Learning An Anaphoricity Determination Model Combining Preceding and Local Contextual Information

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

[COE technical presentation, 25 February 2005]

Background

- A huge amount of text data on the Web
- For handling such a text data on ubiquitous computing network, Natural Language Processing (NLP) techniques are important
  - Machine Translation, Information Extraction and Question Answering
- Previous work on NLP have limited their target onto sentence
  - Linguistic phenomena and problems crossing multiple sentences are relatively unexplored
- We tackle on of the discourse related processing:
  coreference resolution

(Noun phrases) coreference resolution

- **Coreference resolution** is the process of determining whether two expressions in natural language refer to the same entity in the world

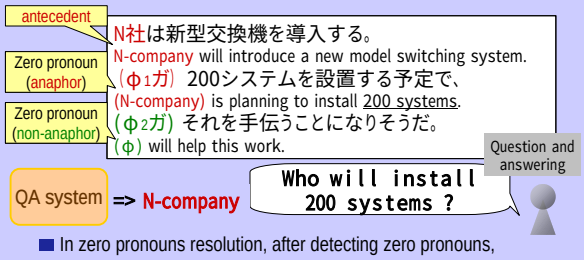
A federal judge in Pittsburgh issued a temporary restraining order preventing Trans World Airlines from buying additional shares of USAir Group Inc. The order, requested in a suit filed by USAir, dealt another blow to TWA's bid to buy the company for $52 a share.

antecedent — anaphor — non-anaphor

- **Coreference resolution** is decomposed into two sub processes
  1. **Anaphoricity determination** is the task of classifying whether a given noun phrase (NP) is *anaphoric* or *non-anaphoric*
  2. **Antecedent identification** is the identification of the antecedent of a given anaphoric NP

Japanese zero pronouns resolution

- In Japanese, anaphors are frequently omitted because of speaker' s and hearer's shared understanding



- In zero pronouns resolution, after detecting zero pronouns, the processes of anaphoricity determination and antecedent identification are needed as well as NP coreference resolution

Anaphoricity determination

- Early corpus-based work on coreference resolution does not address anaphoricity determination (Hobbs `78, Lappin and Leass `94)
  - Assuming that the coreference resolution system knows a priori all the anaphoric noun phrases
- This problem has been paid attention by an increasing number of researchers
  (Bean and Riloff `99, Ng and Cardie `02, Uryupina `03, Ng `04)
  - Determining anaphoricity is not a trivial problem
  - Overall performance of coreference resolution crucially depends on the accuracy of anaphoricity determination
- The problems of anaphoricity determination is even more critical in case of Japanese, because of the absence of articles
- Our aim is improving the performance of anaphoricity determination for the overall performance of coreference resolution in Japanese

Essential information for anaphoricity determination

Two linguistic clues :

1. Preceding contextual information
   - Antecedent information
2. Local contextual information
   - Non-anaphoric information

1. Antecedent information

- Information extracted from pairs between an anaphor candidate and an antecedent candidate

antecedent candidate

A federal judge in Pittsburgh issued a temporary restraining order preventing Trans World Airlines from buying additional shares of USAir Group Inc.
The order, requested in a suit filed by USAir, dealt another blow to TWA's bid to buy the company for $52 a share.

anaphor candidate

USAir Group Inc	USAir	Coreferent

Anaphor candidate "USAir" has the corresponded antecedent "USAir Group Inc"
=> "USAir" is judged as anaphor

2. Non-anaphoric information

- Noun phrases Information that contrast with anaphor information

antecedent

A federal judge in Pittsburgh issued a temporary restraining order preventing Trans World Airlines from buying additional shares of anaphor Inc. non-anaphor
The order, requested in a suit filed by USAir, dealt another blow to TWA's bid to buy the company for $52 a share.

"The order" has an article "The" => anaphor

contrast

"a suit" has an article "a" => non-anaphor

Previous work (learning-based approaches)

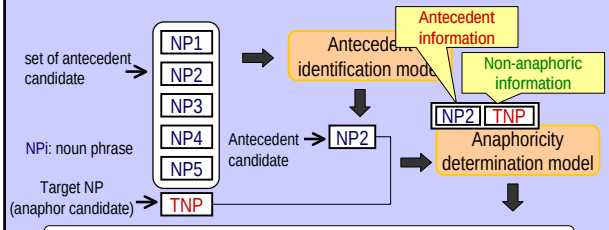
- Search-based approach
  (Soon et al. `01, Ng and Cardie `02, Yang et al. `03)
  - Advantage: solving indirectly the problem of anaphoricity determination by searching the antecedent for a given anaphor
    (⇒ antecedent information)
  - Disadvantage: this model is not designed to learn non-anaphors
    (⇏ non-anaphoric information)
- Classification approach
  (Bean and Riloff `99, Ng and Cardie `02, Uryupina `03, Ng `04)
  - Advantage: learning explicitly the behavior of non-anaphors
    (⇒ non-anaphoric information)
  - Disadvantage: this model does not use the contextual information introduced by the search-based approach
    (⇏ antecedent information)

