Reference Patterns for Discourse Coherence

Thesis Proposal

Micha Elsner

Department of Computer Science
Brown University

May 6, 2010

Getting information: then

The key problem in information retrieval used to be bandwidth.
The solution looked like this:

Bandwidth is no longer the problem

www.inkycircus.com

Getting information: now

Dealing with too much data requires:

- **Searching** for relevant documents
- **Extracting** what you need
- **Summarizing** the good stuff
- **Updating** when something new happens

Getting information: now

Dealing with too much data requires:

- **Searching** for relevant documents
- **Extracting** what you need
- **Summarizing** the good stuff
- **Updating** when something new happens

Computer assistance is critical!

Automatic search widely available...

Google™

- But we still do the rest by hand.

Text at document scale

To read and write **whole documents...**

- we need to understand document structure.

Lots of NLP work on sentences

- Is this sentence grammatical?
- What is its parse tree?

Questions on documents are fuzzier, but still important

- Is this document coherent?
- What is it about?

What to model: document coherence

Coherence

Structure by which a document presents information–
So readers get context they need to understand new points

What to model: document coherence

Coherence

Structure by which a document presents information–
So readers get context they need to understand new points

Gross structural violations:

Coherent:
A scientist gave a lecture on astronomy.
Afterwards, a woman approached him.

Incoherent:
Afterwards, a woman approached him.
A scientist gave a lecture on astronomy.

What to model: document coherence

Coherence

Structure by which a document presents information–
So readers get context they need to understand new points

Stylistic preferences:

Preferred:

In a change of policy, the US will attend nuclear talks with Iran in Geneva.
Previously, the US did not participate in such meetings with Iran.

Dispreferred:

In a departure from the usual US isolation policy of Iran, US diplomats will attend nuclear talks with Iran in Geneva.

Using coherence

Want to build models that:

- Discover this hidden structure
- Use it to evaluate document quality

Such a model will help:

Using coherence

Want to build models that:

- Discover this hidden structure
- Use it to evaluate document quality

Such a model will help:

- Extraction
  - Find where topic shifts occur
  - Pull out complete topical segments

Using coherence

Want to build models that:

- Discover this hidden structure
- Use it to evaluate document quality

Such a model will help:

- Summary
  - Search for coherent order for sentences in summary
  - Rewrite sentences to improve coherence score

Using coherence

Want to build models that:

- Discover this hidden structure
- Use it to evaluate document quality

Such a model will help:

- Updating
    - Search for where to insert new content

In this proposal:

Why work on documents?

Entity-based models of documents

Getting information from referring expressions
- Named entities

Using the information in modeling
- Extending the entity grid

Applying the model to editing
- Preliminary work

Overview:

Why work on documents?

Entity-based models of documents

Getting information from referring expressions

- Named entities

Using the information in modeling

- Extending the entity grid

Applying the model to editing

- Preliminary work

What are documents about?

Entities!

- Things in the world...
- Like *Hillary Clinton* or *Seattle*.
- Or like *Santa Claus* or *Christianity*.

Entity: *Hillary Clinton*

Text (abridged from Wikipedia):
Clinton was elected as a U.S. Senator in 2000. In the Senate, **she** opposed the administration on its conduct of the war in Iraq. **Senator Clinton** was reelected by a wide margin in 2006. In the 2008 presidential nomination race, **Hillary Clinton** won more primaries and delegates than any other female candidate in American history...

How we talk about entities

Referring expression

- Sometimes called a *mention*
- A piece of language that points out an entity (the *referent*)
- Two expressions with the same referent are *coreferent*
- Usually a *noun phrase*

Two properties we care about:

- The form of the expression
  - **Clinton** vs **Hillary Rodham Clinton**
- Where in the sentence it appears
  - **Clinton** *was elected* vs
    the voters elected **Clinton**

Reference patterns for discourse coherence

Thesis statement

Examining the forms of referring expressions can improve the performance of discourse coherence models on real and artificial tasks.

Forms are underused in previous work

- Linguistics: Centering Theory (Grosz+Sidner+al)
  - A set of constraints on how and where in the sentence entities can occur
  - Based on transitions in the way an entity is used between sentences
- Direct computational models of Centering: (Karamanis), (Tetreault), and others...
  - Rules can be vague or overly restrictive in practice
- Our baseline model: the Entity Grid (Lapata+Barzilay)
  - Statistical model using Centering transitions as features

Issue: most models concentrate on positions, not forms, of referring expressions.

Referring expression forms: informative

Looking at *how* a text refers to an entity can be useful:

- Salient (current topic):

  She *opposed the administration*

  vs non-salient:

  Hillary Rodham Clinton *is the Secretary of State*
- Unfamiliar:

  Clinton gave birth to **a daughter, Chelsea**

  vs familiar:

  Chelsea *attended Stanford University*

Referring expression forms: ambiguous

Without the form, we don't know *which* expressions point to which entities.

- Key issue for pronouns and deictics: **she**, **that**
- Some ambiguity even in easy cases:

  As wife of **President Bill Clinton**, *she was First Lady until 2001.*

  In 2008, **Clinton** *ran for president.*

  Clinton: Hillary or Bill?

Reference patterns for discourse coherence

Thesis statement

Examining the forms of referring expressions can improve the performance of discourse coherence models on real and artificial tasks.

