HORUS-NER: A Multimodal Named Entity Recognition Framework for Noisy Text

Diego Esteves
Principal Data Scientist@Farfetch, Portugal
Research Associate@SDA Research, Germany

Data Science Portugal (DSPT#69)
12/12/2019 - Talkdesk

DSPT
DATA SCIENCE PORTUGAL

Named Entity Recognition for Noisy Data

RQ	Can images along with news improve the performance of the named entity recognition models on noisy text?

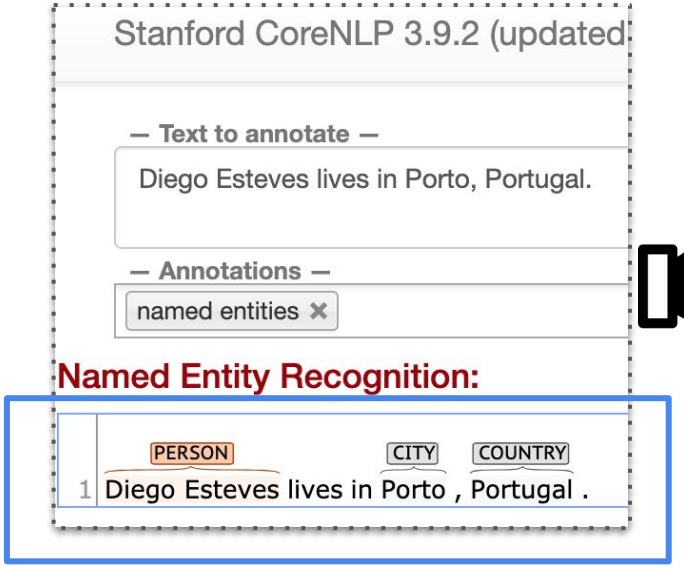
Named-entity recognition (NER) is a subtask of information extraction aiming to locate **named entities** in natural language documents:

S = Diego Esteves lives in Porto, Portugal.

Named Entity Recognition for Noisy Data

Newswire

- Lexical, Shape and Orthographic features
- Gazetteers
- SOTA high performance in formal domains (easily 0.90 F1)



Named Entity Recognition for Noisy Data

Microblogs

- [Derczynski, 2015]
  - Shortness of microblogs, harder to understand
  - Less gramatical
  - More language variations
  - Spellings mistakes
  - Emojis
  - Hashtag
  - Abbreviation
  - Ambiguity

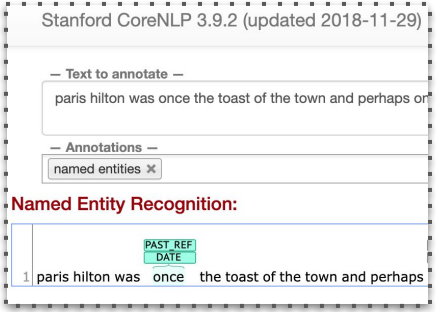
Named Entity Recognition for Noisy Data

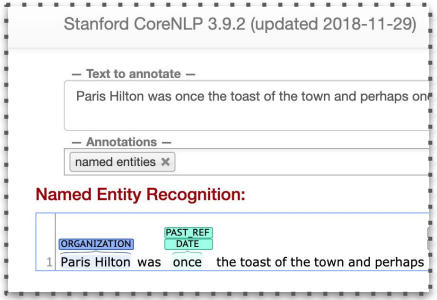
Dealing with ~~real life~~ noisy

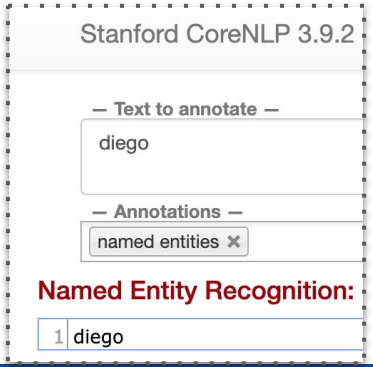
S = Paris Hilton was once the toast of the town and perhaps one of Hollywood's most famous socialites.

What if small variations are applied?
e.g., *S* = paris hilton ?

What about non-english names?
e.g., S = diego ?







Named Entity Recognition for Noisy Data

Dealing with ~~real life~~ noisy

- Look-up strategies and standard local features struggle on noisy data
- **F1 0.20 and 0.60**
  [Ritter et al. 2011]
  [Derczynski et al., 2015]
  [Esteves et al., 2017]
  [Qi Zhang et al., 2018]

'2m', '2ma', '2mar', '2mara', '2maro', '2marrow', '2mor', '2mora', '2moro', '2morow', '2morr', '2morro', '2morrow', '2moz', '2mr', '2mro', '2mrrw', '2mrw', '2mw', 'tmmrw', 'tmo', 'tmoro', 'tmorrow', 'tmoz', 'tmr', 'tmro', 'tmrow', 'tmrrow', 'tmrrw', 'tmrw', 'tmrww', 'tmw', 'tomaro', 'tomarow', 'tomarro', 'tomarrow', 'tomm', 'tommarow', 'tommarrow', 'tommoro', 'tommorow', 'tommorrow', 'tommorw', 'tommrow', 'tomo', 'tomolo', 'tomoro', 'tomorow', 'tomorro', 'tomorrw', 'tomoz', 'tomrw', 'tomz'

