Discourse Analysis

A statistical approach to coreference resolution of noun phrases

Abhijit Mishra (114056002)
Samir Janardan Sohoni (114086001)
What is 'Discourse'? 

- "A mode of organizing knowledge, ideas or experience that is rooted in language and its concrete contexts" - Meriam Webster Dictionary 

- "A continuous stretch of (especially spoken) language larger than a sentence, often constituting a coherent unit such as a sermon, argument, joke or narrative" - Crystal (1992:25)
What is 'Discourse'?

- "A mode of organizing knowledge, ideas or experience that is rooted in language and its concrete contexts" - Meriam Webster Dictionary

- "A continuous stretch of (especially spoken) language larger than a sentence, often constituting a coherent unit such as a sermon, argument, joke or narrative" - Crystal (1992:25)
What happens in Discourse Analysis?

- Analyze the formal and contextual links within a discourse
- Formal links are built into the language rendering
- Contextual links rely upon world knowledge
Formal Links in language

- **Substitution** – use of one, do, so
  - Tom produced a nice painting. I told you so long ago.

- **Ellipsis** – omission of words or clauses
  - Prosperity is a great teacher; adversity is a greater ?.

- **Conjunction** – addition, temporals and causals
  - The travellers had lunch then they rested because they were tired.

- **References** – pronouns and articles draw meaning from other words/contexts
  - Neighbours bought a new car, it is nice.
  - It is raining/It is day/It is night.
Motivation

- Ability to link coreferring NPs within and across sentences is important
- Coreference resolution is important because
  - MT will need it
  - IE, QA and Summarization systems need it
  - In a natural environment, students learning new language need to understand the phenomenon
A coreference is not always limited to a pronoun like *they, it* etc.

It can be a chain of non-pronominals

- Mahatma Gandhi insisted on non-violent means for freedom. He is a key figure in Indian history. Gandhi is also known as 'father of the nation'.
  - Coreferenced chain = (Mahatma Gandhi)-(He)-(Gandhi)

Can we identify coreferenced chains of noun-phrases?
Why Statistical Approach

- Rules-based approaches take time, money and trained personnel to make and test the rules.
- Coreference resolution is a semantic level task which requires a lot of time and effort.
- Statistical methods may not be highly accurate but save a lot of time and money.
- Availability of monolingual corpus motivates us to try out quick statistical systems.
Glossary

- **Markables** – NPs, nested NPs, pronouns etc. that are identities of reference
  - $(((\text{Bill Gates})_B, (\text{the chairman})_C \text{ of } (\text{Microsoft Corp})_D)_A$

- **MUC** – Message Understanding Conference
  - Initiative to US Gov and depts like DARPA
  - Standardize data to be used by participants
Methodology

1) Training data is standardized corpus having chains of coref-annotated markables

2) From an annotated chain each bigram pair as <antecedent, anaphore> is obtained.

3) Basing on the features possessed by such pairs, a decision tree is learnt.

4) For testing, chains of markables are created from test data.

5) Markers are presented to the classifier and coreference chains are extracted.
Processing Pipeline

Free text → Tokenization & sentence segmentation → Morphological processing → POS tagger → Noun phrase identification

Morphological processing

POS tagger

Feature Extraction

Nested noun phrase extraction

Markables

Classifier

Named Entity Recognition

HMM

HMM

HMM
Features

- Properties of a discourse which help to decide whether two markable corefer or not
- Should be domain independent.
- Should not be too difficult to compute.

For a marker_pair<i,j> we consider 12 different kinds of features.

Consider this example:

Separately, Clinton transition officials said that Frank Newman, 50, Vice chairman and chief financial officer of BankAmerica Corp., is expected to be nominated as assistant Treasury secretary for domestic finance.

