# Heuristic-based Korean Coreference Resolution for Information Extraction

**Euisok Chung, Soojong Lim, Bo-Hyun Yun**
Human Information Processing Dept.
Electronics and Telecommunications Research Institute
161, Kajong-Dong, Yusong-Gu, Daejon, 305-350, KOREA
{eschung, isj, ybh}@etri.re.kr
Tel. 82-42-860-1014, Fax. 82-42-860-4889

## Abstract

The information extraction is to delimit in advance, as part of the specification of the task, the semantic range of the output and to filter information from large volumes of texts. The most representative word of the document is composed of named entities and pronouns. Therefore, it is important to resolve coreference in order to extract the meaningful information in information extraction. Coreference resolution is to find name entities co-referencing real-world entities in the documents. Results of coreference resolution are used for name entity detection and template generation. This paper presents the heuristic-based approach for coreference resolution in Korean. We constructed the heuristics expanded gradually by using the corpus and derived the salience factors of antecedents as the importance measure in Korean. Our approach consists of antecedents selection and antecedents weighting. We used three kinds of salience factors that are used to weight each antecedent of the anaphor. The experiment result shows 80% precision.

## 1 Introduction

Information extraction(IE) systems take texts containing natural language as input and produce database templates relevant to a particular application. IE system must create templates describing the relevant entities that are reported on. This requires determining when two or more templates describe the same entity, as templates created from conferencing words to be merged. Thus, it is difficult to resolve coference in order to extract more reliable information.

Results of coreference resolution are used for the clue of template generation. In coreference resolution, there are two kinds of problems such as anaphora resolution and name aliases recognition. The name aliases could be resolved by lexical pattern matching or synonym dictionary(Huyck, C. (1998). Fukumoto, J., Masui (1998)). However, the anaphora resolution has more complexities of natural language. It has been studied conservatively in the discourse part of natural language processing. Recently, several proposals addressed that using limited knowledge is better than using heavy linguistic and domain knowledge(Lappin, S. and Leass, H. (1994). Baldwin, F. B. (1995). Mitmov, R. (1998)).

This paper presents the heuristic-based approach with limited knowledge such as pattern rules, preference rules, and conditional rules. The resolution procedure is to find antecedents and then to evaluate the weight of antecedents with the heuristics. In this paper, we focus on the anaphora resolution. In Korean, an anaphora consists of 'pronoun' and 'demonstrative pronoun + noun phrases'.

## 2 Related Works

Many researches have been performed to solve the problem of coreference resolution. One is the research of anaphora resolution which is based on the discourse theory such as centering theory(Lappin, S. and Leass, H. (1994). Baldwin, F. B. (1995). Mitmov, R. (1998)), another is the information extraction system in order to apply to MUC(message understanding conference)(Huyck, C. (1998). Yangarber, R. and Grishman, R. (1998). Urbanowicz, R. A. and Nettleton, D. J. (1998). Humphreys,K.,

Gaizauskas, R. Azzam, S., Huyck, C., Mitchell, B., Cunningham, H. and Wilks, Y. (1998). Lin, D. (1998). Fukumoto, J., Masui, F., Shimohata, M., and Sasaki, M. (1998). Aone, C., Halverson, L., Hampton, T., and Ramos-Santacruz, M. (1998)). With a view of the methodology, it could be divided with rule-based approaches using limited knowledges such as lexical patterns(Huyck, C. (1998). Fukumoto, J., Masui, F., Shimohata, M., and Sasaki, M. (1998)) and heuristics(Lappin, S. and Leass, H. (1994). Baldwin, F. B. (1995). Mitmov, R. (1998)), knowledge based approach using semantic network(Yangarber, R. and Grishman, R. (1998). Humphreys,K., Gaizauskas, R. Azzam, S., Huyck, C., Mitchell, B., Cunningham, H. and Wilks, Y. (1998)), and Hybrid approaches which integrate knowledge based and machine learning approaches(Urbanowicz, R. A. and Nettleton, D. J. (1998). Lin, D. (1998)). Another is statistical approach(Kehler, A. (1997)). In the case of coreference resolution, the statistical approach is rare since the phenomenon of coreference is generally inter-sentential problem than intra-sentential thing. Thus, it makes the statistical modeling of coreference very difficult.

