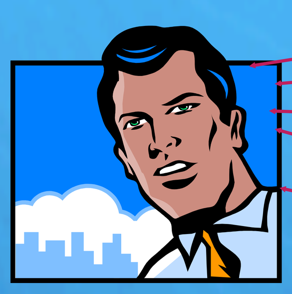
Ensemble-Based Coreference Resolution

Altaf Rahman and Vincent Ng
Human Language Technology Research Institute
The University of Texas at Dallas

Task: Noun Phrase Coreference Resolution

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



John Simon, Chief Financial Officer of Prime Corp. since 1986, saw his pay jump 20%, to $1.3 million, as the 37-year-old also became the financial-services company's president…

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 features 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 & Baldridge, 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 the 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
  - Idea: 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_j as the antecedent, then we assume that it selects the preceding cluster containing NP_j
  - 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 casted 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 learn 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

- Corpus: 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^{light}
- System output is scored using B^3 (Bagga & Baldwin, 1998)

Results and Discussion

- Baselines: 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^3 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

	MP Models			MR Models			CR Models			Ensembles			
Source	conv	lex	comb	conv	lex	comb	conv	lex	comb	M1	M2	M3	M4
bc	50.8	57.4	55.7	52.9	56.5	54.1	55.1	57.7	58.2	59.1	59.7	60.2	**61.9**
bn	53.4	62.3	62.7	55.8	63.5	63.7	62.7	63.3	62.5	63.9	64.6	65.2	**66.9**
cts	57.0	61.1	61.3	58.6	62.7	61.7	62.5	61.1	64.1	66.0	67.0	67.6	**69.7**
nw	57.7	64.9	60.8	60.2	65.4	61.3	61.5	65.3	64.6	65.1	66.2	66.5	**68.3**
un	53.7	54.8	55.4	55.6	56.3	56.0	56.2	55.7	58.1	58.9	59.2	59.5	**61.4**
wb	63.3	65.2	57.6	65.2	68.7	54.5	67.0	63.3	67.9	69.0	69.5	69.9	**71.5**
Overall	56.2	61.2	58.8	58.2	62.4	61.2	61.2	61.5	62.8	63.7†	64.4†	64.8†	**66.8†**

  - Best-performing baseline is CR-comb (F-measure: 62.8), which does not achieve the best performance on each data source among the baselines
- Ensemble approaches: M1, M2, M3, M4 (last 4 rows of the table) correspond to the four methods for applying ensembles
  - All four ensemble methods perform better than CR-comb
    - Ensemble approaches can indeed improve coreference resolution
    - M4 (best ensemble method, F-measure: 66.8) outperforms CR-comb by 4.0% and achieves the best performance on each data source