In this talk:

Reference patterns for discourse coherence

Thesis statement

Examining the forms of referring expressions can improve the performance of discourse coherence models on real and artificial tasks.

In this talk:

- Learning properties of referring expressions
  - In this talk: Named entity type

Reference patterns for discourse coherence

Thesis statement

Examining the forms of referring expressions can improve the performance of discourse coherence models on real and artificial tasks.

In this talk:

- Learning properties of referring expressions
  - In this talk: Named entity type
- Modeling the coherence of documents
  - In this talk: Extending the entity grid

Reference patterns for discourse coherence

Thesis statement

Examining the forms of referring expressions can improve the performance of discourse coherence models on real and artificial tasks.

In this talk:

- Learning properties of referring expressions
  - In this talk: Named entity type
- Modeling the coherence of documents
  - In this talk: Extending the entity grid
- Novel applications
  - In this talk: Learning to edit

Overview

Why work on documents?

Entity-based models of documents

Getting information from referring expressions

- Named entities

Using the information in modeling

- Extending the entity grid

Applying the model to editing

- Preliminary work

Different kinds of entity

Not all entities are equal.

- We expect **Hillary Clinton** to behave differently from **a bill**, **two hundred dollars** or **Dixville Notch, NH**
- Documents are often about **people** or **organizations** (Nenkova)...
- Less often about **places** or **amounts of money**

Given a set of strings, can we cluster them by type of entity?

Different kinds of entity

Not all entities are equal.

- We expect **Hillary Clinton** to behave differently from **a bill**, **two hundred dollars** or **Dixville Notch, NH**
- Documents are often about **people** or **organizations** (Nenkova)...
- Less often about **places** or **amounts of money**

Given a set of strings, can we cluster them by type of entity?

Named entity recognition

A standard NLP task
Only look at *proper noun phrases*...
Group entities into three classes: **PERSON**, **ORGANIZATION**, **LOCATION**

Named Entity Structure

People

	Micha		Elsner
Prof.	Eugene		Charniak
Prof.	Mark	E.	Johnson

Organizations

Brown Lab for Linguistic and Information Processing
Brown University

Places

Providence RI

Consistent phrases

Definition: Consistent

Phrases that could refer to the same entity.
Weaker than coreference.

Non-trivial for named entities.
Inconsistent, same heads:

- Ford Motor **Co.**
- Lockheed Martin **Co.**

Consistent, different heads:

- Professor **Johnson**
- **Mark**

Modeling consistency

Model's concept of consistency follows (Charniak '01):

Phrases are consistent if none of their internal subparts clash.

Ordered template	pers^1	pers^2	pers^3	pers^4
	Prof.	Mark	E.	Johnson

Modeling consistency

Model's concept of consistency follows (Charniak '01):

Phrases are consistent if none of their internal subparts clash.

Ordered template	pers^1	pers^2	pers^3	pers^4
	Prof.	Mark	E.	Johnson
realizations		Mark		Johnson

Modeling consistency

Model's concept of consistency follows (Charniak '01):

Phrases are consistent if none of their internal subparts clash.

Ordered template	pers^1	pers^2	pers^3	pers^4
	Prof.	Mark	E.	Johnson
realizations		Mark		Johnson
	Prof.			Johnson

Modeling consistency

Model's concept of consistency follows (Charniak '01):

Phrases are consistent if none of their internal subparts clash.

Ordered template	pers^1 Prof.	pers^2 Mark	pers^3 E.	pers^4 Johnson
realizations		Mark		Johnson
	Prof.			Johnson
		Mark		

Modeling consistency

Model's concept of consistency follows (Charniak '01):

Phrases are consistent if none of their internal subparts clash.

	pers^1	pers^2	pers^3	pers^4
Ordered template	Prof.	Mark	E.	Johnson
realizations		Mark		Johnson
	Prof.			Johnson
		Mark		
inconsistent		~~Mark~~		~~Steedman~~

Named entity recognition: in more detail

- Input:
  - Set of proper noun phrases from a large corpus
  - Plus features from context (explained later!)
  - No labels (unsupervised)

Named entity recognition: in more detail

- Input:
  - Set of proper noun phrases from a large corpus
  - Plus features from context (explained later!)
  - No labels (unsupervised)
- Output:
  - Three clusters of phrases (ideally *person*, *organization*, *location*)
  - Many clusters of words (ideally *first names*, *middle names*, *last names*...)

Named entity recognition: in more detail

- Input:
  - Set of proper noun phrases from a large corpus
  - Plus features from context (explained later!)
  - No labels (unsupervised)
- Output:
  - Three clusters of phrases (ideally *person*, *organization*, *location*)
  - Many clusters of words (ideally *first names*, *middle names*, *last names*...)
- Scoring against a gold standard:
  - MUC corpus labeled by humans
  - Report overlap between our clusters and truth
  - Phrases not in these categories ignored (no gold labels)
  - Word categories unscored (no gold labels)

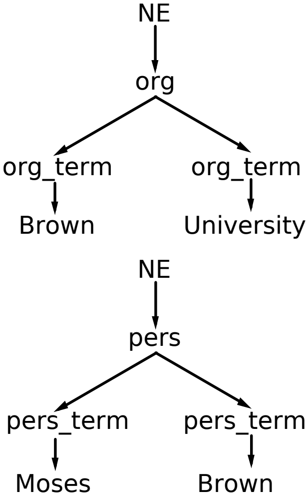
Gathering features

- Nominal modifiers (Collins+Singer '99)
  - Appositive: "Hillary Clinton, the **Secretary** of State
  - Prenominal: "**candidate** Hillary Clinton"
- Prepositional governor (C+S '99)
  - "a **spokesman for** Hillary Clinton"
- Personal pronouns
  - "... Hillary Clinton. **She** said ..."
  - Unsupervised model of pronouns (Charniak+Elsner '09)
- Relative pronouns
  - "Hillary Clinton, **who** said..."