- Joint clustering to minimise the gap between world knowledge and KBs
- Basic idea:
  - Correlation between images and entities
  - Correlation between search textual results and entities
- Combination of text and image features with simple decision trees-based models
- Majority voting committee

Named Entity Recognition for Noisy Data

SOTA architectures

- **[Z. Huang et al., 2015]**
  B-LSTM+CRF
- **[Ma, Xuezhe and Hovy, Eduard, 2016]**
  B-LSTM+CNN+CRF
- **[Lample et al., 2016]**
  Char+B-LSTM+CRF
- **[Qi Zhang et al., 2018]**
  +Attention

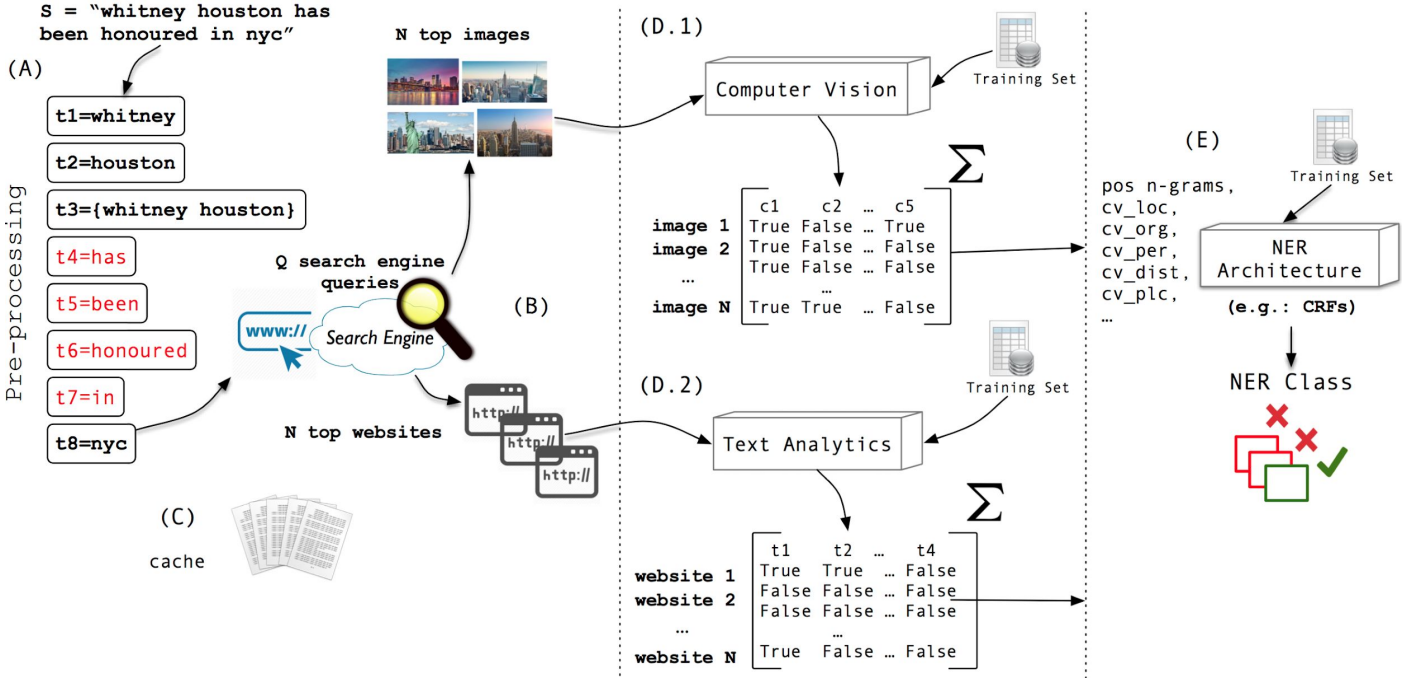
Named Entity Recognition for Noisy Data

SOTA architectures

+bias, -noisy

Named Entity Recognition for Noisy Data

HORUS: The methodology



Named Entity Recognition for Noisy Data

HORUS: The methodology (Computer Vision)

Object detection

SIFT (Scale Invariant Feature Transform): image descriptor extraction

BoF: clustering of feature histograms (k-means)

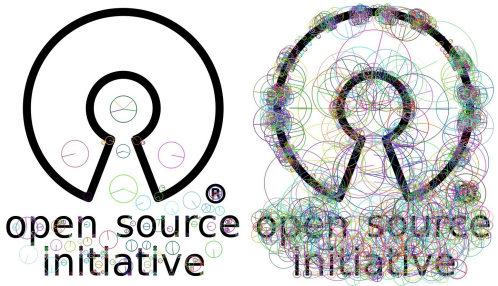
- Image ~ histogram of visual words frequencies
- Some image groups are related to certain named entities

Classifiers: Unsupervised + Supervised learning

Training datasets

- LOC: *Scene 13*
- PER: *Caltech 101 Object Categories*
- ORG: *METU*

NER Images Candidates (number of trained models)	
LOC	Building, Suburb, Street, City, Country, Mountain, Highway, Forest, Coast and Map (10)
ORG	Company Logo (1)
PER	Human Face (1)



Named Entity Recognition for Noisy Data

HORUS: The methodology (Text)

