

# Adventures in Statistical Pronoun Interpretation

Andrew Kehler  
UCSD Linguistics

(Joint work with Douglas Appelt, Lara Taylor, and  
Aleksandr Simma)

1

# Coreference

*Hillary Clinton* has taken the lead among Democratic  
presidential candidates in an *Iowa* poll, a sign of progress  
for *her* campaign of progress toward overcoming a big  
hurdle in the race.

Although *the New York senator* is the clear front-runner in  
national surveys, *Iowa* has remained an elusive prize. *She*  
has been in a tight race with *John Edwards and Barack*  
*Obama in the state that begins the primary campaign voting*  
in three months.

2

# Pronoun Interpretation

*Perdue said circumstances in Maryland are particularly dire, because production costs here already are higher than in other regions of the country. He worried that Glendening's initiative could push his industry over the edge, forcing it to shift operations elsewhere.*

3

# Outline

- ◆ Part I
  - ◆ A supervised model
  - ◆ Using predicate-argument frequencies
- ◆ Part II
  - ◆ A self-trained system

4

## Background

- ◆ State-of-the-art systems for pronoun interpretation typically rely on a variety of morphosyntactic features
- ◆ A common refrain: performance is plateauing, and further progress will require world knowledge and inference
- ◆ Unfortunately, we don't have that...

5

## A More Tractable Version

- ◆ It has been suggested that we might use predicate-argument frequencies as a substitute (Lappin and Leass 1994, Dagan et al. 1995)

*He worries that Glendening's initiative could push his industry over the edge, forcing **it** to shift operations elsewhere*

6

## Dagan et al. (1995)

- ◆ Added a predicate-argument frequency post-processor onto Lappin and Leass' (1994) symbolic pronoun interpretation system
- ◆ Report a modest rise in performance (2.5%, 9 additional correct pronouns out of 360)
- ◆ Suggest that improved performance could result from using a statistical approach in which predicate-argument features were modeled alongside morphosyntactic ones

7

## Goals

- ◆ We set out to evaluate this suggestion
- ◆ We train a statistical model based on a set of morphosyntactic features, and
- ◆ Augment it with predicate-argument statistics in two scenarios:
  - ◆ Using a Dagan et al. postprocessor
  - ◆ Modeling them with features

8

## Corpora Used

- ◆ Training and test sets from the newspaper and newswire segments of the ACE corpus
  - ◆ 2773 pronouns for training (reannot v.4)
  - ◆ 762 pronouns for testing (Feb 02 eval set)
- ◆ Annotated pronouns only included ACE “markables” -- Persons, Organizations, GeoPoliticalEntities, Locations, and Facilities

9

## Algorithms

- ◆ Input from SRI's TextPro system, a chunk-style shallow parser (noun and verb groups)
- ◆ Systems:
  - ◆ Maximum Entropy (MaxEnt)
    - ◆ Trained as a classifier (2 outcomes)
    - ◆ Testing: highest probability of coreference
  - ◆ Naïve Bayes
  - ◆ Adapted Hobbs algorithm baseline

10

## Features

- ◆ Hard constraints based on conservative gender and number tests
- ◆ Five categories of ‘soft’ features:
  - ◆ Number
  - ◆ Gender
  - ◆ Distance between pronoun and antecedent
  - ◆ Grammatical role of antecedent
  - ◆ Linguistic form of antecedent

11

## Predicate-Argument Frequencies

- ◆ Ran TextPro over the TDT-2 newswire corpus
- ◆ Collected stats for three types of relationships:
  - ◆ Subject-Verb (1,321,072)
  - ◆ Verb-Object (1,167,189)
  - ◆ Possessive-Noun (301,477)
- ◆ Lemmas used when available, proper names categorized by ACE entity type

12

## Integrating Pred-Arg Statistics

- ◆ The postprocessor uses two equations
- ◆ Frequency with which a candidate head noun C is found with the predicate word A:

$$\text{stat}(C) = \text{freq}[\text{tuple}(C,A)] / \text{freq}(C)$$

- ◆ The difference in co-occurrence versus the difference in salience assigned by the morphosyntactic model:

$$\ln(\text{stat}(C2) / \text{stat}(C1)) > K \times [\text{sal}(C1) - \text{sal}(C2)]$$

13

## Smoothing

- ◆ Two approaches to smoothing:
  - ◆ Good-Turing (following Dagan et al.)
  - ◆ Pereira, Tishby, and Lee (1993) clustering
- ◆ Final systems performed similarly between the two; will report Good-Turing results

14

## Factoids

- ◆ All development evaluated via jackknifing on training data
- ◆ Statistics won't flip a pronominal antecedent
- ◆ Probabilistic ties broken by Hobbs distance
- ◆ Performance ceiling is 91.6% do to misanalyses

15

## Results (Blind Test)

Features	MaxEnt	Maxent-Features	MaxEnt-Postprocessing	Naive Bayes
none	0.6877	0.6496	0.6627	0.6877
num, gend	0.6667	0.6745	0.6719	0.6654
num, gend, dist	0.7336	0.7415	0.7428	0.7297
num, gend, dist, pos	0.7441	0.7507	0.7520	0.7441
num, gend, dist, pos, lform	0.7572	0.7572	0.7677	0.7415

16

## Error Analysis (I)

- ◆ Negligible improvement could be due to either of two reasons:
  - ◆ Predicate-argument statistics are not good predictors for pronoun interpretation
  - ◆ Data sparsity
- ◆ Performed an error analysis to tease these apart

17

## Error Analysis (II)

- ◆ Analyzed corpus of errors made during jackknifing on training data
- ◆ Filtered cases that:
  - ◆ Did not participate in one of three relations
  - ◆ Had a pronoun or proper name as either the correct or chosen antecedent
  - ◆ Had a headless antecedent