Clustering as parsing

Grammar:

$$\begin{aligned}
\mathit{NE} &\rightarrow \mathit{pers} \\
\mathit{NE} &\rightarrow \mathit{org} \\
\mathit{NE} &\rightarrow \mathit{loc} \\
\mathit{org} &\rightarrow \mathit{org_term}^{+} \\
\mathit{org_term} &\rightarrow \text{Brown} \\
\mathit{org_term} &\rightarrow \text{University} \\
\mathit{pers} &\rightarrow \mathit{pers_term}^{+} \\
\mathit{pers_term} &\rightarrow \text{Moses} \\
\mathit{pers_term} &\rightarrow \text{Brown}
\end{aligned}$$



Internal structure

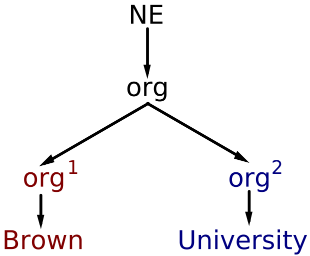
Grammar:

$NE \rightarrow org$

$org \rightarrow org^1\, org^2$

$org^1 \rightarrow$ Brown

$org^2 \rightarrow$ University



Internal structure

Grammar:

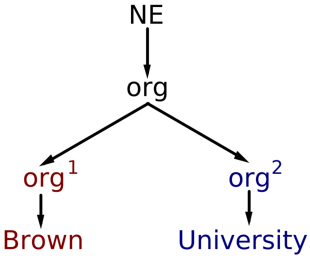
$$NE \rightarrow org$$

$$org \rightarrow org^1\, org^2$$

$$org \rightarrow (org^1)(org^2)(org^3)(org^4)(org^5)$$

$$org^1 \rightarrow \text{Brown}$$

$$org^2 \rightarrow \text{University}$$



Multiword expansions

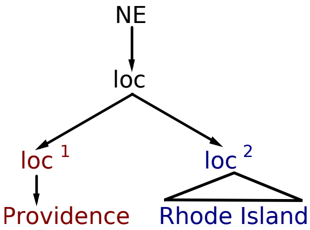
Grammar:

$$NE \rightarrow loc$$

$$place \rightarrow loc^1\, loc^2$$

$$loc^1 \rightarrow \text{Providence}$$

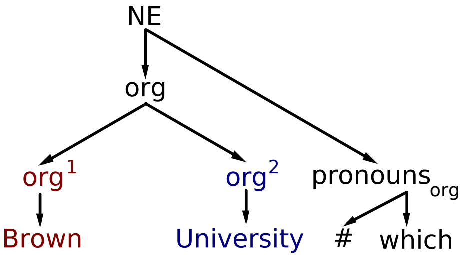
$$loc^2 \rightarrow \textbf{Rhode Island}$$



Adding features

Grammar:

$$
\begin{array}{lcl}
\mathit{NE} & \rightarrow & \mathit{org}\ \mathit{pronouns}_{org} \\
\mathit{org} & \rightarrow & \color{darkred}{\mathit{org}^1}\ \color{darkblue}{\mathit{org}^2} \\
\mathit{pronouns}_{org} & \rightarrow & \#\ \mathit{pronoun}_{org}{}^{*} \\
\mathit{pronoun}_{org} & \rightarrow & \mathit{which} \\
\mathit{pronoun}_{org} & \rightarrow & \mathit{they} \\
\ldots & & \\
\mathit{pronoun}_{org} & \rightarrow & \mathit{he} \\
\ldots & &
\end{array}
$$



Learning the grammar

How to learn rule probabilities?

- Many, many rules:
  - With multiword strings, infinite!
- Most of them useless.

Bayesian model

Sparse prior over rules.
Posterior concentrated around few useful rules.

Adaptor grammars (Johnson+al '07)

- Prior over grammars
- Form of hierarchical *Dirichlet process*
- Black-box inference, downloadable software
  - Development is just writing the grammar
- But standard inference isn't always good enough

Basic results

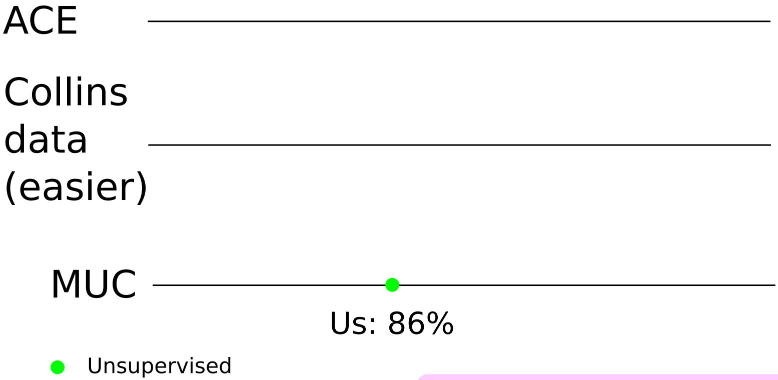
Our model:

Baseline (all ORG):	42%
Our best model:	**86%**

Confusion matrix:

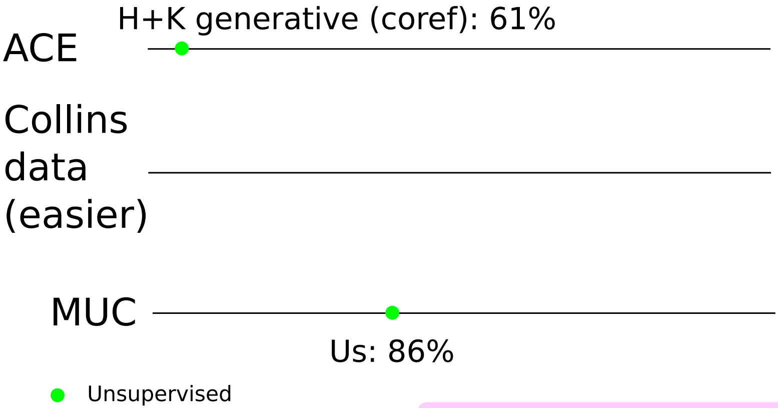
		True label		
		Loc	*Org*	*Per*
Our label	*Loc*	1187	97	37
	Org	223	1517	122
	Per	36	20	820

Comparison



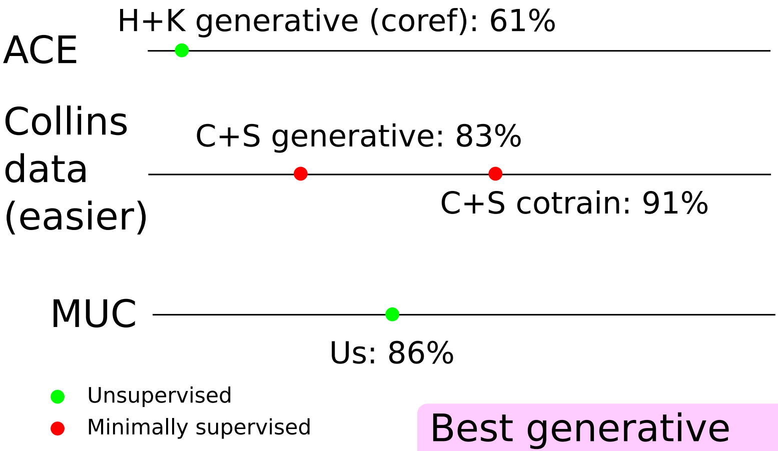
Best generative
Best unsupervised

Comparison



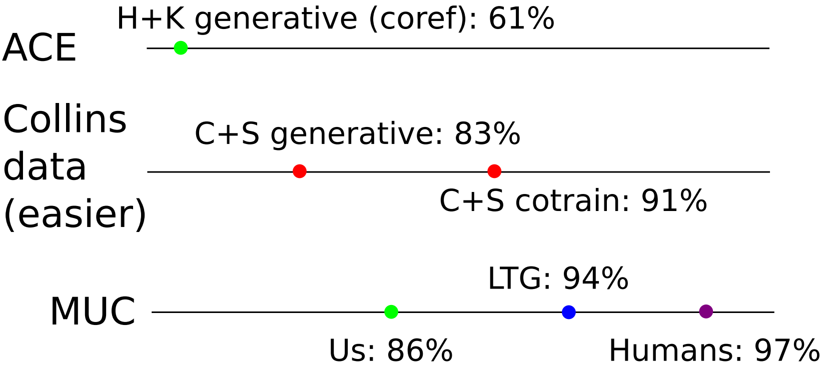
Best generative
Best unsupervised

Comparison



Best generative
Best unsupervised

Comparison



- Unsupervised
- Minimally supervised
- Supervised
- Annotator agreement

Best generative
Best unsupervised

Named entity structure

pers^0	*pers*^{1}	*pers*^{2}	*pers*^{3}	*pers*^{4}
rep.	john	minister	brown	jr.
sen.	robert	j.	smith	a
washington	david	john	b	smith
dr.	michael	l.	johnson	iii

loc^0	*loc*^{1}	*loc*^{2}	*loc*^{3}	*loc*^{4}
washington	the	texas	county	monday
los angeles	st.	new york	city	thursday
south	new	washington	beach	river
north	national	united states	valley	tuesday

Overview

Entity grids: the baseline

Model of transitions from sentence to sentence (Lapata+Barzilay,Barzilay+Lapata):

Text	Syntactic role
Suddenly **a White Rabbit** ran by her.	**subject**
Alice heard **the Rabbit** say “I shall be late!”	**object**
The Rabbit took a watch out of its pocket.	**subject**
Alice started to her feet.	**missing**

Entity grids: the baseline

Model of transitions from sentence to sentence (Lapata+Barzilay,Barzilay+Lapata):

Text	Syntactic role
Suddenly **a White Rabbit** ran by her.	**subject**
Alice heard **the Rabbit** say “I shall be late!”	**object**
The Rabbit took a watch out of its pocket.	**subject**
Alice started to her feet.	**missing**

Treat as a Markov chain:

$$P(subj|\text{<s>})P(obj|subj)P(subj|obj)P(miss|subj)$$

All entities independent.