Text Analytics

Features: term frequency-Inverse document frequency (TF-IDF)
Classifier: bag-of-words based
Training dataset: 15K DBpedia instances annotated with PER, ORG and LOC classes

```
SELECT ?location, ?abstract FROM <http://dbpedia.org> WHERE {?location
rdf:type dbo:Location .
        ?location dbo:abstract ?abstract .
FILTER (lang(?abstract) = 'en')} LIMIT 50000
```

HORUS 1.0

Named Entity Recognition for Noisy Data

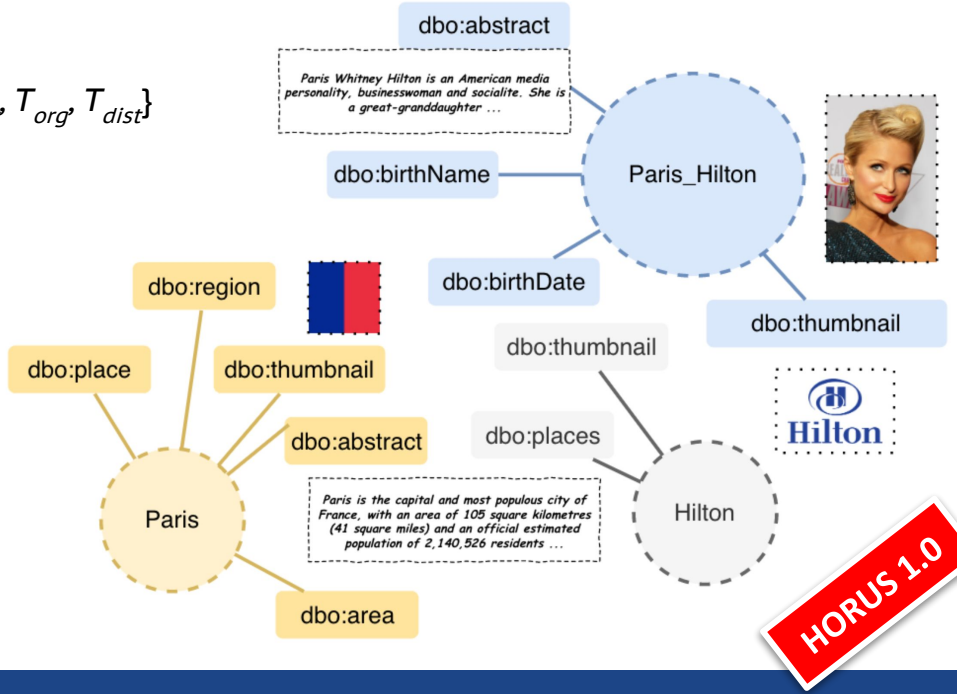
HORUS 1.0: Feature Vector

Heuristic-based DT

$$M_i = \{j, t, ng_{pos}, C_{loc}, C_{per}, C_{org}, C_{dist}, C_{plc}, T_{loc}, T_{per}, T_{org}, T_{dist}\}$$

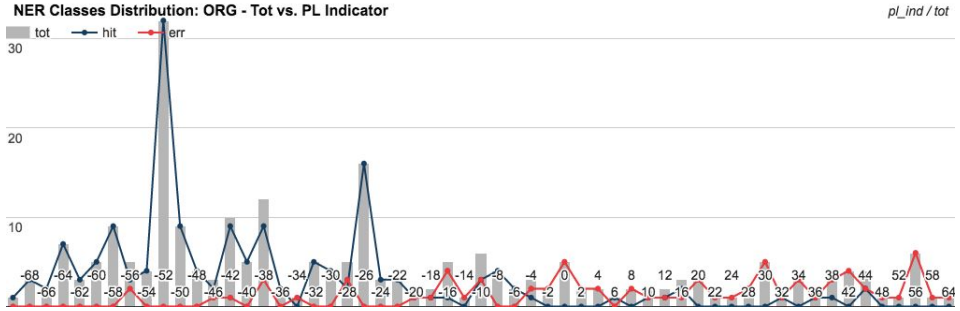
for each sentence *i* and token *t* in position *j*

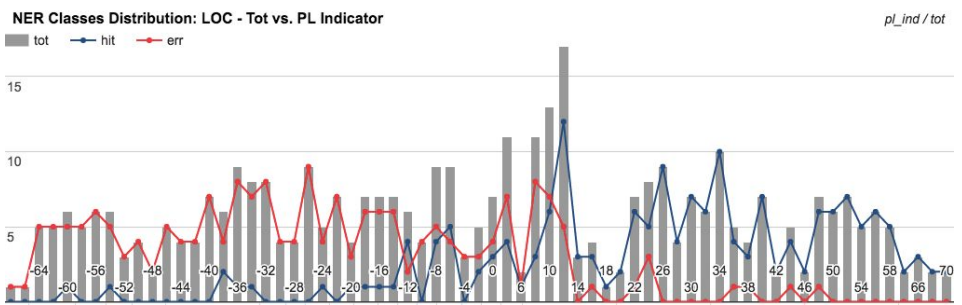
- ng_{pos}= n-gram of POS tag
- C_k,T_k= total objects found by classifier for class *k*
- C_{dist},T_{dist}= distance b/w two top predictions
- C_{plc}= sum of all predictions by all LOC classifiers

