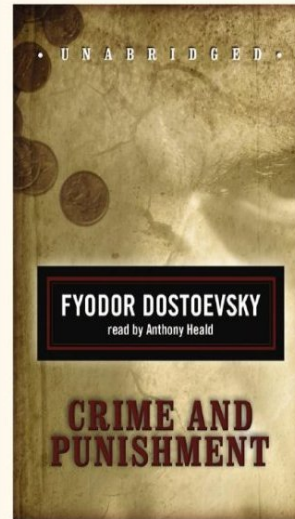# Distant Supervision for Relation Extraction without Labeled Data

Joshua Ehrlich and Michael Stock-Matthews

Background

- Goal is to extract relations between entities

- Previous work uses
  - supervised learning (expensive to get good data)
  - unsupervised learning (limited control over results)
  - bootstrapping (semantic drift)

Distant Supervision

- Collect training data based on sentences containing entities in known relations
- Extract features from these sentences
- (combine features from sentences that have the same two entities)
- Use this to train a multiclass logistic regression classifier

Freebase

- Freebase is a database of binary ordered relations between entitites
- Collected from many sources including text boxes from Wikipedia
- Merged reverse relations and duplicates for a total of 1.8 million instances

Features

- Lexical
- Syntactic
- Named Entity tag

Feature type	Left window	NE1	Middle	NE2	Right window
Lexical	[]	PER	[was/VERB born/VERB in/CLOSED]	LOC	[]
Lexical	[Astronomer]	PER	[was/VERB born/VERB in/CLOSED]	LOC	[,]
Lexical	[#PAD#, Astronomer]	PER	[was/VERB born/VERB in/CLOSED]	LOC	[, Missouri]
Syntactic	[]	PER	$[\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}]$	LOC	[]
Syntactic	[Edwin Hubble $\Downarrow_{lex-mod}$]	PER	$[\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}]$	LOC	[]
Syntactic	[Astronomer $\Downarrow_{lex-mod}$]	PER	$[\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}]$	LOC	[]
Syntactic	[]	PER	$[\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}]$	LOC	$[\Downarrow_{lex-mod}$,]
Syntactic	[Edwin Hubble $\Downarrow_{lex-mod}$]	PER	$[\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}]$	LOC	$[\Downarrow_{lex-mod}$,]
Syntactic	[Astronomer $\Downarrow_{lex-mod}$]	PER	$[\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}]$	LOC	$[\Downarrow_{lex-mod}$,]
Syntactic	[]	PER	$[\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}]$	LOC	$[\Downarrow_{inside}$ Missouri]
Syntactic	[Edwin Hubble $\Downarrow_{lex-mod}$]	PER	$[\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}]$	LOC	$[\Downarrow_{inside}$ Missouri]
Syntactic	[Astronomer $\Downarrow_{lex-mod}$]	PER	$[\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}]$	LOC	$[\Downarrow_{inside}$ Missouri]

Table 3: Features for 'Astronomer Edwin Hubble was born in Marshfield, Missouri'.

Dataset

- Unstructured text from Wikipedia
  - tokenized into sentences by Metaweb Technologies
  - sentences parsed by Minipar
- For training used 800,000 articles and a different 400,00 for testing
  - to generate negative examples randomly select entity pairs that are not in any relationship
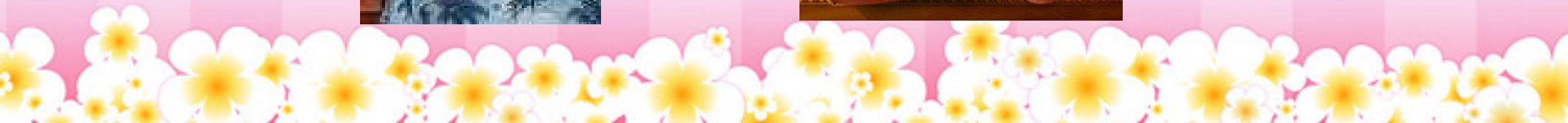
Evaluation