Marker_i = ”Frank Newman” and Marker_j = ”Vice chairman”
**Distance Feature** : $f_{\text{dist}}$

- Possible Values: $<\text{Num}> : 0,1,2,3$
- Captures distance between $i$ and $j$
- If $i$ and $j$ are in the same sentence, $f(i,j)=0$. If they are one sentence apart $f(i,j)=1$ and so on. E.g: $f_{\text{dist}}(\text{Frank Newman, Vice chairman})=0$

**I-Pronoun and J-Pronoun** : $f_{i_{\text{pron}}}$, $f_{j_{\text{pron}}}$

- Possible values $<\text{true},\text{false}>$
- If $i$ is a pronoun then $f_{i_{\text{pron}}}(i,j)="\text{true}"
- Similarly if $j$ is a pronoun then $f_{j_{\text{pron}}}(i,j)="\text{true}"
- Pronouns include reflexive $<\text{herself, himself}>$, personal pronouns $<\text{She, her, you}>$ and possessive pronouns $<\text{her, his}>$. 
- E.g: $f_{j_{\text{pron}}}(\text{Frank Newman, Vice chairman})=\text{false}$
- **Definite and Demonstrative NP**: $f_{\text{def}}$, $f_{\text{dem}}$
  - If "j" is a definite NP (e.g. "the car") or demonstrative NP (e.g. "that boy") then return true.
  - E.g: $f_{\text{def\_NP}}(\text{Frank Newman, Vice chairman}) = \text{false}$
- **Number and Gender**: $f_{\text{num}}$ and $f_{\text{gender}}$
  - If i, j agree in number then $f_{\text{num}}(i, j) = \text{true}$
  - If i, j agree in gender then $f_{\text{gender}}(i, j) = \text{true}$
  - E.g: $f_{\text{num}}(\text{Frank Newman, Vice chairman}) = \text{true}$
  - $f_{\text{gender}}$ can take three values < true, false, unknown >
  - Designators and pronouns such as "Mr", "Mrs", "she", "he" are used to determine the gender.
Features (contd..)

- **Both-Proper-Noun**: If both i and j are proper nouns return true.

- **Alias Feature**: If i is an alias of j return true.

- **Appositive Feature**: If ”j” is an apposition to ”i” return true.
  
  E.g : $f_{\text{appositive}} (\text{Frank Newman}, \text{Vice chairman}) = \text{true}$

- **Semantic Class Agreement feature**: $f_{\text{semclass}}$
  
  - Possible values are <true, false, unknown>
  
  - The marker head words are assigned with one the following classes.
    
    < person, organization, location, time, object >
  
  - Semantic class labeling is done by finding out the class label closest to the first sense of the head word in a marker.

  E.g: $f_{\text{semclass}} (\text{Frank Newman}, \text{Vice chairman})=\text{true}$
  
  since both i and j correspond to persons
(Eastern Air)\textsubscript{1} proposes (date)\textsubscript{2} for (talks)\textsubscript{3} on (pay-cut plan)\textsubscript{4}. ((Eastern Airlines)\textsubscript{5} executives)\textsubscript{6} notified ((union)\textsubscript{7} leaders)\textsubscript{8} that (the carrier)\textsubscript{9} wishes to discuss (selective (wage)\textsubscript{10} reductions)\textsubscript{11} on (Feb. 3)\textsubscript{12}. ((Union)\textsubscript{13} representatives)\textsubscript{14} who could be reached said (they)\textsubscript{15} hadn't decided whether (they)\textsubscript{16} would respond. By proposing (a meeting date)\textsubscript{17} (Eastern)\textsubscript{18} moved (one step)\textsubscript{19} closer towards reopening (current high-cost contracts agreements)\textsubscript{20} with ((its)\textsubscript{21} unions)\textsubscript{22}. 

((union)\textsubscript{7} (unions)\textsubscript{13}) and ((union)\textsubscript{13} (its unions)\textsubscript{22})

(the carrier)\textsubscript{9} (union)\textsubscript{13} and ((wage)\textsubscript{10} (union)\textsubscript{13})
C5 decision tree algorithm is used to learn a decision tree from the training data.