Proposed approach

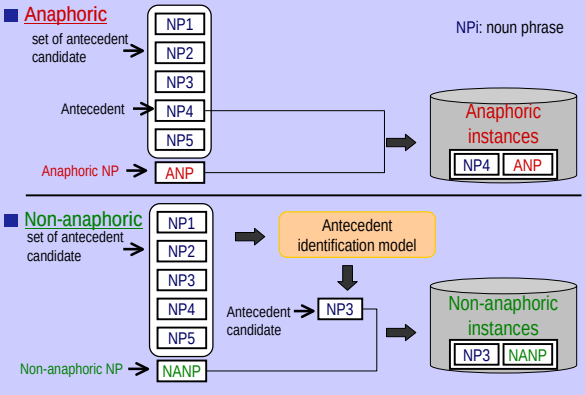
- Combining the advantages of
  - search-based model
  - classification model
- We have an advantage to utilize both
  1. antecedent candidate as the preceding contextual information
  2. non-anaphoric instances

2-step processing:

1. Identifying the most likely antecedent candidate for a target NP
2. Determining anaphoricity of the target NP using a pair of the target NP and the most likely antecedent



Proposed model (training phase)

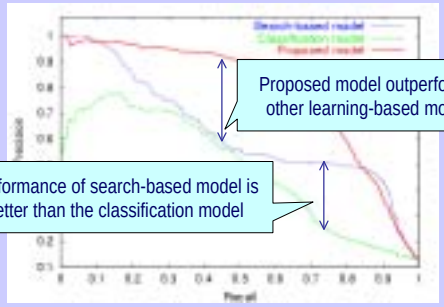


Experiments

- Empirical evaluation on anaphoricity determination of noun phrases and zero pronouns in Japanese
- Data (newspaper article corpus)
  - Noun phrases : 876 anaphors and 6,292 non-anaphors
    -> detecting anaphors
  - Zero pronouns: 4,225 anaphors and 1,957 non-anaphors
    -> detecting non-anaphors
- Conduct 10-fold cross-validation with support vector machines
- Comparison among three models
  1. Search-based model (Soon et al. `01)
  2. Classification model (Ng and Cardie `02)
  3. Proposed model

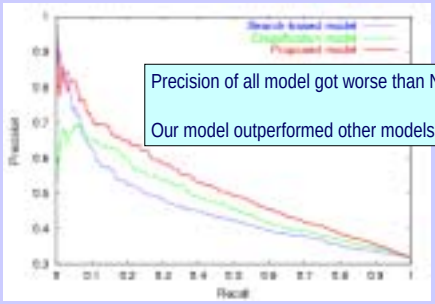
Results on noun phrases anaphoricity determination

- 876 anaphors and 6,292 non-anaphors (detecting anaphors)



Results on zero pronouns anaphoricity determination

- 4,225 anaphors and 1,957 non-anaphors (detecting non-anaphors)



Results

- 9-points average precision (Recall = 0.1, 0.2, ..., 0.9)

	Search-based approach	Classification approach	Proposed approach
Noun phrases	63.6%	49.2%	81.1%
Zero pronouns	44.2%	47.3%	50.9%

- Prec. of zero pronouns << Prec. of noun phrases
  -> Difference of extracted features
  - Noun phrases: string sequence information
    (e.g. antecedent "USAir Group INC." and anaphor "USAir")
  - Zero pronouns: such information is not introduced
    because zero pronouns have no surface strings.

Conclusion

- We proposed an anaphoricity determination model
  - Preceding contextual information
  - Non-anaphoric instances
- Proposed model outperformed previous machine learning-based models
  - Noun phrases: 49.2% -> 81.1%
  - Zero pronouns: 44.2% -> 50.9%

Future work

- Noun phrases:
  - Analysis of the definiteness (whether a target NP is definite or not)
- Zero pronouns:
  - Improvement of the quality of selectional restrictions
  - Analysis of the relation between anaphoricity and discourse structure