The approach of using limited rules does not depend on the massive linguistic knowledge or domain knowledge but it only depends on the simple heuristics. The general resolution procedure is the selection of antecedent candidates, the ranking of the candidates, and the decision of the candidate of an anaphoric word. In each step, it uses the empirical heuristics. The typical heuristic-based approach is the dynamic coreference model generation such as (Lappin, S. and Leass, H. (1994)). It divides the antecedents into the intra and inter sentential types, and evaluates the salience factors of the antecedents. The approach shows 86% precision in the computer manual domain.

Another heuristic-based research is the file card approach. It makes first discourse model of the anaphora and then resolves the model by the file card operation such as antecedents grouping, deleting and weighting with the morphological and syntactic analysis. It shows the 73% precision in the Wall Street Journal(Baldwin, F. B. (1995)). The best report, 89.7% precision, is the approach that uses the antecedent indicator which is antecedents weighting types(Kehler, A. (1997)). It depends on the heuristics and syntactic pattern of the anaphora contexts. However, in MUC, the heuristic approaches did not report affirmative result(Huyck, C. (1998). Fukumoto, J., Masui, F., Shimohata, M., and Sasaki, M. (1998)).

According to the previous researches, the heuristic-based approaches presented good results. Therefore, we use the approach. Furthermore, since a coreference resolution is the part of information extraction and the independence and conciseness of the module is mostly important, the approach of the using limited knowledge is appropriate for the information extraction system.

| *type* | *example* | *num* | *sentence distance* |
|---|---|---|---|
| Formal Coreference | Pronoun | 11 | 1.9 |
| | demonstrative pronoun + noun phrases | 7 | 1.4 |
| | name aliases | 1 | 1 |
| | etc. | 1 | 5 |
| Informal Coreference | antecedent is not name entity | 6 | 2.16 |
| | no antecedent | 2 | - |
| | plural antecedents | 3 | 1 |
| | antecedents is part of name entity | 3 | 1.3 |
| | anaphora is common noun | 7 | 2.28 |
| | etc. | 3 | 6 |

Table 1. the analysis of coreference

## 3 Analysis of Coreference Phenomenon in Korean

In this section, we analyze the anaphora appearance in Korean with 20 documents of the travel/performance domain articles such as table 1. Each article has 13 sentences and 140.7 words(eojol). We detect the 44 coreferences that have 2.19 sentence distance. However, all of the coreferences have

not the antecedent. Only the half of the cases have the antecedent. Therefore, we should divide it into the formal and informal phenomena.

The coreference resolution depends on the name entity since the resolution procedure is the part of the information extraction system and is following the name entity recognition step. Therefore, the antecedents of the coreference are the name entities. In table 1, the formal coreference is the ordinary case, but the informal coreference could not be resolved with the formal approach. Thus, the informal coreference resolution approach is different with the formal approach. In this paper, we focus on the formal phenomenon, but we consider the informal coreference resolution, too.

## 4 Coreference Resolution Based on Limited Knowledge

This paper presents the heuristic-based approach for coreference resolution in Korean. It is similar to the previous approaches in the viewpoint of resolution procedure(Lappin, S. and Leass, H. (1994). Baldwin, F. B. (1995). Mitmov, R. (1998)). However, we devised the heuristics to expand gradually by using the corpus and the derived salience factors of antecedents in Korean.

### 4.1 Coreference Resolution Procedure

The procedure for coreference resolution has three essential steps. First is the name entity recognition which is performed in the name entity module of information extraction system. It recognizes the antecedents and anaphora. In this paper, we don't described this process. Second, antecedents selection process is to select antecedent list per each anaphor. It consists of two steps such as antecedents grouping and eliminating using lexical patterns and disused lexical list. Finally, in the antecedent weighting process, each antecedent is weighted with salience factors. The most weighted antecedent, summation of all weights, is selected as the result. This steps is described in table 2.

| *step* | *description* |
|---|---|
| name entity recognition | person name, location name, organization name detection |
| name aliases grouping | similar name entities grouping with the lexical pattern and syntactic pattern |
| anaphora recognition | anaphora words detection using lexical information |
| antecedent candidates selection | using lexical pattern and name entity type(semantic information) antecedents candidates detection |
| weighting using heuristics | comparing anaphora and antecedents each other with salience factors, then give weight to each antecedent |
| coreference resolution | select the most weighted antecedent |

Table 2. the coreference resolution steps

### 4.2 Heuristics for Coreference Resolution

The heuristics is derived empirically from the training data. First, we find the anaphora candidates and name entities. In each anaphor, we choose the features which is the criteria to select antecedent. The features is derived and is structured. It could make us to find the salience factors and selectional restrictions to the antecedents in Korean. Furthermore, the count of features is used as the weight of each salience factor. With this analysis, we devised the coreference resolution approach for the antecedents selection and antecedents weighting.

