The Same-head Heuristic for Coreference

Micha Elsner
joint work with Eugene Charniak and Mark Johnson

Department of Computer Science
Brown University

January 19, 2010

Same-head coreference

Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it, ‘and what is the use of a book,’ thought Alice ‘without pictures or conversation?’

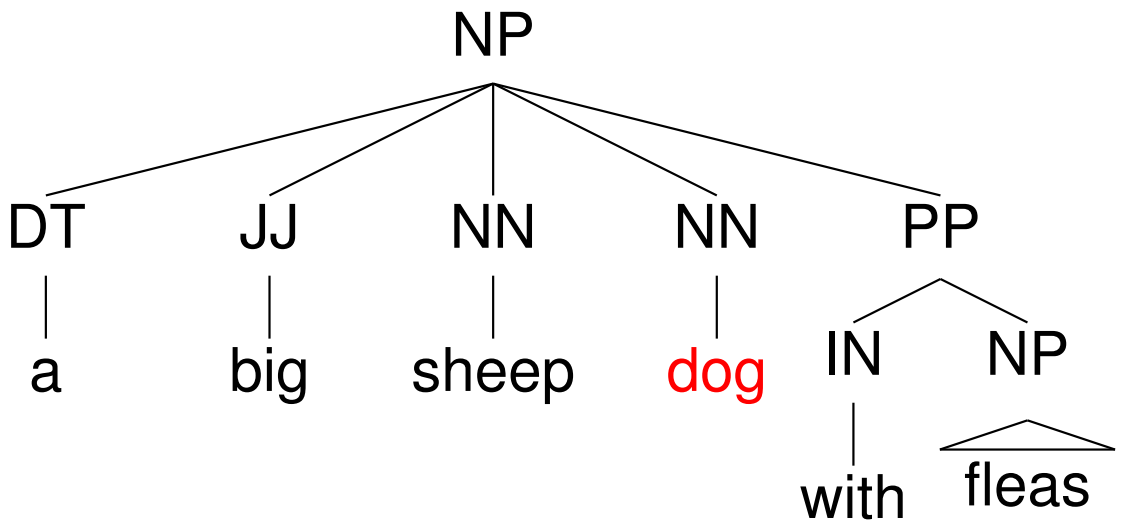
Same-head coreference

Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it, ‘and what is the use of a book,’ thought Alice ‘without pictures or conversation?’

Head

Head word

The “main word” in a phrase.



Same-head coreference

Same-head heuristic

If two NPs have the same head, they are coreferent.

A natural starting point!

- Easy to code
- Works pretty well
- Can be *very* good in some experimental conditions
- Most work focuses on hard cases
  - Non-matching NPs
  - Pronouns

Overview

Introduction

Mention detection and scoring matter

Non-coreferent same-head pairs

Conversational speech is different

Modeling

Overview

Introduction

Mention detection and scoring matter

Non-coreferent same-head pairs

Conversational speech is different

Modeling

Related work

We *know* same-head pairs don't always corefer.

- (Poesio+Vieira) do some counts.
- (Stoyanov+al) system scores (MUC):
  - NPs where all words match: .82
  - Some words match: .53
  - No words match: .27
- Same head is the easy case...
- But not *that* easy

Unsupervised systems

Unsupervised work uses the same-head heuristic.

- (Haghighi+Klein '07): sparse prior on p(word|entity)
- (Poon+Domingos '08): head-prediction clause
- (Haghighi+Klein '09): direct assumption
- partial exception: (Ng '08)

Why can they get away with this?

Mention detection

Gold mentions

- Anything marked by a MUC annotator
- Small subset of NPs
- Used by most unsupervised systems

Annotators don't mark singleton NPs!

- Most of the exceptions are singletons
- This setting is too easy (Stoyanov+al)

Example

However, the Multiplication Table doesn't signify: let's try Geography. London is the capital of Paris, and Paris is the capital of Rome, and Rome— no, THAT'S all wrong, I'm certain!

More realistic option

All NPs

- Reasonable alternative
- Could improve recall by parsing into NPs (Vadas+Curran)

Example

However, the Multiplication Table doesn't signify: let's try Geography. London is the capital of Paris, and Paris is the capital of Rome, and Rome— no, THAT'S all wrong, I'm certain!