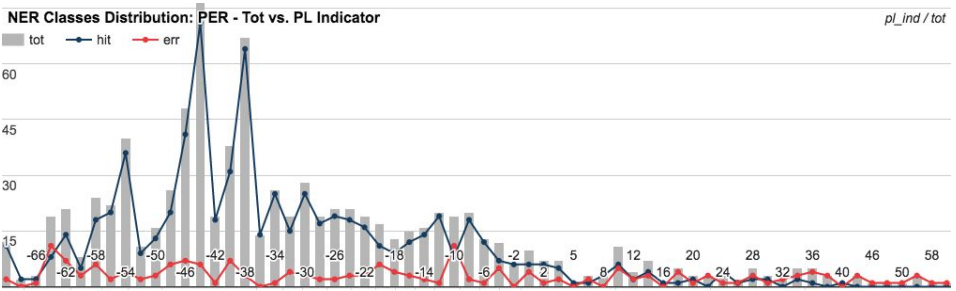


Named Entity Recognition for Noisy Data

Place bias: hits x errors over sample







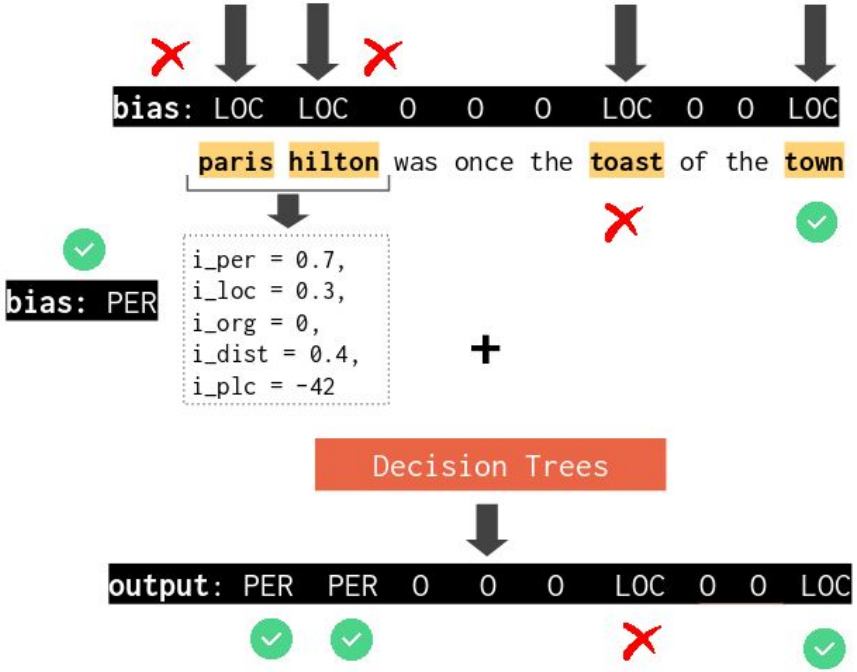
$$\Phi_l^{\mathrm{LOC}}(i) = \begin{cases} 1, & \text{if } \sum_{l=1}^{L} \Phi_l^{\mathrm{LOC}}(i) \geq \theta \\ -1, & \text{otherwise} \end{cases}$$

if $c \neq \mathrm{LOC}$ we get $R^c_{\mathcal{I}_t} = \sum_{i \in \mathcal{I}_t} \sum_{l=1}^{L} \Phi_l^c(i)$

HORUS 1.0

Named Entity Recognition for Noisy Data

HORUS 1.0: Performance over first noisy challenge (Ritter)



NER Class	Precision	Recall	F-measure
Person (PER)	0.86	0.53	0.66
Location (LOC)	0.70	0.40	0.51
Organisation (ORG)	0.90	0.46	0.61
None	0.99	1.0	0.99
Average (PLO)	0.82	0.46	**0.59**

Table 2: Performance measure for our approach in Ritter dataset: 4-fold cross validation

NER System	Description	Precision	Recall	F-measure
Ritter et al., 2011 [19]	LabeledLDA-Freebase	0.73	0.49	0.59
Bontcheva et al., 2013 [3]	Gazetteer/JAPE	0.77	0.83	0.80
Bontcheva et al., 2013 [3]	Stanford-twitter	0.54	0.45	0.49
Etter et al., 2013 [6]	SVM-HMM	0.65	**0.49**	0.54
our approach	Cluster (images and texts) + DT	**0.82**	0.46	**0.59**

Table 3: Performance measures (PER, ORG and LOC classes) of state-of-the-art N
for short texts (Ritter dataset). Approaches which do not rely on hand-crafted r
Gazetteers are highlighted in gray. Etter et al., 2013 trained using 10 class

