Task: Noun Phrase Coreference Resolution

- Identify the noun phrases (NPs) that refer to the same real-world entity in a text or dialogue

An Ensemble-Based Approach

- Employ an ensemble of models for making coreference decisions

Why an Ensemble-Based Approach?

- Hypothesis: Existing coreference models have complementary strengths and weaknesses, i.e., no single model is the best!

Our Goal

- Investigate new methods for creating and applying ensembles for coreference resolution

Related Work: Creating Ensembles for Coreference

- Munson et al. (2005) employ different learning algorithms
- Ng (2005) employs different clustering algorithms
- Ng & Cardie (2003), Kouchnir (2004), Vemulapalli et al. (2009) perturb the training set using bagging and boosting

Creating an Ensemble: Two Methods

Method 1: Employ 3 different linguistic feature sets

- Conventional feature set
  - contains 39 commonly-used coreference features, which can be divided into four categories:
    - String-matching features: exact and partial string match, ...
    - Grammatical features: gender and number agreement, ...
    - Semantic features: alias, semantic class compatibility, ...
    - Positional features: distance between two NPs in sentences, ...

- Lexical feature set
  - contains word pairs collected from coreference-annotated documents
    - for lexical features to be effective, need to combat data sparsity, e.g.,
      - by replacing a named entity with its named entity tag
      - by replacing a common noun phrase with its head noun

- Combined feature set
  - is the union of the Conventional and the Lexical features

Method 2: Employ 3 different supervised coreference models

- Mention-pair (MP) model (Soon et al., 2001; Ng & Cardie, 2002)
- a classifier that determines whether two NPs are coreferent

- Mention-ranking (MR) model (Denis & Baldwin, 2008)
- a ranker that ranks the candidate antecedents for each anaphor

- Cluster-ranking (CR) model (Rahman & Ng, 2009)
- a ranker that ranks the preceding clusters for each anaphor
  - employs cluster-level features:
    - defined over any subset of NPs in a preceding cluster
    - derived from the combined features by applying logical predicates

Given these two methods, we create a 9-member ensemble

- Since each of these three models can be trained in combination with each of the three feature sets, we can create nine coreference systems

Applying the Ensemble

- Challenge: since our ensemble is model-heterogeneous comprising both pairwise models (e.g., the MP model) and a cluster-based model (i.e., the CR model), combining the coreference decisions made by different models is not straightforward
- Consequently, we propose 4 methods for applying our ensemble

Four Methods for Applying the Ensemble

Method 1: Applying Best Per-NP-Type Model

- Motivation: different members of the ensemble are good at resolving different types of NPs
  - So, for each type of NPs, we identify the member that is best at resolving NPs of this type using held-out development data
  - When resolving an NP in a test text, we first identify its NP type, and then resolve it using the best model given this NP type

Method 2: Antecedent-Based Voting

- Given an NP to be resolved, NP_k, each member independently selects an antecedent for NP_k
  - The candidate antecedent that receives the largest number of votes will be selected as the antecedent for NP_k
  - Caveat: since CR-members select preceding clusters, we force each CR-based member to select an antecedent by assuming that the antecedent it selects is the last NP in the preceding cluster it selects

Method 3: Cluster-Based Voting

- A natural alternative to Method 2
  - Instead of forcing the CR-based members to select antecedents, we force the MP- and MR-based members to select preceding clusters
  - E.g., if the MP model selects NP, as the antecedent, then we assume that it selects the preceding cluster containing NP
  - Every NP in the selected preceding cluster gets one vote
  - The NP with the largest number of votes wins

Method 4: Weighted Cluster-Based Voting

- Motivation: In Method 3, all the votes cast for a candidate antecedent have equal weights; in practice, however, some members are more important than the others, so their votes should have higher weights
  - So, we weight the weights on held-out development data using a hill-climbing algorithm that optimizes the weight of one member at a time
  - We then perform cluster-based voting, except that votes are weighted
  - The NP with the largest number of weighted votes wins

Experimental Setup

- Dataset: ACE 2005, which has 6 data sources, including broadcast news (bn), broadcast conversations (bc), newswire (nw), webblog (wb), usenet (un), and conversational telephone speech (cts)
  - For each data source, use 80% of data for training; 20% for testing
  - Extract NPs using a mention detector trained on training texts
  - All coreference models are trained using SVM (Bagga & Baldwin, 1998)
  - System output is scored using B^2 (Bagga & Baldwin, 1998)

Results and Discussion

- Baseline: Since our goal is to determine the effectiveness of ensemble approaches, the baselines are non-ensemble-based
  - 9 baselines, corresponding to the 9 members of the ensemble
  - First 9 columns in the table below are baseline B^2 F-measure scores
    - Each row corresponds to a data source; last row has aggregate results
  - Conv., lex., and comb are Conventional, Lexical, & Combined feature sets

<table>
<thead>
<tr>
<th>Source</th>
<th>MP Models</th>
<th>MR Models</th>
<th>CR Models</th>
<th>Ensembles</th>
</tr>
</thead>
<tbody>
<tr>
<td>conv.</td>
<td>lex.</td>
<td>comb.</td>
<td>lex.</td>
<td>comb.</td>
</tr>
<tr>
<td>M1</td>
<td>M2</td>
<td>M3</td>
<td>M4</td>
<td>M5</td>
</tr>
<tr>
<td>bn</td>
<td>59.8</td>
<td>60.0</td>
<td>59.8</td>
<td>59.8</td>
</tr>
<tr>
<td>bc</td>
<td>67.4</td>
<td>67.4</td>
<td>67.4</td>
<td>67.4</td>
</tr>
<tr>
<td>hs</td>
<td>62.3</td>
<td>62.3</td>
<td>62.3</td>
<td>62.3</td>
</tr>
<tr>
<td>h3</td>
<td>62.3</td>
<td>62.3</td>
<td>62.3</td>
<td>62.3</td>
</tr>
<tr>
<td>cts</td>
<td>67.4</td>
<td>67.4</td>
<td>67.4</td>
<td>67.4</td>
</tr>
<tr>
<td>nw</td>
<td>64.9</td>
<td>64.9</td>
<td>64.9</td>
<td>64.9</td>
</tr>
<tr>
<td>us</td>
<td>59.8</td>
<td>59.8</td>
<td>59.8</td>
<td>59.8</td>
</tr>
<tr>
<td>x0</td>
<td>63.2</td>
<td>63.2</td>
<td>63.2</td>
<td>63.2</td>
</tr>
<tr>
<td>x1</td>
<td>63.2</td>
<td>63.2</td>
<td>63.2</td>
<td>63.2</td>
</tr>
<tr>
<td>Overall</td>
<td>62.6</td>
<td>62.6</td>
<td>62.6</td>
<td>62.6</td>
</tr>
</tbody>
</table>

Best-performing baseline is CR-comb (F-measure: 62.8), which does not achieve the best performance on each data source among the baselines