It's an updated version of ID3 algorithm in which the feature to be selected is the one which provides maximum information gain.

\[ Gain(S, A) = Entropy(S) - \left( \frac{|A|}{|S|} \right) \times Entropy(A) \]

\[ Entropy(X) = - \sum_{i} \left( Pr(x_i) \times \log(Pr(x_i)) \right) \]

C5 has a better pruning mechanism and also handles training data with missing attribute values.
Algorithm: (Document D, Decision_Tree T) : List

M = get_markers_from_document (D)

for ( j = 2; j<M ; j++):
    for ( i = 1; i<j ; i++ ):
        F = get_feature_vector ( i, j )
        /********Get the class from Decision Tree******/
        corfer = get_corefer ( F , T)
        if (corfer):
            j.antecedent = i

for ( j = M ; j > 1; j-- ):
    chain = back_track ( j )
    List.add ( chain )

return List
(Ms. Washington)’s candidacy is being championed by several lawmakers including (her boss), (chairman John Dingell) (D., (Mich.) of (the House Energy and Commerce Committee). (She currently is (a counsel) to (the committee). (Ms. Washington) and (Mr. Dingell) have been considered allies of (the securities exchanges), while (banks) and (futures exchanges) have often fought with them.
Coreferenced chain is \((\text{Ms. Washington})_{73} \)-(\text{her})_{76}-(\text{She})_{80}\)
# Evaluation

<table>
<thead>
<tr>
<th>System ID</th>
<th>Recall</th>
<th>Prec</th>
<th>F</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Complete systems</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DSO</td>
<td>58.6</td>
<td>67.3</td>
<td>62.6</td>
<td>Our system</td>
</tr>
<tr>
<td>DSO.TRG</td>
<td>52.6</td>
<td>67.6</td>
<td>59.2</td>
<td>Our system using RESOLVE’s method of generating positive and negative examples</td>
</tr>
<tr>
<td>RESOLVE</td>
<td>44.2</td>
<td>50.7</td>
<td>47.2</td>
<td>The RESOLVE coreference system at the University of Massachusetts</td>
</tr>
<tr>
<td><strong>Baseline systems using just one feature</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DIST</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>Only “distance” feature is used</td>
</tr>
<tr>
<td>SEMCLASS</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>Only “semantic class agreement”</td>
</tr>
<tr>
<td>NUMBER</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>Only “number agreement”</td>
</tr>
<tr>
<td>GENDER</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>Only “gender agreement”</td>
</tr>
<tr>
<td>PROPER_NAME</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>Only “both proper names”</td>
</tr>
<tr>
<td>ALIAS</td>
<td>24.5</td>
<td>88.7</td>
<td>38.4</td>
<td>Only “alias”</td>
</tr>
<tr>
<td>J_PRONOUN</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>Only “j-pronoun”</td>
</tr>
<tr>
<td>DEF_NP</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>Only “definite noun phrase”</td>
</tr>
<tr>
<td>DEM_NP</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>Only “demonstrative noun phrase”</td>
</tr>
<tr>
<td>STR_MATCH</td>
<td>45.7</td>
<td>65.6</td>
<td>53.9</td>
<td>Only “string match”</td>
</tr>
<tr>
<td>APPOSITIVE</td>
<td>3.9</td>
<td>57.7</td>
<td>7.3</td>
<td>Only “appositive”</td>
</tr>
<tr>
<td>I_PRONOUN</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>Only “i-pronoun”</td>
</tr>
<tr>
<td><strong>Other baseline systems</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALIAS_STR</td>
<td>51.5</td>
<td>66.4</td>
<td>58.0</td>
<td>Only the “alias” and “string match” features are used</td>
</tr>
<tr>
<td>ALIAS_STR_APPPOS</td>
<td>55.2</td>
<td>66.4</td>
<td>60.3</td>
<td>Only the “alias,” “string match,” and “appositive” features are used</td>
</tr>
<tr>
<td>ONE_CHAIN</td>
<td>89.9</td>
<td>31.8</td>
<td>47.0</td>
<td>All markables form one chain</td>
</tr>
<tr>
<td>ONE_WRD</td>
<td>55.4</td>
<td>36.6</td>
<td>44.1</td>
<td>Markables corefer if there is at least one common word</td>
</tr>
<tr>
<td>HD_WRD</td>
<td>56.4</td>
<td>50.4</td>
<td>53.2</td>
<td>Markables corefer if their head words are the same</td>
</tr>
</tbody>
</table>

Courtesy: Soon, Ng, Lim (2001)
Evaluation (contd..)