HORUS 1.0

Named Entity Recognition for

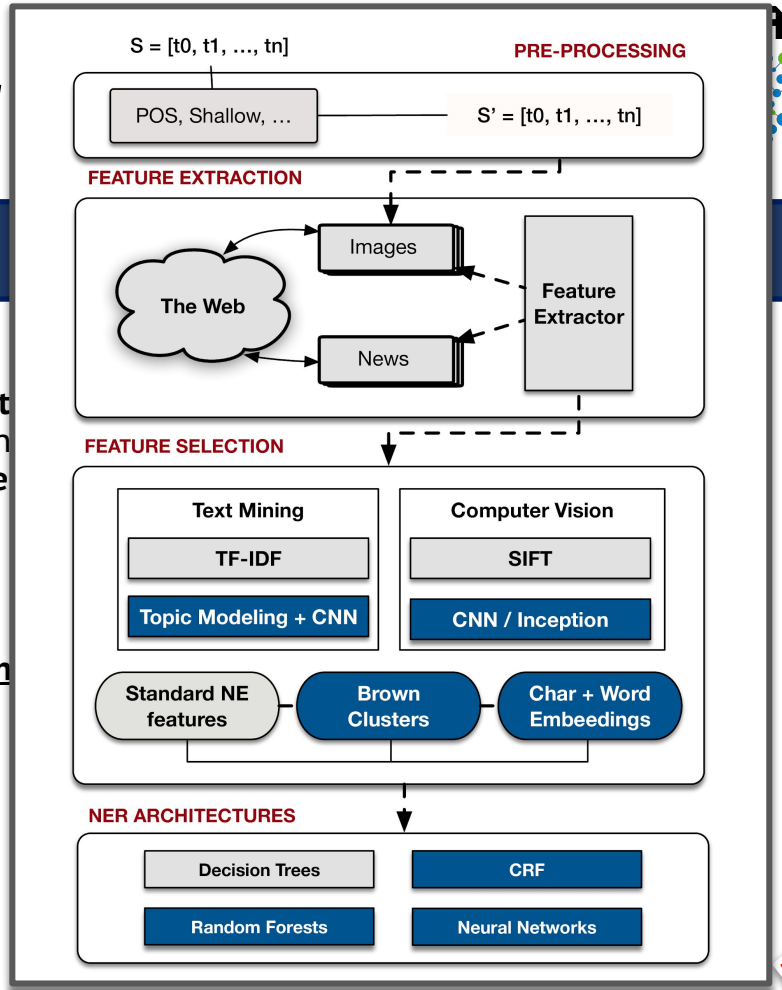
HORUS 2.0

Advantages

- CV module makes the approach **language agnost**
- Each text snipped is **automatically translated** (en
- Very **simple algorithms (DT) performing really we** SOTA)
- **NO Gazetteers**!

Disadvantages

- Still **NOT achieving similar to SOTA in formal dom**
- Do **NOT scale** well!



HORUS 2.0

Named Entity Recognition for Noisy Data

HORUS 2.0

Brown Clusters ($\mathcal{B}$)

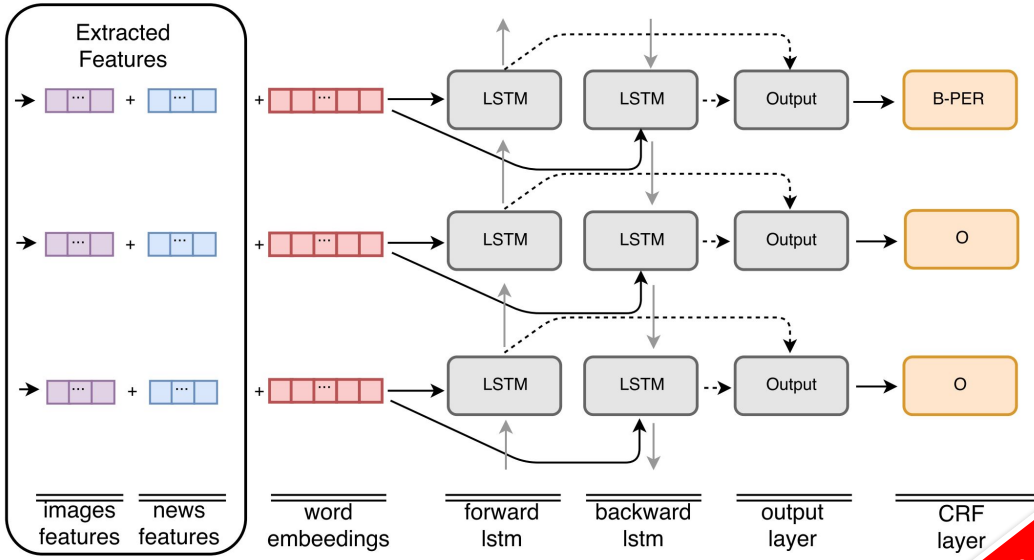
Standard Features: ($\mathcal{S}$)

Topic Modeling + CNNs ($\mathcal{TX}_{nn}$)

Seeds x Word2Vec ($\mathcal{TX}_{emb}$)

Text Correlation ($\mathcal{TX}_{stats}$)

Convolutional Neural Nets ($\mathcal{CV}_{nn}$)