The antecedents selection process consists of two steps, antecedents grouping and antecedents eliminating. The antecedent grouping is applied when splitting antecedents exists per the anaphor. Therefore, the antecedents should be grouped to link up with the anaphor. For example, the antecedents

having parallel structure such as "미국, 한국, 일본(*miguk, hanguk, ilbon*, U.S., Korea, Japan)" could be grouped to the anaphor "이 나라들(*i naradeul*, these nations)".

The antecedents eliminating is applied to exclude an anaphora that has not the antecedent obviously. It usually occurred when the antecedent is not kind of name entity. We found the lexical items that could be clues to refer to sentences or events such as "그러한(*greohan*, like that)", "그런(*greon*, like that)", etc. In figure 1, antecedents grouping and anaphor eliminating is represented. We described antecedents group as NE-SET. NE-SET create by the syntactic pattern such as parallel structure and the derived NE-SETs merge by the inner common NEs. Then, anaphor eliminating is processed. Exactly, it is to filter informal anaphoric lexical items such as "그러한(*greohan*, like that)", "그런(*greon*, like that)".

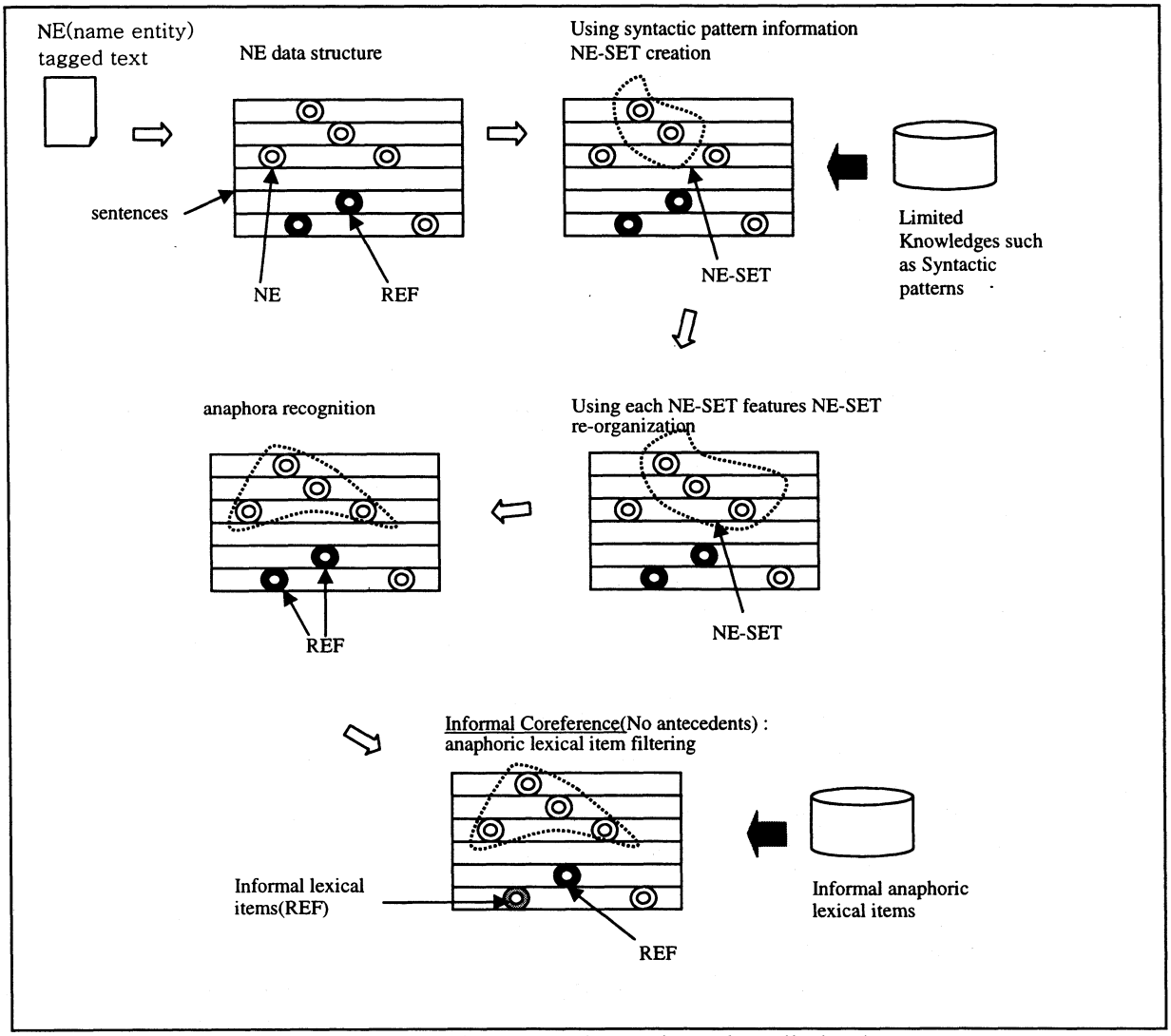


Figure 1. antecedents grouping and anaphor eliminating

The antecedent weighting step is processing by using heuristics. We derived three kinds of salience factors that is to weight each antecedent of the anaphor. In table 3, the heuristics for the antecedent weighting is described. The morphological pattern rules using the similarity between antecedent and anaphor have three types of patterns such as case marker, affix and partial lexical item. The weight is determined in each pattern whether matched or not. The preference rules represent antecedent features as the sentence constituents such as subject and object, the distance from the anaphor and the frequency of antecedent itself. The constituents factor is from the centering theory. Thus, if the antecedent is subject or object, it could be topic word and it has a possibility of the antecedent. The conditional rules

are used in the antecedents selection step. Thus, if an anaphor is person type, only the antecedents of person type is selected above all. In this paper, we use name entity types for the semantic compatibility such as person, location, organization, artifact and titile. In the following we shall illustrate them by examples.