Option maximizing recall

All nouns

- Including premodifiers, like “a Bush spokesman”
- Highest possible recall rates

Example

However, the Multiplication Table doesn’t signify: let’s try Geography. London is the capital of Paris, and Paris is the capital of Rome, and Rome— no, THAT’S all wrong, I’m certain!

Comparison

Oracle system

Links NP pairs:

- Same heads
- Within 10 sentences
- **Actually coreferent**

Link all

Links NP pairs:

- Same heads
- Within 10 sentences
- **Always!**

Comparison

	Mentions	Linked	b^3 pr	rec	F
Gold mentions					
Oracle	1929	1164	100	32.3	48.8
Link all	1929	1182	80.6	31.7	45.5
NPs					
Oracle	3993	864	100	30.6	46.9
Link all	3993	1592	67.2	29.5	41.0
Nouns					
Oracle	5435	1127	100	41.5	58.6
Link all	5435	2541	56.6	40.9	45.7

Comparison

	Mentions	Linked	b^3 pr	rec	F
Gold mentions					
Oracle	1929	**1164**	100	32.3	48.8
Link all	1929	**1182**	80.6	31.7	45.5
NPs					
Oracle	3993	864	100	30.6	46.9
Link all	3993	1592	67.2	29.5	41.0
Nouns					
Oracle	5435	1127	100	41.5	58.6
Link all	5435	2541	56.6	40.9	45.7

Comparison

	Mentions	Linked	b^3 pr	rec	F
Gold mentions					
Oracle	1929	1164	100	32.3	48.8
Link all	1929	1182	80.6	31.7	45.5
NPs					
Oracle	3993	**864**	100	30.6	46.9
Link all	3993	**1592**	67.2	29.5	41.0
Nouns					
Oracle	5435	1127	100	41.5	58.6
Link all	5435	2541	56.6	40.9	45.7

Comparison

	Mentions	Linked	b^3 pr	rec	F
Gold mentions					
Oracle	1929	1164	100	32.3	48.8
Link all	1929	1182	80.6	31.7	45.5
NPs					
Oracle	3993	864	100	30.6	46.9
Link all	3993	1592	67.2	29.5	41.0
Nouns					
Oracle	5435	**1127**	100	41.5	58.6
Link all	5435	**2541**	56.6	40.9	45.7

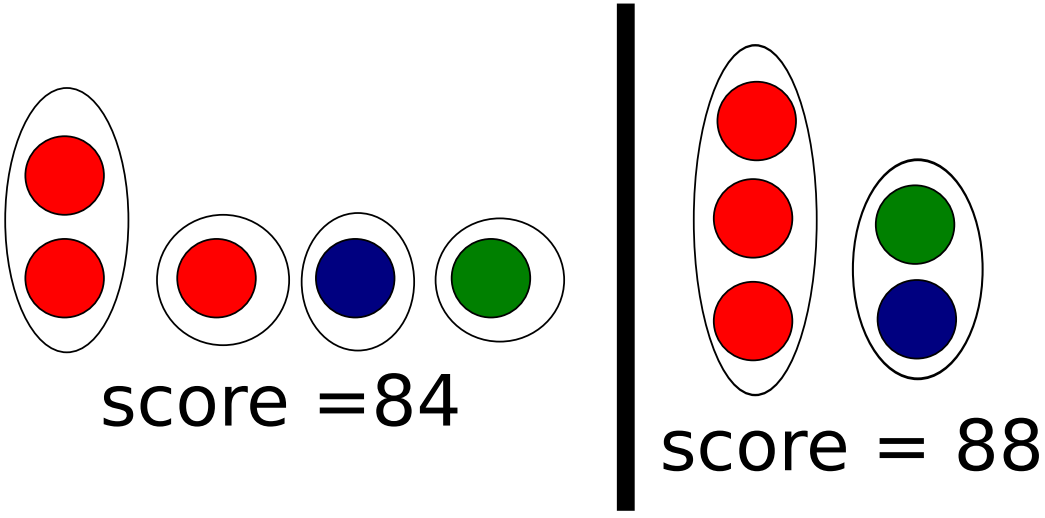
Comparison

	Mentions	Linked	b^3 pr	rec	F
Gold mentions					
Oracle	1929	1164	100	32.3	**48.8**
Link all	1929	1182	80.6	31.7	**45.5**
NPs					
Oracle	3993	864	100	30.6	**46.9**
Link all	3993	1592	67.2	29.5	**41.0**
Nouns					
Oracle	5435	1127	100	41.5	**58.6**
Link all	5435	2541	56.6	40.9	**45.7**