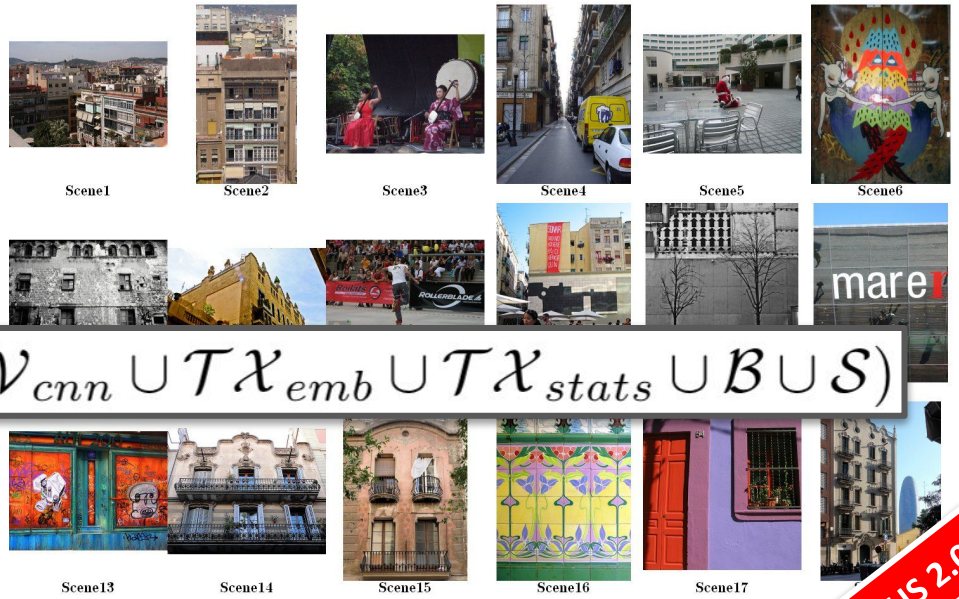


Named Entity Recognition for Noisy Data

HORUS 2.0

- DT + HORUS
- CRF + HORUS
- B-LSTM + CRF + HORUS
- B-LSTM + CNN + CRF + HORUS
- Char + B-LSTM + CRF + HORUS

$$\mathcal{F} = (\mathcal{TX} \cup \mathcal{CV} \cup \mathcal{TX}_{cnn} \cup \mathcal{CV}_{cnn} \cup \mathcal{TX}_{emb} \cup \mathcal{TX}_{stats} \cup \mathcal{B} \cup \mathcal{S})$$



Named Entity Recognition for Noisy Data

Experiment Configurations

Cfg[4]	Features
cfg01	$\mathcal{S}$
cfg02	$\mathcal{S}$ + $\mathcal{TX}$ (TF-IDF+SVM)
cfg03	$\mathcal{S}$ + $\mathcal{CV}$ (SIFT+K-means+SVM)
cfg04	$\mathcal{S}$ + $\mathcal{TX}$ + $\mathcal{CV}$ [15]
cfg05	$\mathcal{S}$ + Lemma
cfg06	$\mathcal{S}$ + Lemma + $\mathcal{TX}$
cfg07	$\mathcal{S}$ + Lemma + $\mathcal{CV}$
cfg08	$\mathcal{S}$ + Lemma + $\mathcal{TX}$ + $\mathcal{CV}$

HORUS 1.1

Cfg	Features
cfg09	$\mathcal{S}$ + Brown 64M c320
cfg10	$\mathcal{S}$ + Brown 64M c640 ($\mathcal{B}_{best}$)
cfg11	$\mathcal{S}$ + Brown 500M c1000
cfg12	$\mathcal{S}$ + Lemma + Brown 64M c320
cfg13	$\mathcal{S}$ + Lemma + Brown 64M c640
cfg14	$\mathcal{S}$ + Lemma + Brown 500M c1000
cfg15	$\mathcal{S}$ + $\mathcal{B}_{best}$ + $\mathcal{CV}$
cfg16	$\mathcal{S}$ + $\mathcal{B}_{best}$ + $\mathcal{TX}$
cfg17	$\mathcal{S}$ + $\mathcal{B}_{best}$ + $\mathcal{CV}$ + $\mathcal{TX}$

Brown Clusters

Cfg	Features	Cfg	Features
cfg18	$\mathcal{S}$+$\mathcal{CV}_{cnn}$	cfg30	=18+$\mathcal{B}_{best}$
cfg19	$\mathcal{S}$+$\mathcal{TX}_{cnn}$	cfg31	=19+$\mathcal{B}_{best}$
cfg20	$\mathcal{S}$+$\mathcal{TX}_{emb}$	cfg32	=20+$\mathcal{B}_{best}$
cfg21	$\mathcal{S}$+$\mathcal{TX}_{stats}$	cfg33	=21+$\mathcal{B}_{best}$
cfg22	$\mathcal{S}$+$\mathcal{TX}_{cnn}$+$\mathcal{TX}$	cfg34	=22+$\mathcal{B}_{best}$
cfg23	$\mathcal{S}$+$\mathcal{TX}_{cnn}$+$\mathcal{TX}$+$\mathcal{TX}_{e}$ +$\mathcal{TX}_{stats}$	cfg35	=23+$\mathcal{B}_{best}$
cfg24	$\mathcal{S}$+$\mathcal{TX}_{cnn}$+$\mathcal{CV}_{cnn}$	cfg36	=24+$\mathcal{B}_{best}$
cfg25	$\mathcal{S}$+$\mathcal{TX}_{cnn}$+$\mathcal{TX}$+$\mathcal{CV}$	cfg37	=25+$\mathcal{B}_{best}$
cfg26	$\mathcal{S}$+$\mathcal{CV}_{cnn}$+$\mathcal{CV}$	cfg38	=26+$\mathcal{B}_{best}$
cfg27	$\mathcal{S}$+$\mathcal{CV}_{cnn}$+$\mathcal{CV}$+$\mathcal{TX}$	cfg39	=27+$\mathcal{B}_{best}$
cfg28	$\mathcal{S}$+$\mathcal{CV}_{cnn}$+$\mathcal{CV}$+$\mathcal{TX}_{cnn}$ +$\mathcal{TX}$	cfg40	=28+$\mathcal{B}_{be}$
cfg29	$\mathcal{S}$+$\mathcal{CV}_{cnn}$+$\mathcal{CV}$+$\mathcal{TX}_{cnn}$ +$\mathcal{TX}$+$\mathcal{TX}_{emb}$+$\mathcal{TX}_{stats}$	cfg41	=2[illegible]

