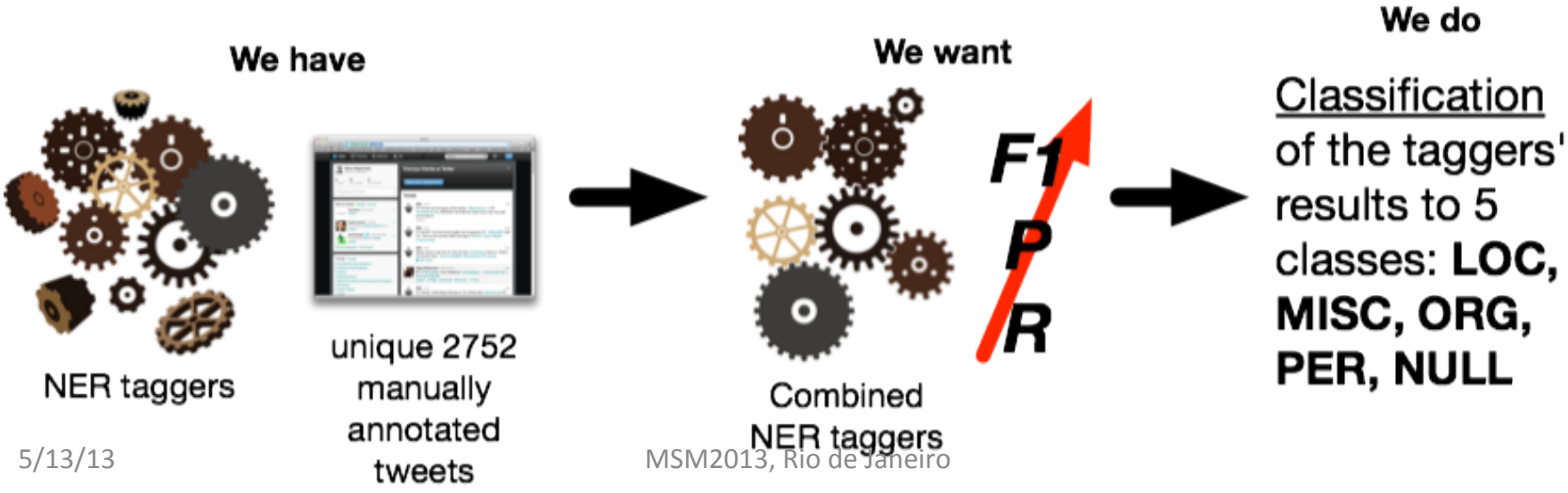
- ANNIE
- Apache OpenNLP
- Illinois NER
- Illinois Wikifier
- OpenCalais
- Stanford NER
- Wikipedia Miner

- Not specially configured, trained or tweaked for Microposts.
- Default settings and the most suitable official models were used.
- mapping to target entities was the only "hack" to these tools

- We have created one special tagger "Miscinator" for MISC extraction of entertainment award and sport events. It was based on a gazetteer created from MISC entities found in the training set and enhanced by Google Sets.

How to combine the taggers

- Not simply take the best tagger for each NE class, i.e. extract LOC, MISC, ORG by OpenCalais and PER by Illinois NER (we called it “Dummy model”)
- Transform to Machine Learning task



Features for ML 1/4

- Micropost features:
  - describe micropost text globally (as a whole)
  - considered only words with length > 3

    - **awc** – all words capital
    - **awuc** – all words upper case
    - **awlc** – all words lower case

Features for ML 2/4

- Annotation features
  - Annotations of underlaying NER taggers
  - Considered annotation types: LOC, MISC, ORG, PER, NP, VP, OTHER
  - Describe each annotation found by underlaying NER tagger (reference/ML instance annotation)

    - **ne** - annotation class
    - **flc** - first letter capital
    - **aluc** - all letters upper cased
    - **allc** - all letters lower cased
    - **cw** - capitalized words
    - **wc** - word count

Features for ML 3/4

- Overlap features
  - Describe, how the reference annotation overlaps with other annotations

    - **ail** – average intersection length of the reference annotation with other annotations of the same type*
    - **aiia** – how much the reference annotation covers the other annotations of the same type* in average
    - **aiir** – how much % of the reference annotation length is covered by other annotations of the same type* in average

    * reference annotation can be of different type

Features for ML 4/4

- Confidence features
  - Some taggers return their confidence about the annotation (OpenCalais, Illinois NER)
    - **E(p)** - mean value of the confidence values for overlapping annotations
    - **var(p)** - variance of the confidence values for overlapping annotations
- Answer – NE class of manual annotation which overlaps the instance/reference annotation

