Ensemble-Based Coreference Resolution

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

Coreference Resolution

- Identify all noun phrases (**mentions**) that refer to the same real world entity

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...

Ensemble Approach

- What ?
  - Employ an ensemble of models for making coreference decisions
- Why ?
  - Hypothesis: Existing coreference models have complementary strengths and weaknesses, i.e., no single model is the best!
- Goal
  - Investigate new methods for creating and applying ensembles for coreference resolution

Related Works

- Existing methods for creating ensemble for coreference resolution:
  - 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 new methods
  - Method 1: employs different linguistic feature sets
  - Method 2: employs different supervised coreference models

Ensemble Creation : Method 1

- 3 different feature set
  1. Conventional Feature Set
     - It 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, …
  2. **Lexical Feature Set**
     - It contains word-pairs collected from coreference-annotated documents
       - For example : his-president, Simon-his, Prime Corp-his
       - Additionally, to improve generalizability we replace a named entity with its named entity tag
         - "John Simon" is replaced with "PERSON" to create a new feature like PERSON-his
  3. **Combined Feature Set**
     - Union of the above two features.

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...

Ensemble Creation : Method 2

- 3 different supervised models
  1. **Mention Pair (MP) model (Soon et al., 2001; Ng & Cardie, 2002)**
  2. **Mention Ranking (MR) model (Denis & Baldridge, 2008)**
  3. **Cluster Ranking (CR) model (Rahman & Ng, 2009)**

- A classifier that determines whether two NPs are coreferent
- Weaknesses:
  - Each candidate antecedent is considered independently of the others.
  - Insufficient information to make an informed coreferenced decision based on two NPs only.

- A ranker that ranks the candidate antecedents for each anaphor
- Advantage:
  - Considers all the candidate antecedents simultaneously.

- A ranker that ranks the preceding clusters for each anaphor
- It employs cluster-level features
  - defined over any subset of NPs in a preceding cluster
  - derived from the Combined features by applying logical predicates
- Advantage:
  - Considers all the candidate antecedents simultaneously.
  - It also improves expressiveness by using cluster level features.

Creating the Ensemble

- 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:
  - Our ensemble is model-heterogeneous, so comprising both pair-wise 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.

Method 1: Applying Best Per-NP-Type Model

- Motivation: different members of the ensemble are good at resolving different types of NPs
- Identify the best model resolving each type of NPs by using a held-out dev-set.
- Resolving an NP :
  - Identify the type of the NP
  - Resolve it using the model that was determined to be the best at handling this NP type.

Method 1: Applying Best Per-NP-Type Model (cont.)

1. How many NP types should be used?

- Three super types (*Name, Nominal and Pronoun*) are further divided into total 10 subtypes :
  - Name and Nominal
    - e (exact string match)
    - p (partial string match)
    - n (no string match)
  - Pronoun
    - 1+2 (1^{st} and 2^{nd} person pronoun)
    - G3 (gendered 3^{rd} person)
    - U3 (ungendered 3^{rd} person)
    - oa (other anaphoric pronoun)

2. How can we determine which model performs the best for an NP type on the development set ?

- For each type C of NP we use a model and rest of the NPs are resolved by the oracle.
- Compute F-measure score only on the NPs belong to type C

Method 2: Antecedent-Based Voting

- Given an NP to resolve, NP_k, each of the 9 models selects an antecedent NP_k independently -
- The candidate antecedent that receives the largest number of votes will be selected as the antecedent for NP_k
- Caveat: since Cluster Ranking (CR) members select preceding clusters, we force them to select the last NP of the cluster as the antecedent.

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
  - if the MP and MR 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.
- Dev-set : we learn the weights on held-out development data using a hill-climbing algorithm which optimizes the weight of one member at a time, selecting the weight from the set {−4, −3, −2, −1, 0, 1, 2, 3, 4}
- Testing : we then perform cluster-based voting, except that votes are weighted
  - The antecedent NP with the largest number of weighted votes wins

Experimental Setup

- Corpus: ACE 2005, which has 6 data sources
  - 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)

Evaluation

- 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.

Baseline Results

src	MP Models			MR Models			CR Models		
	conv	lex	comb	conv	lex	comb	conv	lex	comb
bc	50.8	57.4	55.7	52.9	56.5	54.1	55.1	57.7	58.2
bn	53.4	62.3	62.7	55.8	63.5	63.7	62.7	63.3	62.5
cts	57.0	61.1	61.3	58.6	62.7	61.7	62.5	61.1	64.1
nw	57.7	64.9	60.8	60.2	65.4	61.3	61.5	65.3	64.6
un	53.7	54.8	55.4	55.6	56.3	56.0	56.2	55.7	58.1
wb	63.3	65.2	57.6	65.2	68.7	54.5	67.0	63.3	67.9
all	**56.2**	**61.2**	**58.8**	**58.2**	**62.4**	**61.2**	**61.2**	**61.5**	**62.8**

- 9 baseline systems on the test set, reported in terms of B^3 F-measure

- Columns labeled 'conv', 'lex', and 'comb' correspond to the *Conventional, Lexical, and Combined* feature sets, respectively.

- Aggregate results are in the last row

- The best performing baseline is CR-comb, which achieves comparable performance to Haghighi & Klein's (2010) system on the same test set.

Ensemble Results

src	MP Models			MR Models			CR Models			Ensembles			
	cnv	lex	cmb	cnv	lex	cmb	cnv	lex	cmb	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
all	**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**

1 Ensemble approaches: M1, M2, M3, M4 correspond to the 4 methods for applying ensembles.

- All four ensemble methods perform better than CR-comb
- Ensemble approaches can indeed improve coreference resolution (M1 < M2 < M3 < M4)
- M4 (best ensemble method, F-measure: 66.8) outperforms CR-comb by 4.0% and achieves the best performance on each data source.

Ensemble Results

	CR-comb			M1			M2			M3			M4		
	R	P	F	R	P	F	R	P	F	R	P	F	R	P	F
ll	**54.4**	**74.8**	**62.8**	**55.1**	**75.6**	**63.7**	**55.5**	**76.6**	**64.4**	**55.7**	**77.5**	**64.8**	**57.6**	**79.5**	**66.8**

- M1, M2, M3 and M4 - all improve on both recall and precision over CR-comb model.

Summary

- New methods for creating and applying ensembles of learning-based coreference systems
  - Uses different supervised models (pair-wise and cluster-based) and different feature sets.
- Experimental results on the ACE 2005 data set show that all four ensemble methods outperform the best baseline.
  - The best result was achieved by applying weighted cluster-based voting.

Thank You !!!