Can we use what we learned before?

Why should we expect:

$$P(\textbf{Hillary Clinton} = \textit{subj}|\textit{subj}) = P(\textbf{ten minutes} = \textit{subj}|\textit{subj})$$

Can we use what we learned before?

Why should we expect:

$$P(\textbf{Hillary Clinton} = \textit{subj}|\textit{subj}) = P(\textbf{ten minutes} = \textit{subj}|\textit{subj})$$

We know that entities have:

- Different named entity type
- Different number/gender/affinity for pronouns
- Preference to corefer/not corefer with similar phrases

Let's use this information!

Modeling

Let's use a log-linear model to learn:

$P($**Hillary Clinton** $=$ *subj*$|$

previous role was *subj*

(actually, two previous roles)

occurs 3 times

type is person

singular

high affinity for pronouns

probably corefers with **Hillary Clinton**$)$

Results

Discrimination task

Binary classification: tell an **original document** (assumed coherent) from a **randomly permuted document** (assumed incoherent).

Results

Discrimination task

Binary classification: tell an **original document** (assumed coherent) from a **randomly permuted document** (assumed incoherent).

Discrimination on Wall Street Journal:

Random	50.00
Entity Grid	74.41
Entity Grid + Type Features	**80.27**

Results

Discrimination task

Binary classification: tell an **original document** (assumed coherent) from a **randomly permuted document** (assumed incoherent).

Discrimination on Wall Street Journal:

Random	50.00
Entity Grid	74.41
Entity Grid + Type Features	**80.27**

Can we do better?

Multiple kinds of entity– multiple generative processes?
Incorporate topic variables to predict some entity types?

Overview

Why work on documents?

Entity-based models of documents

Getting information from referring expressions
- Named entities

Using the information in modeling
- Extending the entity grid

Applying the model to editing
- Preliminary work

The data

500 article pairs processed by professional editors:

Novel dataset courtesy of Thomson Reuters

THOMSON REUTERS

Journalist wrote:
Opponents of gay marriage then placed their hopes on an initiative, called Proposition 8, that would limit weddings to opposite sex couples.

Editor altered:
Opponents of gay marriage then placed an initiative to amend the constitution on the November ballot.
"Proposition 8" declares that marriage will be limited to one man with one woman.

Why study editing?

So far, results on discrimination:

- Assume all human-authored documents equally coherent
- Manufacture fake incoherent documents

Can we measure the relative coherence of real documents?

Previous methods:

- Standardized testing essays
- Paid annotators
- Grade level assessments

Editing

Can discover what editors think about coherence...
by what they change

What editors do

What edits occur?

Input sentences	9007	100%
Datelines	513	5.7%
Unchanged	4815	53.5%
Edited inline	2999	33.3%
Deleted	509	5.7%
Split	175	1.9%
Merged	132	1.5%
Inserted	433	4.8%
Output sentences	8974	99.6%

We predict that input sentences editors choose to **alter** are harder to read than those they leave **unchanged**...
Supporting our claim that editing improves coherence.

Features

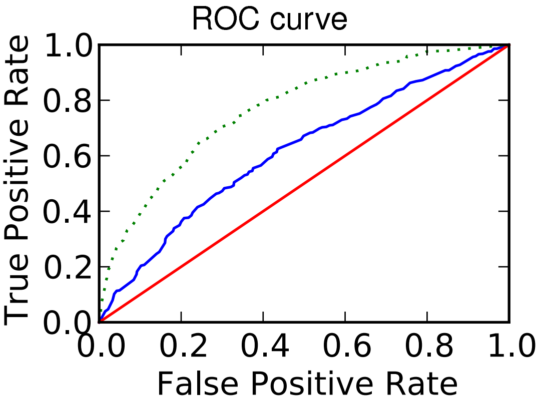
Features previously used in readability prediction (mostly (Chae+Nenkova))
To predict whether a sentence will be edited.

Features with significant information gain:

How many *NPs* had modifiers?
What fraction of words in *NP, VP, PP*?
How many words?
Sentence is a quote?
Where in document is sentence?

Can't extract from raw input:
Were nearby sentences edited?

Predictions more accurate than chance



Red: chance
Blue: practically useable features
Green: +nearby sentence features

Readability features do predict edits

Proposed work

Editors improve coherence...
Can our models predict high-level changes they make?

- Reordered sentences?
  - Like artificial ordering tasks, but realistic source of negative examples!
- Sentence splits and merges?
  - Similar to sentence fusion, (Filipova+Strube),(Barzilay+McKeown)

Conclusion

Thesis statement

Examining the forms of referring expressions can improve the performance of discourse coherence models on real and artificial tasks.

In support of this thesis, we describe work on:

- Linking referring expressions to entities
  - (Elsner+Charniak ACL '10), (Charniak+Elsner EACL '09)
- Distinguishing types of referring expressions
  - (Elsner+Charniak+Johnson NAACL '09)
- Modeling the coherence of documents
  - (Elsner+Austerweil+Charniak NAACL '07), (Elsner+Charniak ACL '08)
- Novel applications
  - (Elsner+Charniak Journal of CL (to appear)), (Elsner+Schudy ILP-NLP '09), (Elsner+Charniak ACL '08)

Acknowledgements

Brought to you by:

- Eugene Charniak, Mark Johnson, Regina Barzilay
- the BLLIP lab, past and present
- Thomson Reuters (and my father, Alan Elsner!)
- everyone who sat through the practice talks
- NSF PIRE, DARPA GALE, the Google Fellowship
- ...and viewers like you!