What about metrics?

b^3 (Bagga+Baldwin)

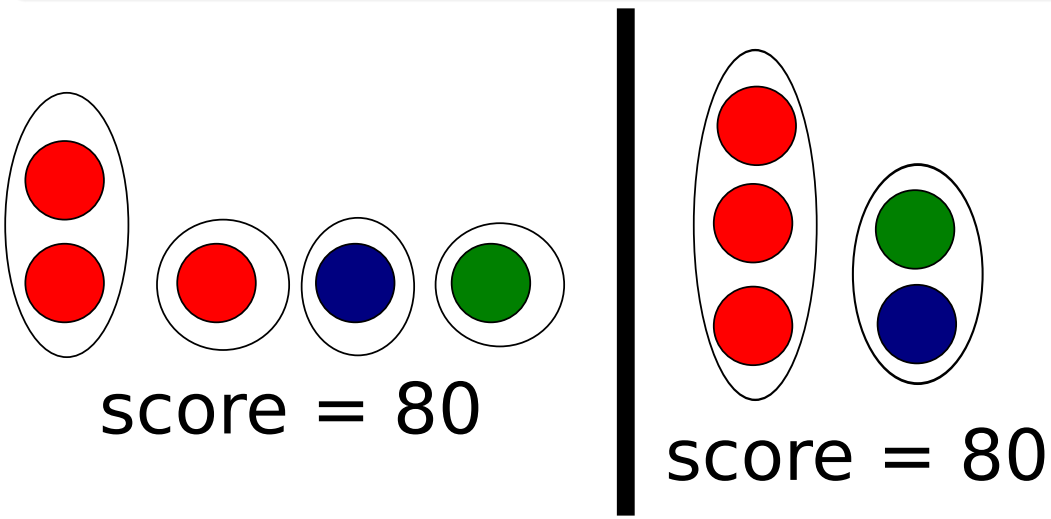
- Precision: correct coreferent NPs / proposed coreferent NPs
- Recall: correct coreferent NPs / true coreferent NPs
- More important to get the big clusters right
- Easier to get high precision
- So best to work on maximizing recall



CEAF

CEAF (Luo)

- Same as one-to-one match for clustering
- Map proposed clusters to actual clusters
- No precision/recall tradeoff



Comparison (again)

	b^3 pr	rec	F	mention CEAF
Gold mentions				
Oracle	100	32.3	**48.8**	**54.4**
Link all	80.6	31.7	**45.5**	**53.8**
NPs				
Oracle	100	30.6	**46.9**	**73.4**
Link all	67.2	29.5	**41.0**	**62.2**
Nouns				
Oracle	100	41.5	**58.6**	**83.5**
Link all	56.6	40.9	**45.7**	**67.0**

What we've learned

- You can get away with using the same-head heuristic...
- Because it works reasonably well
- Using gold mention boundaries
- Using metrics that count links (b^3, link F)

Overview

Introduction

Mention detection and scoring matter

Non-coreferent same-head pairs

Conversational speech is different

Modeling

Quick survey: the MUC data

Did some counting:

- MUC-6 dev
- 100 random pairs: same head, not coreferent
- Ad-hoc categories

Results

Two different entities	39

Different entities

Both NPs refer, but not to the same thing.

- “Recent employees”; “long-time employees”
- “American... the company”; “Hormel... the company”

Results

Two different entities	39
Time/measure phrase	24

Time/measure

- “Last week”; “this week”; “for a week”
- “a billion dollars”; “2.5 billion dollars”

Almost never coreferent.