18

## Error Analysis (III)

- ◆ Used a variant of Keller and Lapata's (2003) technique for unseen bigram estimation
- ◆ Collected the number of AltaVista search hits for each predicate argument combination (subject-verb, verb-object, possessive-noun) and its morphological variants
- ◆ Used normalized counts

19

## Error Analysis (IV)

- ◆ The paper reports on a 20-example sample; stats helped in 10 cases and hurt in 10
- ◆ Less restrictive filtering has since produced 89 examples, which break down as follows:
  - ◆ Stats Helped: 52 cases
  - ◆ Stats Hurt: 32 cases
  - ◆ Tie: 5 cases

20

## Error Analysis (V)

- ◆ A 'textbook' positive case:

*After the endowment was publicly excoriated for having the temerity to award some of **its** money to art that addressed changing views of...*

21

## Error Analysis (VI)

- ◆ A negative case:

*The dancers were joined by about 70 supporters as **they** marched around a fountain not far from the mayor's office, chanting "Giuliani \_ scared of sex! Who's he going to censor next?"*

22

## Error Analysis (VII)

- ◆ Further, many of positive cases seemed more due to fortuity than anything else:

*..., the sources said. Sullivan took the money despite having previously voiced suspicions that Chung was acting as a conduit for illegal contributions from Chinese business executives, **they** added.*

23

## Bean And Riloff (2004)

- ◆ Report an improvement of 6% recall and 1-2% precision for pronouns using lexical statistics in two domains
- ◆ A difference: Their system declines to make a resolution when belief < 50%, "which is often the case"
- ◆ Stats appear to push more pronouns over the threshold, hence the recall improvement over their baselines (42% for terrorism domain and 50% for disasters domain)

24

## Bean and Riloff (2004)

- ◆ Implemented a similar strategy in our system
  - ◆ In baseline, held back pronouns for which the prob of no candidate beat a threshold (prob=0.5)
  - ◆ In B&R version, took highest-probability candidate in holdouts that had a positive pred-arg count
- ◆ Results:
  - ◆ Baseline: R=.520, P=.792, F=.628
  - ◆ B&R: R=.676, P=.749, F=.710
- ◆ Big gain, but still worse than our baseline system (F=.757)

25

## Conclusion

- ◆ Experimental results and error analysis suggest that the predictive power offered by predicate-argument statistics is very modest
- ◆ A majority of cases can be resolved using morphosyntactic features
- ◆ Pred-arg stats appear to be a poor substitute for the knowledge needed to interpret the rest

26

## Outline

- ◆ Part I
  - ◆ A supervised model
  - ◆ Using predicate-argument frequencies
- ◆ Part II
  - ◆ A self-trained system

27

## Background

- ◆ Pronoun interpretation systems with parameters that are tuned manually or by supervised learning require a substantial corpus of annotated data
- ◆ We report on a self-trained system which annotates its own data for training

28

## Details

- ◆ Most details are the same as in the supervised system just described:
  - ◆ Same input (shallow parses from TextPro)
  - ◆ Same hard constraints and features
  - ◆ Same supervised algorithm (MaxEnt)
  - ◆ Same blind evaluation data (762 pronouns)

29

## Details

- ◆ The difference:
  - ◆ Supervised: Training corpus of 2773 annotated pronouns (ACE newswire/newspaper)
  - ◆ Self-trained: MaxEnt embedded in a loop during which the algorithm annotates its own data (2773 pronouns from raw text from the TDT-2 newswire corpus)

30

## The Self-Trained System: Initial Pass

1. For each third-person pronoun:
  - a. Collect possible noun-group antecedents
  - b. Filter them by applying hard constraints
  - c. If only one antecedent remains, label coreference as True
  - d. Otherwise label each possible antecedent as False
2. Train a MaxEnt classifier on this data

31

## The Self-Trained System: Loop

1. For each pronoun:
  - a. Apply the current MaxEnt model to each pronoun-antecedent pair
  - b. Label the most likely antecedent as True, and all the rest (if any) False
2. Retrain the MaxEnt model
3. Repeat until convergence

32

## Simplifying Assumptions

- ◆ The idea that each pronoun will be associated with exactly one correct antecedent is obviously bogus:
  - ◆ There may be more than one antecedent
  - ◆ Some of the pronouns will not be entity-referring (pleonastics, events, situations, etc.)
  - ◆ Misparsing will miss correct antecedents and create nonexistent ones
- ◆ Yet the question remains of how far we can get

33

## Results

(Ceiling: 91.6%; Hobbs baseline: 68.8%)

	Basic	Pred-Arg Stats	Newswire Data Only	Both
Supervised	75.7%	76.8%	75.8%	76.7%
Self-Trained (14 runs)	73.4%	75.1%	75.0%	76.4%

34

## Effect of Training Data Size

# of Pronouns	Blind Test Accuracy
55	71.4%
138	72.3%
277	72.5%
554	72.6%
1386	73.5%
2773	73.4%
5546	73.5%
Full Segment	73.7%

35

## Conclusion

- ◆ A self-trained pronoun interpretation system can achieve performance that approaches supervised performance, even though the self-labeled data is no doubt highly noisy
- ◆ The distributions found in unambiguous cases appear to apply well to ambiguous cases (at least for classification)

36

## What This Means

- ◆ We can now consider running experiments on large data sets with very specific features to see if self-trained systems can overtake supervised ones
- ◆ Anyone up for that?

37

## Interpretation is not Classification

- ◆ Twin Candidate Model (Yang et al., 2003)
- ◆ Reranking (Denis and Baldrige, 2007a)
- ◆ More global approaches to coreference (McCallum and Wellner, 2004; Denis and Baldrige, 2007b; Haghghi and Klein, 2007)

38

That's it!

39