Overview

Learning about pronouns

Adaptor grammars: framework for Bayesian grammar learning

Implementing Consistency

Inference: a general problem for this approach

Overview

Learning about pronouns

Adaptor grammars: framework for Bayesian grammar learning

Implementing Consistency

Inference: a general problem for this approach

Motivation

The White Queen looked timidly at Alice, who felt she ought to say something kind, but really couldn't think of anything at the moment.

- Pronouns are potentially ambiguous.
- Does she mean Alice, or the White Queen?
- Technically could be either, but strong intuitions.

Pronoun resolution

Pronouns

He, *she*, *it*, *they*...

Most have an *antecedent*:

- Coreferent phrase
- Not a pronoun (we assume an NP)
- Occurring earlier in the text

Pronoun resolution

Pronouns

He, *she*, *it*, *they*...
Most have an *antecedent*:

- Coreferent phrase
- Not a pronoun (we assume an NP)
- Occurring earlier in the text

Our task: learn to link pronouns to their antecedents.
Training data is expensive... do this *without supervision*.

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 a pronoun...
  - Via a hidden alignment.
  - And a hidden translation model

Source: `The White Queen looked at Alice who...`

Target: `...felt she ought to...`

Translation parameters

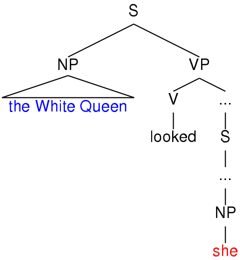
For each word, need to learn:

- Singular or plural?
- Masculine, feminine, or neuter?

Some results:

	Masc	**Fem**	**Neut**
Paul	**.96**	.002	.035
Paula	.003	**.915**	.082
pig	**.445**	.170	.385
piggy	.001	**.853**	.146
wal-mart	.016	.007	**.976**
waist	.380	.155	**.465**

Alignment features



- syntactic role: subject
- position: beginning of sentence
- proximity: same sentence
- within-sentence proximity: 6 words away
- phrase type: proper noun phrase
- determiner: “the”
- head word: “Queen”

Learning

Learn using EM algorithm:

- Finds a local maximum of likelihood

Key insight: some pronouns very unambiguous...

- Like very beginning of article:
  Senator Hillary Clinton *announced that* ***she****...*
- Model learns these quickly...
  - Which improves more difficult cases

Somewhat surprising that EM/Max-likelihood works...
Many NLP cases where it doesn't.

Results

Metric *roughly* percent of pronouns attached to a correct antecedent.
Dataset: Hand-annotated news text (Ge+al), 1119 pronouns/

Performance: 68.6% pronouns correct
Best publically available system: 59.3%

Comparable results described in:

- (Cherry+Bergsma)
- (Kehler+al)
- (No released software, so no direct comparisons)

Modeling document coherence

Can we tell a coherent document from an incoherent one... by looking at how they use pronouns?

Discrimination task

Binary classification: tell an **original document** (assumed coherent) from a **randomly permuted document** (assumed incoherent) (Lapata+Barzilay).

Modeling document coherence

Can we tell a coherent document from an incoherent one...
by looking at how they use pronouns?

Discrimination task

Binary classification: tell an **original document** (assumed coherent) from a **randomly permuted document** (assumed incoherent) (Lapata+Barzilay).

Results on Wall Street Journal:

Random	50.00
Entity Grid	74.41
Pronouns	64.41
Entity Grid + Pronouns	76.83

Adaptor grammars (Johnson+al '07)

- A prior over grammars
- Some nonterms are *Dirichlet processes* over subtrees
  - Previously used expansions gain probability
- Black-box inference, downloadable software
  - Development is just writing the grammar
- But standard inference isn't always good enough
  - More on this later...

Adaptor grammars (Johnson+al '07)

Data:

Prior grammar:

count	*rule*	
1	*words* $\rightarrow$	*word words*
1	*words* $\rightarrow$	*word*
1	*word* $\rightarrow$	Rhode
1	*word* $\rightarrow$	Island
1	*word* $\rightarrow$	Colorado
...		
1	$\underline{loc^2}$ $\rightarrow$	*words*

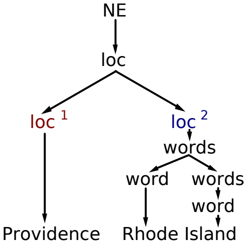
Providence Rhode Island

Boulder Colorado

Newport Rhode Island

Adaptor grammars (Johnson+al '07)

Data:



Boulder Colorado

Newport Rhode Island

Posterior grammar:

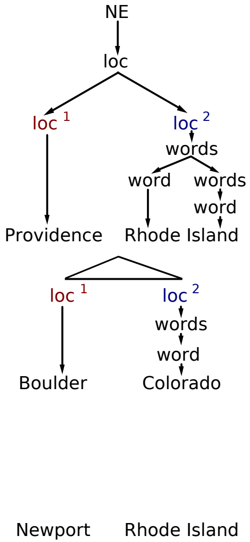
count	*rule*	
2	*words* $\rightarrow$	*word words*
2	*words* $\rightarrow$	*word*
2	*word* $\rightarrow$	Rhode
2	*word* $\rightarrow$	Island
1	*word* $\rightarrow$	Colorado
...		
1	$\underline{loc^2}$ $\rightarrow$	*words*
1	$\underline{loc^2}$ $\rightarrow$	Rhode Island