Deep Learning

HORUS 2.0

Named Entity Recognition for Noisy Data

Experiment Configurations

	Description	Configurations	Note
1	Standard	`cfg01, cfg05, cfg09-14`	usual features
2	Brown Clusters	`cfg09-14`	usual features + Brown
3	Images	`cfg03, cfg07, cfg15 cfg18, cfg26`	computer vision (only)
4	Text	`cfg02, cfg06, cfg16 cfg19-23`	text mining (only)
5	Images	`cfg03, cfg07, cfg15`	HORUS 1.0
6	Text	`cfg02, cfg06, cfg16`	HORUS 1.0
7	Images and Text	`cfg04, cfg08, cfg17`	HORUS 1.0
8	Images	`cfg18`	HORUS 2.0
9	Text	`cfg19-23`	HORUS 2.0
10	Images and Text	`cfg24, cfg08, cfg17`	HORUS 2.0

Table 9: Different experiment dimensions: the impact of images and textual features.

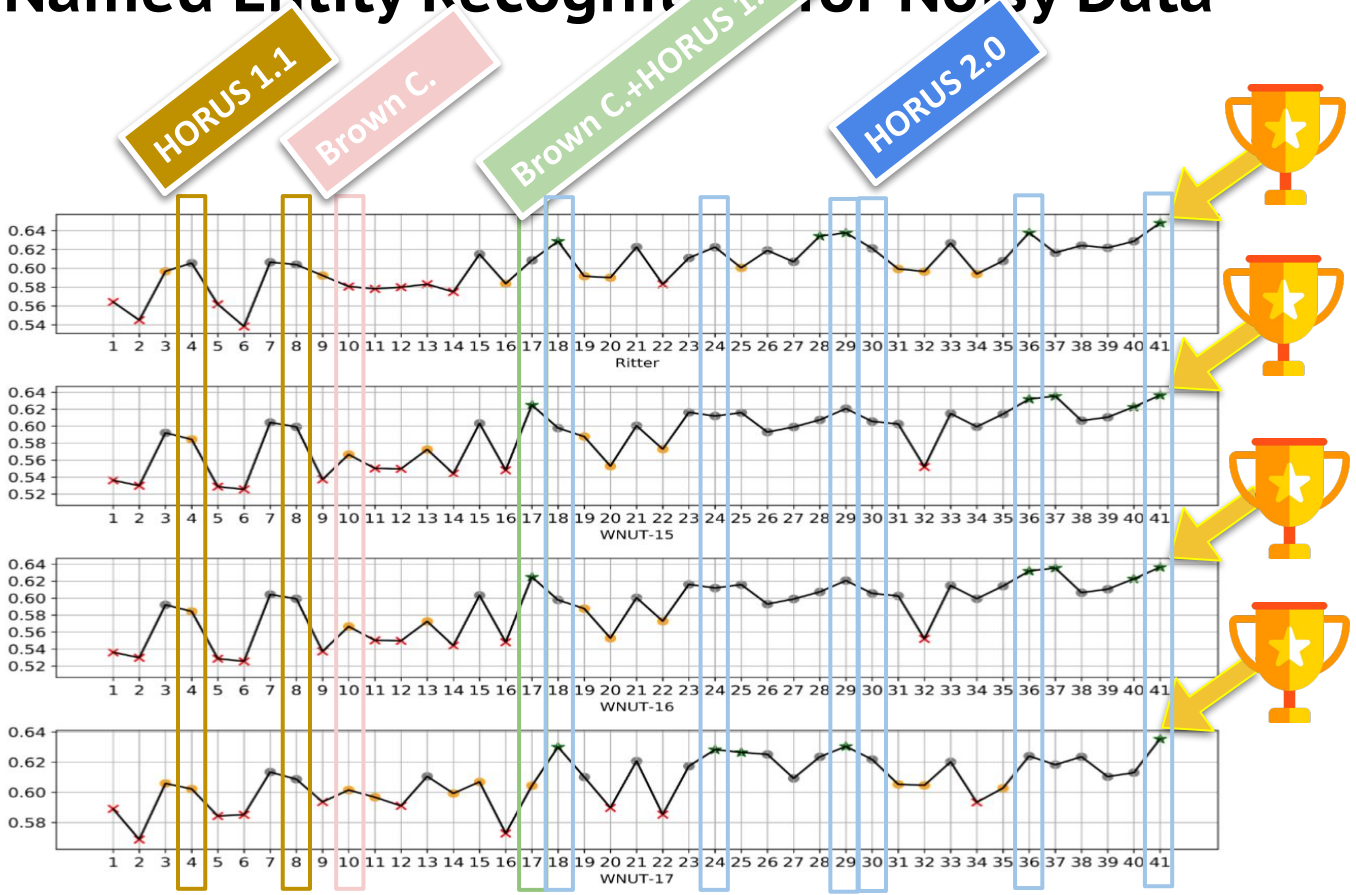
Named Entity Recognition for Noisy Data

Results: off-the-shelf frameworks

	WNUT-15				WNUT-16				WNUT-17			
Tool	PER	LOC	ORG	AVG	PER	LOC	ORG	AVG	PER	LOC	ORG	AVG
NLTK	0.3429	0.2872	0.0711	**0.2961**	0.2943	0.4154	0.1681	**0.3021**	0.3959	0.1954	0.0886	**0.3106**
Spacy	0.3486	0.3799	0.04926	**0.328**	0.3197	0.5488	0.1694	**0.3627**	0.388	0.3221	0.076	**0.3391**
MITIE	0.4623	0.211	0.0465	**0.3353**	0.3881	0.3245	0.1473	**0.2858**	0.3167	0.2349	0.0845	**0.2714**
OSU	0.6078	0.4915	0.3011	**0.5372**	0.4529	0.6566	0.2982	**0.4855**	0.4374	0.4068	0.0813	**0.3938**
Stanford	0.6296	0.27	0.1859	**0.4632**	0.5862	0.5807	0.2841	**0.4878**	0.5723	0.3561	0.1548	**0.4712**

Table 8: Off the shelf NER performance: F1-measure.

Named Entity Recogniti or Noisy Data



Below the baseline

HORUS 1.1

Above the baseline

TOP

HORUS 2.0

Named Entity Recognition for Nois[illegible]

Benchmarking

10 = Baseline

04 = HORUS 1.1

41 = HORUS 2.0

Dataset		Decision Trees			Random Forest			CRF			B-LSTM CRF [20]			B-LSTM C+CRF [23]			B-LSTM C+CRF+CNN [30]		
cfg ⟶		10	04	41	10	04	41	10	04	41	10	04	41	10	04	41	10	04	41
Ritter	P	0.48	+2%	+4%	0.51	+1%	+24%	0.73	+5%	+7%	0.77	+1%	−3%	0.81	−5%	−1%	0.81	−5%	−5%
	R	0.49	+1%	+3%	0.48	−1%	−2%	0.58	−8%	−2%	0.63	+5%	+5%	0.59	+5%	+4%	0.62	+3%	+5%
	F	0.49	+1%	+3%	0.49	+4%	+7%	0.58	+2%	+7%	0.68	+1%	+1%	0.67	+1%	+1%	0.69	−1%	+1%