| *Heuristics* | *Similarity clues* |
|---|---|
| Morphological pattern rules | Case marker<br>Affix<br>Partial lexical item |
| Preference rules | Subject/Object<br>Recency<br>Frequency |
| Conditional rules | Reflexive pronoun<br>Syntactic pattern<br>Category restriction |

Table 3. the heuristics for antecedent weighting

- **Morphological pattern rules**
  - *Case marker* : in Korean, a case is determined by grammatical morphemes, case marker such as '-가(-이)(*-ga(-i)*, subject marker)' and '-을(*-eul*, object marker), etc. If the anaphor and antecedent have the same case marker, it is possible to consider them as coreference.
    - antecedent : "몽골인들이(*monggol-in-deul-**i***, Mongolians)"
    - anaphor : "그들이(*geudeul-**i***, they)"

  - *Affix* : Number(*sl* or *pl*) is determined by suffixes such as "-들(-을)(*deul(eul)*, -s(es)). If antecedent and anaphor have same suffix, we regard them as the same. However, this is not absolute rule since there are many cases to conflict with the rule even if they are discorded.
    - ✓ Positive case:
      - antecedent : "몽골인들이(*monggol-in-**deul**-i*, Mongolians)"
      - anaphor : "그들이(*g-**deul**-i*, they)"
    - ✓ Negative case:
      - antecedent: "몽골 국민이(*monggol gukmin-i*, Mongolians)"
      - anaphor: "그들이(*geu**deul**-i*, they)"

  - *Partial lexical item* : Korean is an agglutinative language, thus the compound noun is overflowed. Therefore, it needs the partial lexical matching.
    - antecedent: "몽골국민들이(***monggol**-gukmin-deul-i*, Mongolians)"
    - anaphor: "그 몽골인들은(*geu **monggol**-in-deul-en*, the Mongolians)"

- **Preference rules**
  - *Subject* : this is similar to English. If the antecedent is subject or object, it gains more weight. Thus, if the antecedent is subject, the word could be the topic in text. It means that the topic is possible to be antecedent.
    - 철수가 영희에게 책을 주었다. (*Cheol-su-ga Young-Hee-ege Chak-ul Ju-ot-da* , Cheolsu gives a book to Younghee)

  - *Recency* : this is similar to English. The distance between antecedents and anaphor is important factor.

- *Frequency* : the most important word is usually repetitive in text. This could be topic word. The antecedent having high frequency gains more weight.