Results

Two different entities	39
Time/measure phrase	24
Partitive/quantified/property	12

Partitive/quantified/property

Entity defined relative to complement phrase.

- “members of the union”
- “most Senators”
- “the idea that someone is guilty”

Results

Two different entities	39
Time/measure phrase	24
Partitive/quantified/property	12
Generic	12

Generic

- “In a corporate campaign, a union tries to...”
- “Everyone coming in goes through the drug test”

Results

Two different entities	39
Time/measure phrase	24
Partitive/quantified/property	12
Generic	12
Annotator error	9

Annotator error

Just what it sounds like.

Results

Two different entities	39
Time/measure phrase	24
Partitive/quantified/property	12
Generic	12
Annotator error	9
Proper name	4

Proper names

- “Inc.” and “Co.”
- Pretty well-understood (cf (Ng ‘08), (Stoyanov ‘09))

What knowledge can help us?

- Notion of “compatible” modifiers
  - As in (Elsner+al ‘09) for named entities
- Lexical heads of time/measure/partitive
- Syntactic environment
  - Emphatic discourse position? (Grosz+al)
  - Phrase modifiers?
  - Has complement phrase?
  - Generics: determiner, aspect of governing verb (Gelman)

Overview

Introduction

Mention detection and scoring matter

Non-coreferent same-head pairs

Conversational speech is different

Modeling

Data: Switchboard corpus

Annotated for coreference (Calhoun+al '09), (Nissim '04)

	Linked	Correct
Oracle	454	454
Link all	**2281**	487

Disfluency markup causes annotation errors, but same-head is still a huge problem.

Hand-labeled pairs from SWBD

Two different entities	17	⇓
Time/measure phrase	7	⇓
Partitive/quantified/property	19	⇑
Generic	12	
Annotator error/unmarked	21	⇑
Proper name	0	⇓
Indefinite	9	⇑
Abstract	14	⇑
Q/A	1	

- Lots of errors!
- Less time/measure
- More partitive/quantified
- A few new types...

Indefinites

Mostly “Something”, “everything”, “things”

Abstract NPs

“What happened to pollution?”

Question-Answer

“Do you have a big family?”
“I have kind of a big family”

Overview

Introduction

Mention detection and scoring matter

Non-coreferent same-head pairs

Conversational speech is different

Modeling

Starting point: machine translation

IBM model 2

Generate German from English:

- *Align:* pick a random English word to translate.
- *Translate:* pick an appropriate German word.

English:	He	can	sing	well
German:	Er	kann	gut	singen

Our generative setting

- “Translate” the context into an anaphor...
  - Via a hidden alignment.

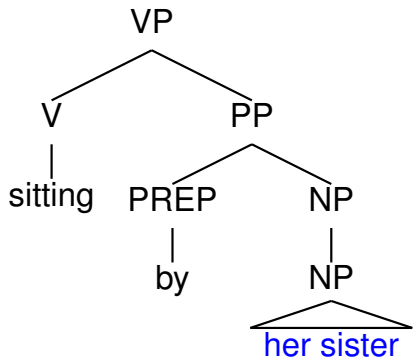
Source text:	Alice	sitting by	her sister	..other NPs..
Target text:		the book	TARGET	was reading
Generated:			her sister	

Generative process

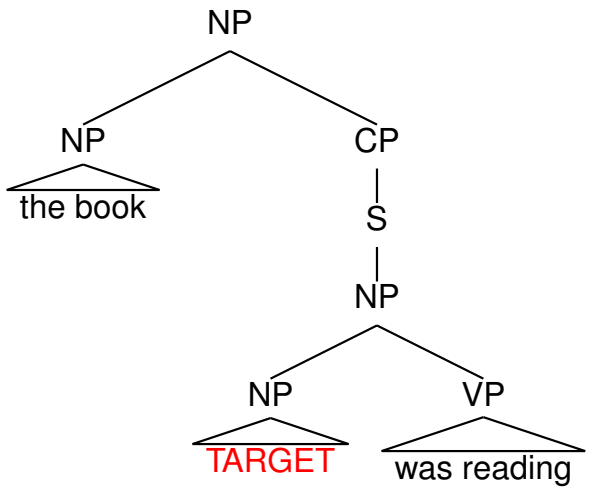
- Input: available NPs, syntactic skeleton around next NP
- Will the next NP corefer with an antecedent? (Alice)
  - Pick an antecedent from the alignment
  - And generate an NP with the same head
  - ...or pick the null antecedent
  - And generate an NP with a random head
- Or will the next NP corefer with nothing? (five minutes)
  - Pick an antecedent uniformally at random
  - And generate an NP with the same head