Model training

Tried algorithms to train a classification model:

- ***C4.5***
- *LMT (Logistic Model Trees)*
- *NBTree (Bayess Network Tree)*
- *REPTree (Fast decision tree learner)*
- *SimpleCart LADTree (LogitBoost Alternating Decision Tree)*
- ***Random Forest***
- *AdaBoostM1*
- *MultiLayer Perceptron Neural Network*
- *Bayes Network*
- *Bagging Tree*
- *FT (Functional trees)*

Input data:

- ~36,000 instances,
- 200 attributes

Input data preprocessing:

- removed duplicate instances
- removed attributes where values changed for insignificant number of instances

Preprocessed input data, ready for training:

- ~31,000 instances,
- 100 attributes

Validation:

- 10-fold cross validation,
- holdout

Evaluation over the MSM test dataset

- Annotowatch 1, 2 and 3 are our submissions to the challenge (v3 used special post-processing)
- RandomForest 21 and C4.5 M13 are new models trained after the submission deadline
- Dummy model is our baseline (simply built of the best taggers for particular NE class)
- Test dataset was cleared from duplicates and there were some corrections made (results may vary from the official challenge results)
- Comparison of the response and gold standard was strict

Tagger	LOC			MISC			ORG			PER			Macro			Micro		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
RandomForest 21	0.51	0.61	0.56	0.39	0.19	0.26	0.50	0.47	0.48	0.86	0.88	0.87	0.56	0.54	0.54	0.77	0.76	0.76
C4.5 M13	0.53	0.61	0.57	0.59	0.25	0.35	0.41	0.33	0.36	0.87	0.87	0.87	0.60	0.51	0.54	0.78	0.73	0.75
Annotowatch 2	0.39	0.54	0.46	0.38	0.25	0.30	0.39	0.40	0.40	0.85	0.85	0.85	0.50	0.51	0.50	0.72	0.72	0.72
Annotowatch 1	0.44	0.58	0.50	0.39	0.26	0.31	0.39	0.40	0.39	0.83	0.84	0.83	0.51	0.52	0.51	0.71	0.72	0.71
Annotowatch 3	0.44	0.58	0.50	0.39	0.25	0.31	0.37	0.45	0.41	0.83	0.84	0.83	0.51	0.53	0.51	0.70	0.72	0.71
Illinois NER	0.46	0.57	0.51	0.05	0.08	0.07	0.26	0.36	0.30	0.86	0.82	0.84	0.41	0.46	0.43	0.64	0.69	0.66
Stanford NER	0.46	0.60	0.52	0.01	0.01	0.01	0.25	0.31	0.28	0.83	0.80	0.82	0.39	0.43	0.41	0.65	0.66	0.66
Open Calais	0.75	0.52	0.61	0.54	0.20	0.29	0.62	0.19	0.30	0.66	0.73	0.69	0.64	0.41	0.47	0.66	0.60	0.63
Dummy	0.29	0.72	0.41	0.09	0.35	0.15	0.32	0.62	0.42	0.63	0.92	0.75	0.33	0.65	0.43	0.47	0.82	0.60
ANNIE	0.47	0.49	0.48	-	-	-	0.23	0.17	0.19	0.72	0.64	0.68	0.61	0.32	0.34	0.63	0.52	0.57
Illinois Wikifier	0.28	0.44	0.34	0.07	0.13	0.09	0.53	0.41	0.46	0.88	0.55	0.67	0.44	0.38	0.39	0.63	0.49	0.55
Apache OpenNLP	0.36	0.41	0.38	-	-	-	0.14	0.12	0.13	0.78	0.54	0.64	0.57	0.27	0.29	0.62	0.43	0.51
Wikipedia Miner	0.25	0.50	0.33	0.03	0.18	0.05	0.29	0.38	0.33	0.76	0.58	0.66	0.33	0.41	0.34	0.41	0.52	0.46
LingPipe	0.09	0.45	0.15	-	-	-	0.03	0.22	0.05	0.34	0.44	0.38	0.37	0.28	0.14	0.15	0.38	0.21

Thank you!

Questions?

Annotowatch 3 post-processing

- If a micropost identical to one in the training set was annotated, we extended the detected concepts by those from manually annotated training data (affecting three microposts)
- A gazetteer built from a list of organizations found in the training set has been used to extend the ORG annotations of the model (affecting 69 microposts).