A Method for Resolving Japanese Zero Pronouns

Ryu IIDA
Computational Linguistics Lab.
ryu-i@is.naist.jp

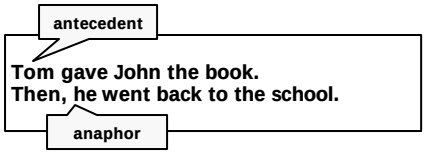
[COE technical presentation, 24 August 2004]

Background

- the accessibility to Web documents
  - most part of information on WWW is transferred by **text**
- Toward intelligent treatment of **text data**
  - we are working on natural language processing (NLP)
  - Development of NLP techniques
    - Machine translation, information extraction
- The process which crosses sentence boundaries, such as **anaphora resolution**, remains a major obstacle to further improvements

Motivation

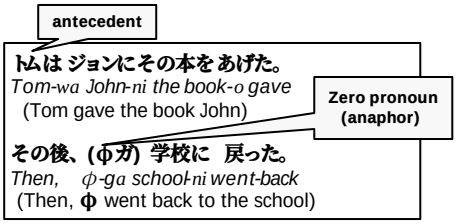
- Developing the performance of anaphora resolution suitable for real world applications
- **Anaphora resolution**
  - **process tracking entities in text**



  - **Detecting antecedents against anaphors**

Zero pronouns resolution

- In Japanese, anaphors are omitted because of speaker's and hearer's recognition



Previous work

Two approaches to anaphora resolution

- Rule-based approach
  [Mitkov 97, Baldwin 95, Nakaiwa 96, Okumura 95, Murata 97]
  - Many attempted to encode linguistic cues into rules

Problem: Further manual refinement is needed in this study but it will be prohibitively costly

  - Best-achieved performance in MUC: Precision roughly 70%
    (Message Understanding Conference) Recall roughly 60%
- Corpus-based machine learning approach

Problem: These previous work tend to lack an appropriate reference to the theoretical linguistic work on coherence and coreference

ng rule-based systems

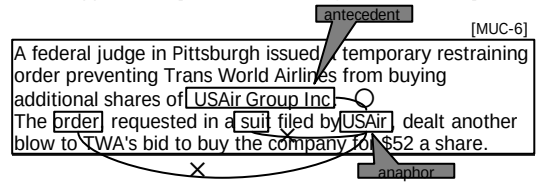
Challenging issue

- Achieving a good union between theoretical linguistic findings and corpus-based empirical methods

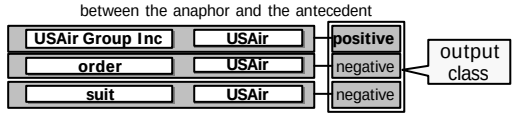
Statistical approaches [Soon et al. '01, Ng and Cardie '02]

- Reach a level of performance comparable to state-of-the-art rule-based systems
- Recast the task of anaphora resolution as a sequence of classification problems

Statistical approaches [Soon et al. '01, Ng and Cardie '02]

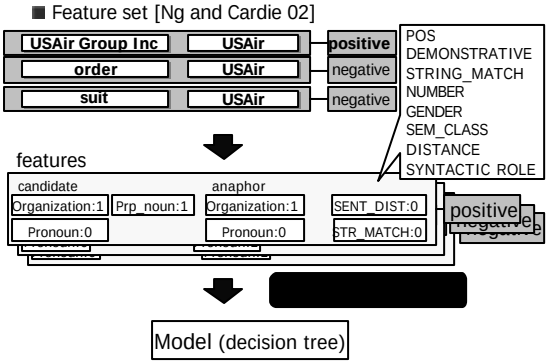


- the task is to classify these pairs of noun phrases into positive or negative
  - positive instance: Pair of an anaphor and the antecedent
  - negative instance: Pairs of an anaphor and the NPs located between the anaphor and the antecedent



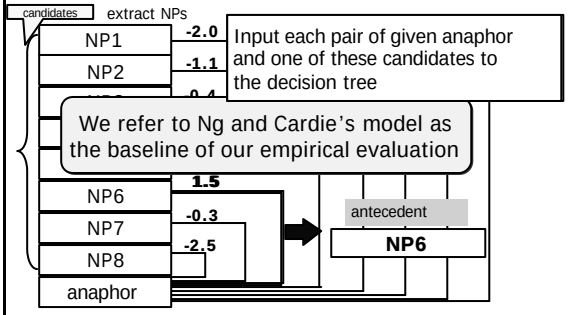
Statistical approaches [Soon et al. '01, Ng and Cardie '02]

- Feature set [Ng and Cardie 02]



Statistical approaches [Soon et al. '01, Ng and Cardie '02]

- Test Phase [Ng and Cardie, 02]



- Precision 78.0%, Recall 64.2%
  - Slightly better than best-performing rule-based model at MUC-7

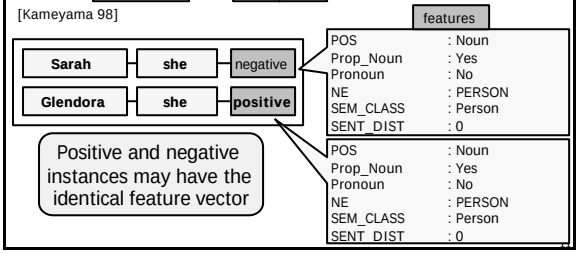
A drawback of the previous statistical models

Sarah went downstairs and received another curious shock, for when [illegible] made mocca[illegible] to her room,

The previous models do not capture local context appropriately

antecedent anaphor

[Kameyama 98]



Positive and negative instances may have the identical feature vector

Proposed model

- Inspired by Centering Theory that captures the local contextual factors
- Improve the search algorithm: *tournament model*
  - A new model which makes pair-wise comparisons between candidates

Centering Theory [Grosz `95]

- Capturing the salience of sentences

Topic > Subj > I-obj > Obj > Others

Sarah went downstairs and received another curious shock, for when Glendora flapped into the dining room in her home made moccasins, Sarah asked her when she had brought coffee to her room, and Glendora said she hadn't.

