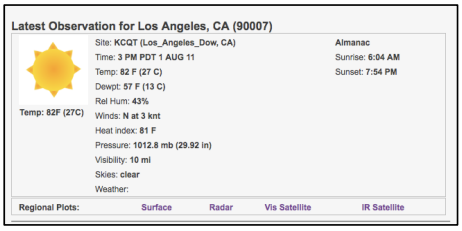
Using Conditional Random Fields to Exploit Token Structure and Labels for Accurate Semantic Annotation

Aman Goel, Craig A. Knoblock, Kristina Lerman
Information Sciences Institute and Computer Science Department
University of Southern California

Problem:

Data Sources



42103003710000	Crier McElroy	31.392944	-102.314314	3050.0
42103003840000	J.T. McElroy	31.409058	-102.342453	3015.0
42103004630000	J.T. McElroy	31.381781	-102.323448	3159.0

Extraction

Extracted Data

90292	76° F	50%	10 miles

Tokenization

90292	76 ° F	50 %	10 miles

? The identity of the data is unknown

Semantic Labels

Zip	TemperatureF	Humidity	Visibility
ZipToken	TemperatureFValue DegreeSymbol TemperatureFUnit	HumidityValue PercentSymbol	VisibilityValue VisibilityUnit

Solution:

Exploit structure within fields for accurate labeling.

Contributions:

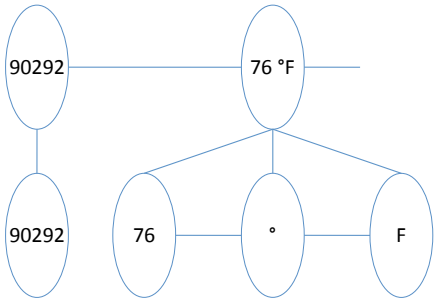
✓ We exploit latent structure within fields that makes the model robust to changes in structure.
✓ Our complexity of inference remains low even while using cyclic graphs.
✓ Our model achieves higher labeling accuracy than two different baseline approaches.

Future Work:

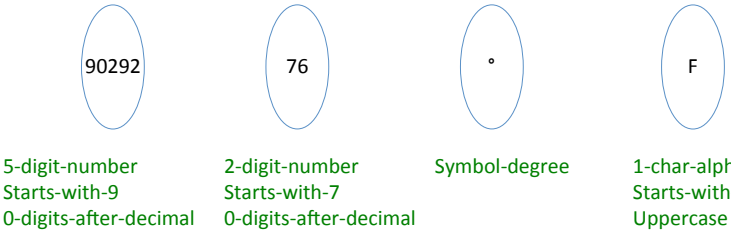
✓ Pruning feature function space to select only those feature functions that discriminate a semantic type from others.
✓ Discovering new features from the training examples.
✓ Composing elementary feature functions into complex feature functions using conjunction and disjunction for greater expressiveness.

Approach:

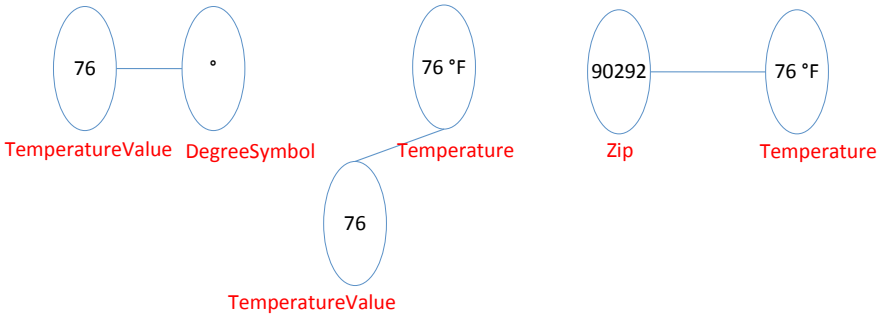
Step1: Convert the data into a CRF graph



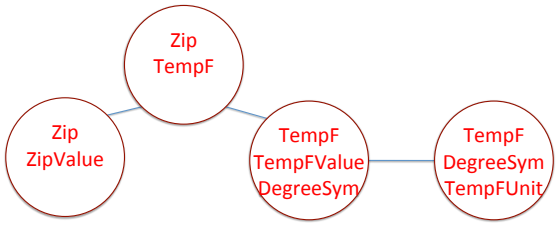
Step2: Extract syntactic features from tokens



Step 3: Generate feature functions from labeled examples



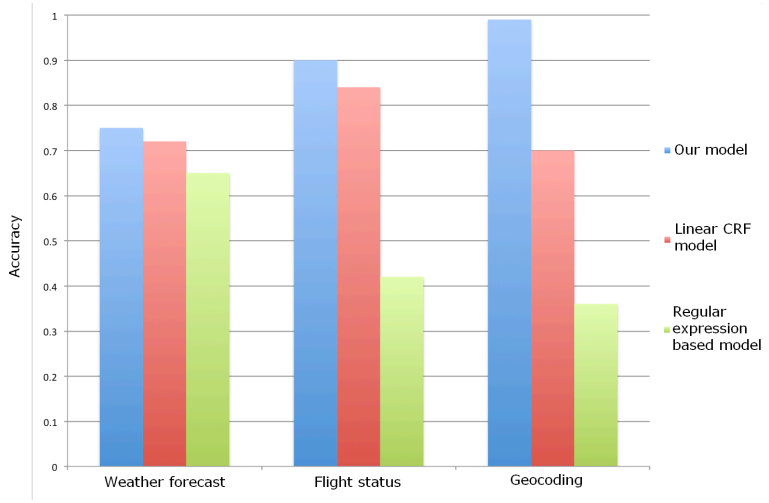
Step 4: Train the model on junction trees



Results:

Domain	Data source	Field accuracy	Token accuracy
Weather forecast	wunderground.com	0.89	0.92
	weather.unisys.com	0.43	0.75
	weather.com	0.70	0.79
	noaa.gov	1.00	0.86
Flight status	flytecomm.com	0.89	0.82
	flightview.com	0.96	0.97
	delta.com	0.81	0.78
	continental.com	0.96	0.55
Geocoding	geocoder.us	1.00	0.85
	geocoder.ca	1.00	0.82
	geonames.com	0.98	0.68
	worldkit.com	1.00	0.89

Comparison with baseline approaches:



Related Work:

- 2D conditional random fields for web information extraction, Zhu et al. (ICML, 2005)
    - We exploit the internal structure of the fields. Also, our graphs are cyclic.
- Tree-structured conditional random fields for semantic annotation, Tang et al. (ISWC, 2006)
    - Our approach is applicable to sources with different field ordering because we exploit the internal structure of the fields. lexity of inference remains low even while using cyclic graphs.
- A survey of approaches to automatic schema matching, Rahm and Bernstein (VLDB, 2001)
    - Our approach builds one model of each semantic type, instead of combining matching scores with different formats.