Modeling alignment

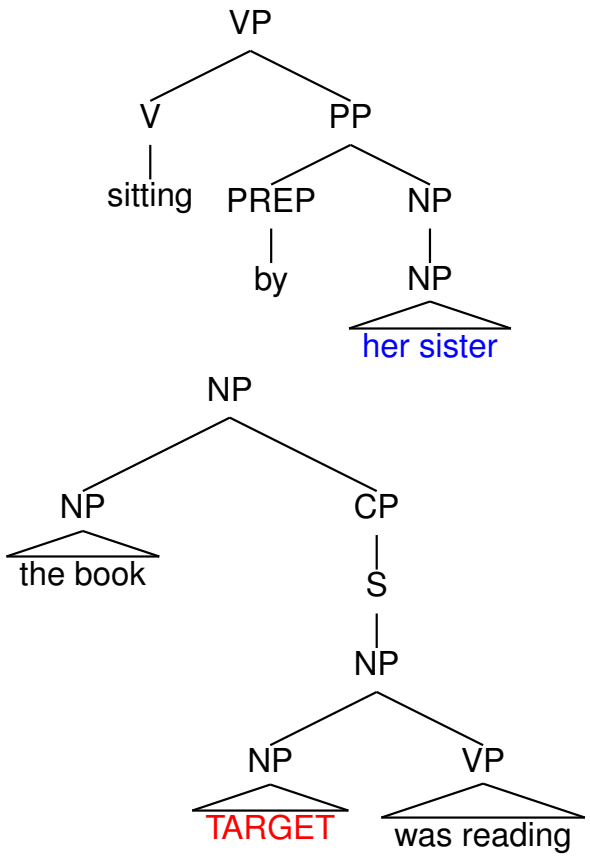
Input to the alignment function:
A possible antecedent:



The slot for the new NP:



Features



- syntactic roles
  (ante: oblique, target: subj)
- positions in sentence
  (between words 5-10)
- proximity in document
  (same sentence)
- proximity in sentence
  (over 10 words apart)
- antecedent phrase type
  (non-proper nominal)
- antecedent determiner (possessive)
- antecedent modifiers (none)

Learning

- Generative model; estimated by EM
- Mixture weight between coreferent and not: set by hand
- Alignment function: log-linear
  - Allows arbitrary features
  - Requires gradient optimization in M-step
  - Or batch updates (as in (Liang+Klein '09))

Initialize parameters for NPs at parameters for pronouns.
Similar preference for NPs likely to refer.

Results

	Mentions	Linked	b^3 pr	rec	F
		NPs			
Oracle	3993	864	100	30.6	46.9
Alignment	3993	518	**87.2**	**24.7**	38.5
Link all	3993	1592	67.2	29.5	41.0

- Precision is up; recall is down.

More results

	Mentions	Linked	Mention CEAF
	NPs		
Oracle	3993	864	73.4
Alignment	3993	518	**67.0**
Link all	3993	1592	62.2

- Overlap of clusterings improves.

SWBD results

	Linked	Correct
Oracle	454	454
Alignment	1168	283
Link all	2281	487

- Favorable precision-recall tradeoff
- But still proposing too many links
- And missing many legitimate ones

Conclusions from analysis

- Experimental setup matters:
  - Use realistic mention detector
  - Report multiple measures
- Domain matters:
  - In conversation, same-head *is* the important case

Conclusions about model

The model is weak.
Future work:

- Translation component that produces modifiers
- Lexicalization

Acknowledgements

- Funded by a Google Fellowship for NLP
- Discussed with the BLLIP group
- Thanks to Jean Carletta for the annotated Switchboard
  - And Dan Jurafsky for telling me about it
- ...and thanks to all of you!