Sarah went downstairs and received another curious shock,
Salience: CHAIN(Cb = Cp = Sarah)

.....

transition

she hadn't.
Salience: CHAIN(Cb = Cp = Glendora) ⇒ antecedent: Glendora

13

Tournament model

- What we want to do is to answer a question which is more likely to be coreferent, Sarah or Glendora

Sarah went downstairs and received another curious shock, for when Glendora flapped into the dining room in her home made moccasins, Sarah asked her when she had brought coffee to her room, and Glendora said she hadn't.

- Conduct a tournament consisting of a series of matches in which candidates compete with each other
  - Match victory is determined by a pairwise comparison between candidates as a binary classification problem
  - Most likely candidate is selected through a single-elimination tournament of matches

14

Tournament model

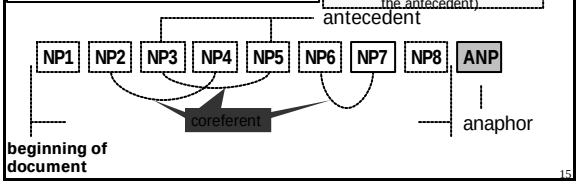
- Training Phase

- In the tournament, the correct antecedent NP5 must prevail over any of the other four candidates
- Extract four training instances
- Induce a pairwise classifier from a set of extracted training instances
- The classifier classifies a given pair of candidates into left or right

Training instances

features			class
NP1	NP5	ANP	right
NP4	NP5	ANP	right
NP5	NP7	ANP	left
NP5	NP8	ANP	left

the right hand side of a given pair wins (is more likely to be the antecedent)

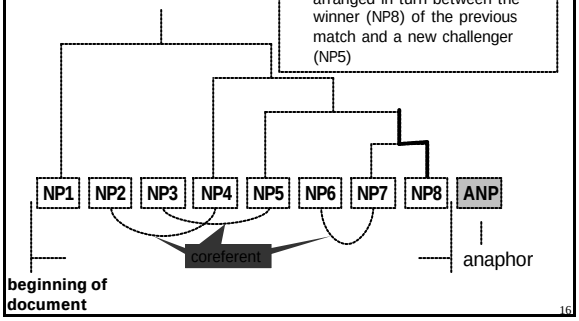


15

Tournament model

- Test Phase

1. the first match is arranged between the nearest candidates (NP7 and NP8)
2. each of the following matches arranged in turn between the winner (NP8) of the previous match and a new challenger (NP5)

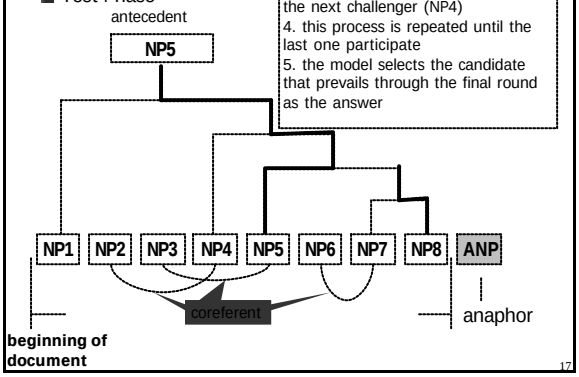


16

Tournament model

- Test Phase

3. the winner is next matched against the next challenger (NP4)
4. this process is repeated until the last one participate
5. the model selects the candidate that prevails through the final round as the answer



17

Experiments

- Empirical evaluation on Japanese zero pronouns resolution
- Comparison among three models
  1. Rule-based model [Nariyama `02]
  2. Previous statistical model (baseline model) [Ng and Cardie `02]
  3. Tournament model (proposed model)

18

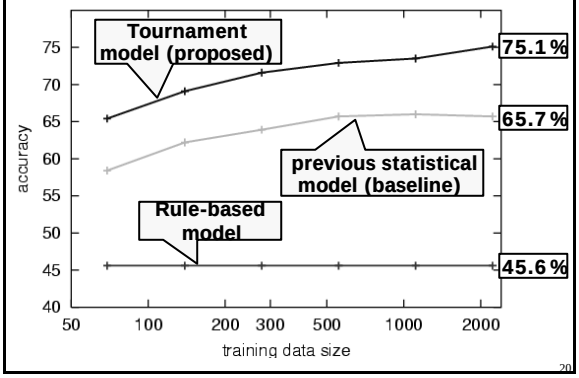
Method

- Data
  - Zero pronouns tagged Japanese newspaper article corpus

	GDA	MUC-6
Texts	: 2,176	60
Sentences	: 24,475	-
Tags of anaphoric relation	: 2,781	8,946

- As a preliminarily test, only resolving subject zero-anaphors, 2,781 instances in total
- Conduct five fold cross-validation on that data set with support vector machines

Results



Conclusions

- Our concern is achieving a good union between theoretical linguistic findings and corpus-based empirical methods
- We presented a trainable anaphora resolution model that is designed to incorporate contextual cues by means of a tournament-based search algorithm
- Future work

  In Japanese zero-anaphora resolution,

  1. Identification of relations between the topic and subtopics
  2. Analysis of complex and quoted sentences
  3. Refinement of the treatment of selectional restrictions