Adaptor grammars (Johnson+al '07)

Posterior grammar:

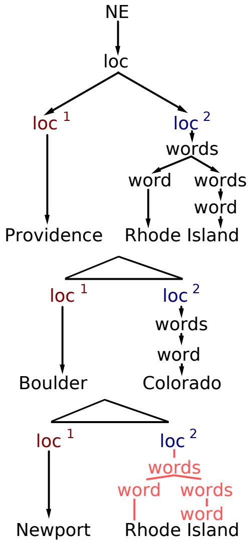
count	*rule*	
2	*words* $\rightarrow$	*word words*
3	*words* $\rightarrow$	*word*
2	*word* $\rightarrow$	Rhode
2	*word* $\rightarrow$	Island
2	*word* $\rightarrow$	Colorado
...		
1	$\underline{loc^2}$ $\rightarrow$	*words*
1	$\underline{loc^2}$ $\rightarrow$	Rhode Island
1	$\underline{loc^2}$ $\rightarrow$	Colorado

Data:



Adaptor grammars (Johnson+al '07)

Data:



Posterior grammar:

count	*rule*	
2	*words* $\rightarrow$	*word words*
3	*words* $\rightarrow$	*word*
2	*word* $\rightarrow$	Rhode
2	*word* $\rightarrow$	Island
2	*word* $\rightarrow$	Colorado
...		
1	$\underline{loc^2}$ $\rightarrow$	*words*
2	$\underline{loc^2}$ $\rightarrow$	Rhode Island
1	$\underline{loc^2}$ $\rightarrow$	Colorado

Overview

Learning about pronouns

Adaptor grammars: framework for Bayesian grammar learning

Implementing Consistency

Inference: a general problem for this approach

Implementing consistency

Grammar:

$$NE \rightarrow org$$

$$org \rightarrow org_{Brown} \ldots$$

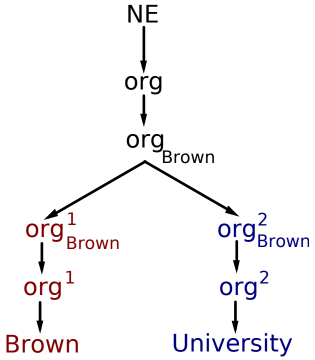
$$org_{Brown} \rightarrow org^1_{Brown} \; org^2_{Brown}$$

$$\underline{org^1_{Brown}} \rightarrow org^1$$

$$\underline{org^2_{Brown}} \rightarrow org^2$$

$$\underline{org^1} \rightarrow \text{Brown}$$

$$\underline{org^2} \rightarrow \text{University}$$



Underlined nonterminals are Dirichlet processes.

$\underline{org^1_{Brown}}$ and $\underline{org^2_{Brown}}$ get only one expansion.

Yet another infinity

How many entities (like org_{Brown}) are there?

- Grows with the data size...
- Again, use Bayesian methods.

Allow an infinite number...

and constrain with a sparse prior.

Simple in principle (special case of “Infinite PCFG”, Liang+al ‘07)
Requires some code changes.

Overview

Learning about pronouns

Adaptor grammars: framework for Bayesian grammar learning

Implementing Consistency

Inference: a general problem for this approach

Basic inference by sampling

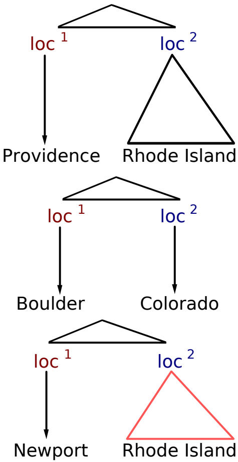
Gibbs sampling:

- Start with arbitrary trees
- Repeat forever
  - Erase a random tree
  - Sample a tree from the current grammar
  - Update the grammar given the new tree

Rules for $\underline{loc}^2$:

1	$\underline{loc}^2$	$\rightarrow$	*words*
1	$\underline{loc}^2$	$\rightarrow$	Colorado
2	$\underline{loc}^2$	$\rightarrow$	Rhode Island

Data:



Basic inference by sampling

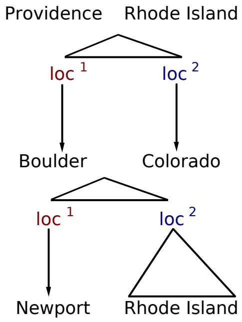
Gibbs sampling:

- Start with arbitrary trees
- Repeat forever
  - Erase a random tree
  - Sample a tree from the current grammar
  - Update the grammar given the new tree

Rules for $\underline{loc}^2$:

1	$\underline{loc}^2$	$\rightarrow$	*words*
1	$\underline{loc}^2$	$\rightarrow$	Colorado
1	$\underline{loc}^2$	$\rightarrow$	Rhode Island

Data:



Basic inference by sampling

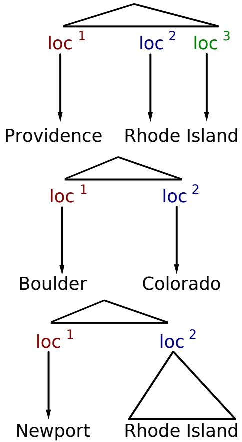
Gibbs sampling:

- Start with arbitrary trees
- Repeat forever
  - Erase a random tree
  - Sample a tree from the current grammar
  - Update the grammar given the new tree

Rules for $\underline{loc^2}$:

1	$\underline{loc^2}$	$\rightarrow$	*words*
1	$\underline{loc^2}$	$\rightarrow$	Colorado
1	$\underline{loc^2}$	$\rightarrow$	Rhode Island

Data:



Basic inference by sampling

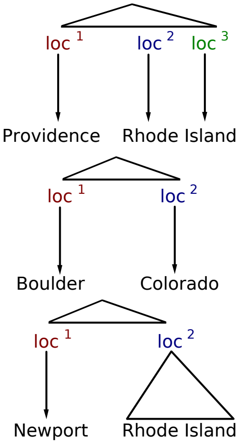
Gibbs sampling:

- Start with arbitrary trees
- Repeat forever
  - Erase a random tree
  - Sample a tree from the current grammar
  - Update the grammar given the new tree

Rules for $\underline{loc^2}$:

1	$\underline{loc^2}$	$\rightarrow$	*words*
1	$\underline{loc^2}$	$\rightarrow$	Colorado
1	$\underline{loc^2}$	$\rightarrow$	Rhode Island
1	$\underline{loc^2}$	$\rightarrow$	Rhode

Data:



Basic inference by sampling

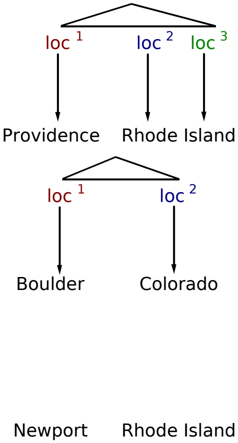
Gibbs sampling:

- Start with arbitrary trees
- Repeat forever
  - Erase a random tree
  - Sample a tree from the current grammar
  - Update the grammar given the new tree

Rules for $\underline{loc}^2$:

1	$\underline{loc}^2$	$\rightarrow$	*words*
1	$\underline{loc}^2$	$\rightarrow$	Colorado
1	$\underline{loc}^2$	$\rightarrow$	Rhode Island
1	$\underline{loc}^2$	$\rightarrow$	Rhode

Data:



Basic inference by sampling

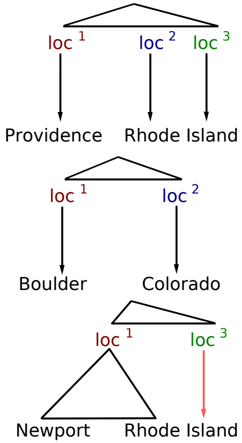
Gibbs sampling:

- Start with arbitrary trees
- Repeat forever
  - Erase a random tree
  - Sample a tree from the current grammar
  - Update the grammar given the new tree

Rules for $\underline{loc}^2$:

1	$\underline{loc}^2$	$\rightarrow$	*words*
1	$\underline{loc}^2$	$\rightarrow$	Colorado
1	$\underline{loc}^2$	$\rightarrow$	Rhode

Data:



Basic inference by sampling

Gibbs sampling:

- Start with arbitrary trees
- Repeat forever
  - Erase a random tree
  - Sample a tree from the current grammar
  - Update the grammar given the new tree

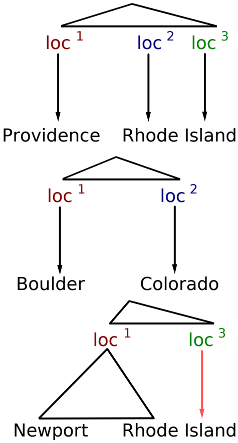
Rules for $\underline{loc^2}$:

1 $\underline{loc^2} \rightarrow$ *words*

1 $\underline{loc^2} \rightarrow$ Colorado

1 $\underline{loc^2} \rightarrow$ Rhode

Data:



Issue 1: efficiency

Sampling a new parse

- Via CKY algorithm: $O(n^3)$
  - ... times a grammar constant!
- One set of nonterminals for each entity
- Scales poorly

Can be dealt with (Metropolis-Hastings algorithm):

- Proposal distribution:
  - Easy-to-calculate approximation to the grammar
- Worse approximations, slower runtimes.

Issue 2: mobility

Local maxima are still a problem

- Gibbs sampling converges in the limit...
- Not in real life!
- What you'd expect – clustering is often NP-hard

- Resampling one tree at a time means lots of local maxima
- Better moves:
  - Split and merge entities
  - Reparse multiple strings at once
- Tricky to implement...
- Correct algorithms can be very slow in practice

Compromise: heuristic inference

What we actually do:

- Propose only a subset of entities for each string:
  - Must have at least one word in common
  - Less likely if shared word is frequent
- *Ignore* the Hastings correction term!

Not theoretically valid, but faster.

- Even so, inference remains a problem.
  - Too many clusters for the same entity

Judging consistency

Sometimes right:

- Dr. Seuss
- Dr. Quinn

... correctly judged inconsistent.

Judging consistency

Sometimes right:

- Dr. Seuss
- Dr. Quinn

... correctly judged inconsistent.

Sometimes wrong:

- Dr. William F. Gibson
- Dr. William Gibson

... judged inconsistent.

- Bruce Jarvis
- Ellen Jarvis

... judged consistent.