WNUT-15	P	0.49	+2%	+5%	0.52	+7%	+25%	0.72	+7%	+9%	0.72	−4%	−2%	0.77	−3%	−4%	0.78	−4%	−5%
	R	0.50	+0%	+5%	0.49	+0%	+1%	0.48	−1%	+6%	0.69	+1%	+1%	0.65	+2%	+2%	0.66	+2%	+2%
	F	0.50	+0%	+5%	0.50	+5%	+9%	0.56	+2%	+8%	0.68	+0%	+0%	0.69	+0%	−1%	0.71	−1%	−2%
WNUT-16	P	0.49	+1%	+6%	0.52	+14%	+23%	0.72	+7%	+9%	0.72	−4%	−2%	0.77	−3%	−3%	0.78	−4%	−6%
	R	0.50	+1%	+6%	0.48	+0%	+2%	0.48	−1%	+6%	0.69	+0%	+1%	0.65	+2%	+2%	0.66	+2%	+2%
	F	0.49	+1%	+6%	0.50	+5%	+10%	0.56	+2%	+8%	0.69	−1%	+0%	0.69	+0%	+0%	0.71	−1%	−2%
WNUT-17	P	0.44	+3%	+7%	0.47	+13%	+24%	0.76	+2%	+1%	0.76	−2%	−2%	0.76	+0%	−2%	0.77	−3%	−3%
	R	0.45	+4%	+6%	0.44	+3%	+4%	0.50	+0%	+5%	0.63	+1%	+1%	0.64	+0%	+1%	0.62	+1%	+1%
	F	0.44	+4%	+6%	0.45	+6%	+12%	0.60	+0%	+4%	0.67	+0%	+0%	0.69	+0%	−1%	0.67	+0%	−1%

HORUS 2.0

Named Entity Recognition for Noisy Data

Conclusions

Contributions made:

- Novel NER Architecture based on Images and News (**NO Gazetteer!**)
- **Language Agnostic** NER Framework for Noisy Data (English and Portuguese)
- **Improved Recall** for NNs, but at cost of precision
- **Great improvement for CRF-based** models; results comparable to SOTA NNs

Named Entity Recognition for Noisy Data

Can images along with news improve
the performance of the named entity recognition mod… …oisy te…

04 = HORUS 1.0

41 = HORUS 2.0

	+cfg10	+cfg04	+cfg41
B-LSTM+CRF	0.5217	0.5352 ↑	0.5376 ↑

Table 5: B-LSTM+CRF F1-measure with expanded training/dev/test data over different feature sets.

Named Entity Recognition for Noisy Data

Future Work

HORUS 3.0

HORUS-NER:
A Multimodal (and **Multilingual**) Named Entity Recognition (and **Linking**) Framework for Noisy Text

- Benchmark segmentation
- Benchmark the method over different languages
- Extend it to perform linking (e.g., DBPedia)
- Possibly leverage other (complex) NLP tasks (e.g., WSD, EL, QA, etc..)
- Explore contextual word embeddings (instead of “static” ones)
- Check normalization methods, slangs, geographic context

Thank you!

www.linkedin.com/in/diegoestevesde/

github.com/SmartDataAnalytics/horus-ner

diegoesteves at gmail dot com