Courtesy: Soon, Ng, Lim (2001)
## Precision Errors (false +ve)

<table>
<thead>
<tr>
<th>No</th>
<th>Type of error</th>
<th>Example</th>
<th>Frequency</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Prenominal modifier string match</td>
<td>David Bronczek, (vice) president of ... was named senior (vice) president.</td>
<td>16</td>
<td>42.1</td>
</tr>
<tr>
<td>2</td>
<td>Strings match but NP refer to different things</td>
<td>... the House Energy and Commerce ...(the committee) ...(the committee) ... Senate Finance Committee ...(the committee) ...</td>
<td>11</td>
<td>28.9</td>
</tr>
<tr>
<td>3</td>
<td>Errors in NP identification</td>
<td>”...May and June...” should actually be (May) and (June) ... but it is treated as nouns in apposition</td>
<td>4</td>
<td>10.5</td>
</tr>
<tr>
<td>4</td>
<td>Errors in Apposition Determination</td>
<td>.. Metaphor, a company, that IBM bought in 1991, also named (Chris Grejtak) (_1), (43 year old) (_2), currently a SVP, president &amp; CEO...</td>
<td>5</td>
<td>13.2</td>
</tr>
<tr>
<td>5</td>
<td>Error in Alias Determination</td>
<td>Ms. Washington, a long time (House) (_1) staffer and an expert in securities laws, ... Ms. Washington's candidacy is championed by ... John Dingell (D., Mich.) of (the House Energy and Commerce Committee) (_2).</td>
<td>2</td>
<td>5.3</td>
</tr>
</tbody>
</table>
# Recall Errors (false -ve)

<table>
<thead>
<tr>
<th>No</th>
<th>Type of error</th>
<th>Example</th>
<th>Frequency</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Inadequacy of current surface features</td>
<td>Mr. X, (general manager)(_1) of ..., was named (president)(_2) of ...</td>
<td>38</td>
<td>63.3</td>
</tr>
<tr>
<td>2</td>
<td>Errors in NP identification</td>
<td>The NP identification module completely misses candidate phrases in coreference chain</td>
<td>7</td>
<td>11.7</td>
</tr>
<tr>
<td>3</td>
<td>Errors in semantic class determination</td>
<td>Error made in applying semantic class ”Date”. ”...Losses for (fiscal 2(^{nd}) period)(_1) faltering sales will result in (second-quarter)(_2) loss....</td>
<td>7</td>
<td>11.7</td>
</tr>
<tr>
<td>4</td>
<td>Errors in POS assignment</td>
<td>Nouns are not being marked correctly. Mr. X, who is ED in (Canada), succeeds Mr. Y as VP and GM (there)</td>
<td>5</td>
<td>8.3</td>
</tr>
<tr>
<td>5</td>
<td>Error in apposition Determination</td>
<td>Not able to detect coreference due to apposition (Bill Gates)(_1) ,(the chairman of Microsoft)(_2)</td>
<td>2</td>
<td>3.3</td>
</tr>
<tr>
<td>6</td>
<td>Errors due to tokenization</td>
<td>...debt-to-equity ration changed to (1-to-2)(_1) from (15-to-1)(_2)</td>
<td>1</td>
<td>1.7</td>
</tr>
</tbody>
</table>
Conclusion and further Improvements

- Works on a small annotated corpus
- Domain and language independent
- Resolves noun phrase coreferences in general and not limited to pronominal coreference resolution.
- We can consider verb suffixes to determine gender in morphologically rich languages. Similarly, other language specific properties can be taken into consideration.
- This is a sequence labelling problem. We can apply techniques like HMM and CRF instead of scalar classifiers like decision trees.


Thank you