- Measuring precision of relation instances extracted
- Automatic evaluation using held-out evaluation
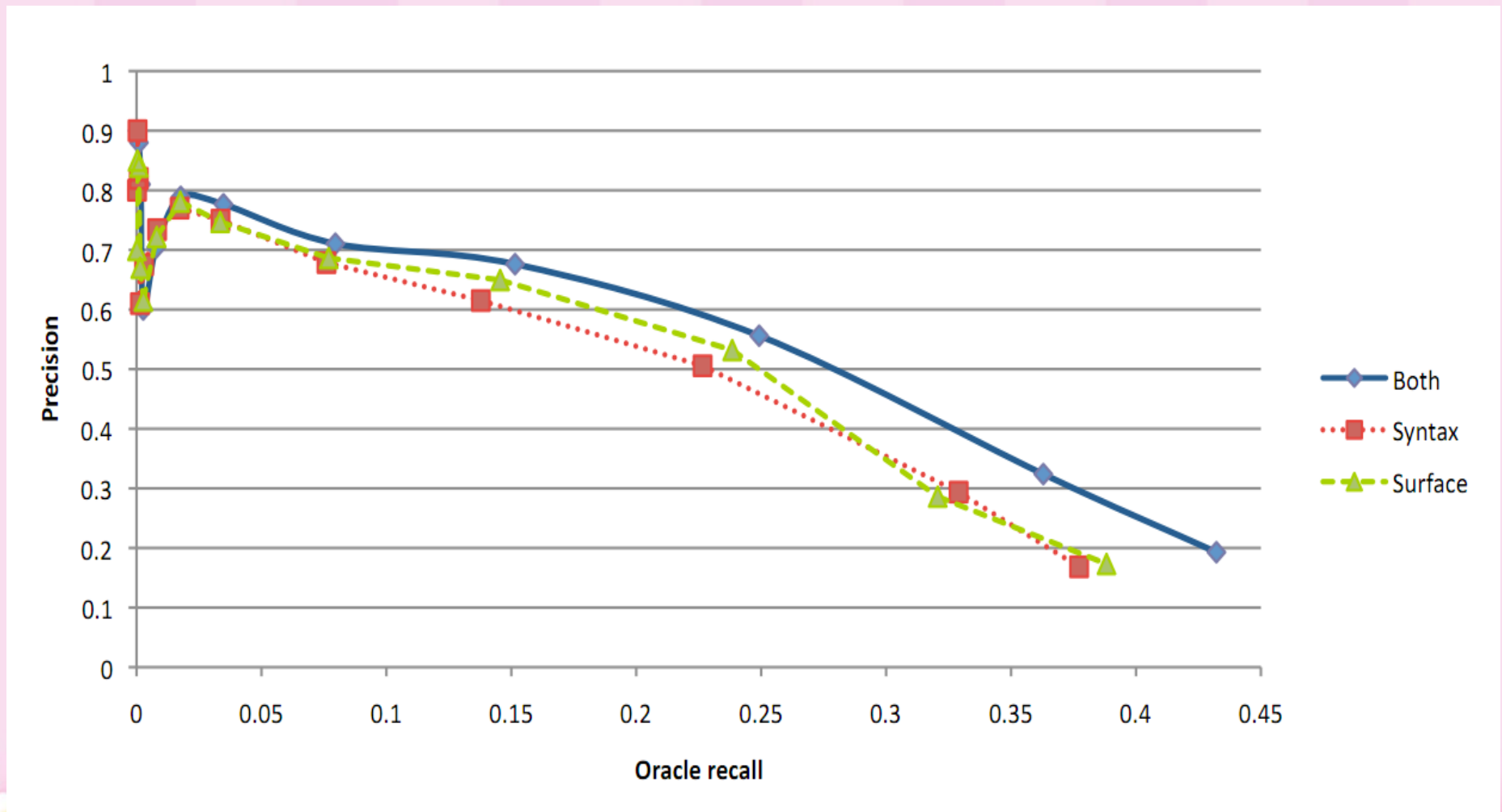
Evaluation (continued)

- Human evaluation using Mechanical Turk

Relation name	100 instances			1000 instances		
	Syn	Lex	Both	Syn	Lex	Both
/film/director/film	**0.49**	0.43	0.44	**0.49**	0.41	0.46
/film/writer/film	**0.70**	0.60	0.65	**0.71**	0.61	0.69
/geography/river/basin_countries	0.65	0.64	**0.67**	**0.73**	0.71	0.64
/location/country/administrative_divisions	0.68	0.59	**0.70**	**0.72**	0.68	**0.72**
/location/location/contains	0.81	**0.89**	0.84	**0.85**	0.83	0.84
/location/us_county/county_seat	0.51	0.51	**0.53**	0.47	**0.57**	0.42
/music/artist/origin	0.64	0.66	**0.71**	0.61	**0.63**	0.60
/people/deceased_person/place_of_death	0.80	0.79	**0.81**	0.80	**0.81**	0.78
/people/person/nationality	0.61	0.70	**0.72**	0.56	0.61	**0.63**
/people/person/place_of_birth	**0.78**	0.77	**0.78**	0.88	0.85	**0.91**
Average	0.67	0.66	**0.69**	**0.68**	0.67	0.67

Discussion

- Extract high precision relationships
- Syntactic features are generally more useful although the combination may be better

Critique

- There is no comparison of results with previous work
  - Unclear if these results are actually high precision
- Their claim that syntactic features are better is not clearly justified given the data
- Their explanation of their automatic evaluation and the results is unclear

Questions