- **Conditional rules**
  - *Reflexive pronoun* : if the anaphor is reflexive pronoun, the recency salience factor is more important, since there is no inter-sentential case.
  - *Syntactic pattern* : this is based on the similarity of syntactic pattern between the contexts of anaphor and antecedents such as "ANT **and** x ... ANA **and** y ➔ ANT = ANA"
  - *Category restriction* : this is semantic compatibility. It could be determined by name entity module.
    - antecedent : 김수희씨/PERSON (*kim-su-hee-ssi,* person name)
    - anaphor : 그 사람/PERSON (*geu saram*, the man)

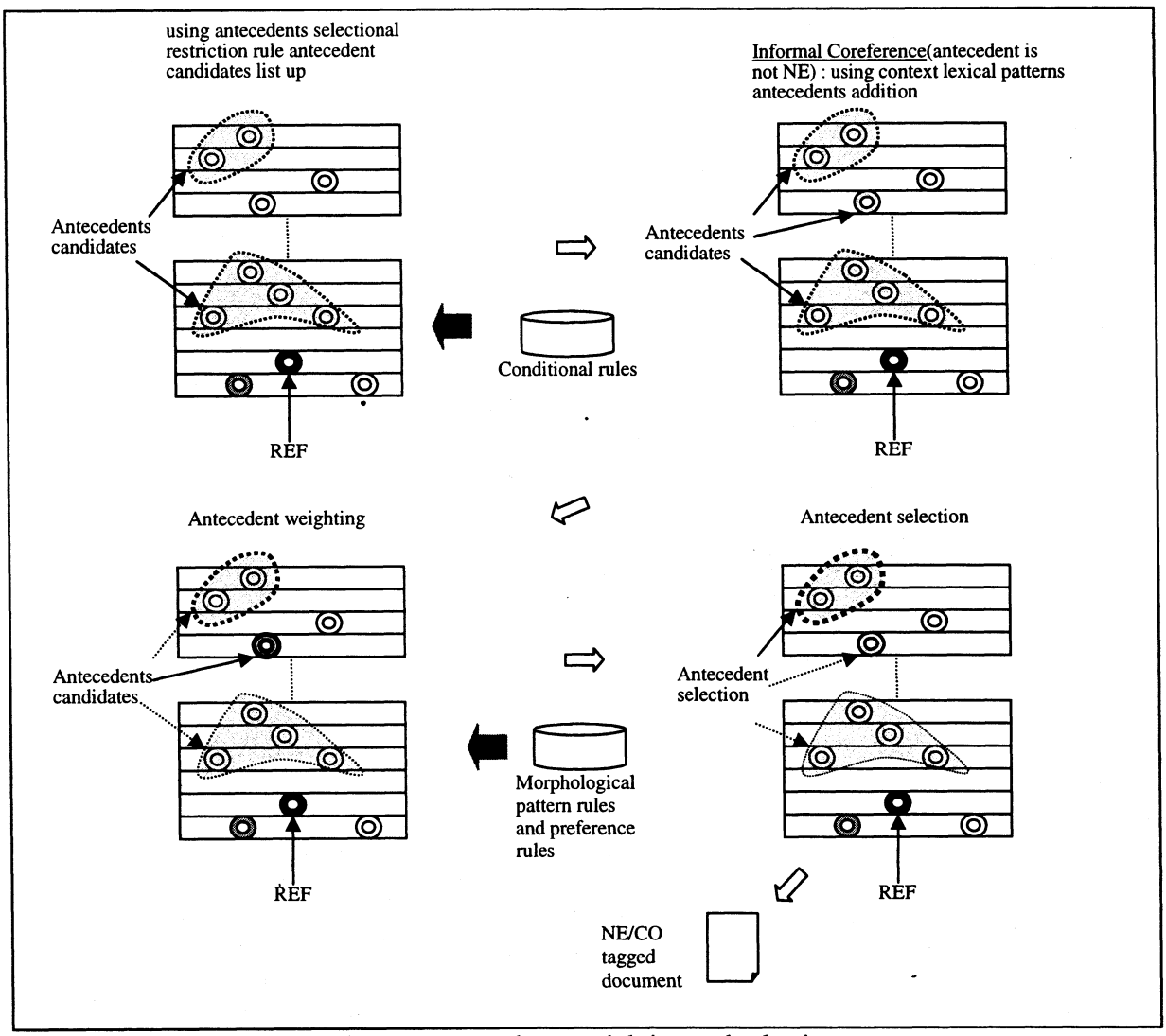


Figure 2. antecedents weighting and selection

In figure 2, antecedents weighting and selection is described. After antecedent grouping and anaphor eliminating, the coreference resolution uses heuristics to determine the appropriate antecedent. First, it list up the antecedents candidates considering semantic compatibility with conditional rules. Then, informal antecedents such as common noun is added to the antecedents by using context lexical patterns.

The context lexical pattern is the lexical items surrounding the name entity and anaphor such as trigram or bigram lexical patterns. Finally, antecedent weighting is processed by salience factors that are morphological pattern rules and preference rules. After the weighting, the most weighted antecedent selection is the coreference resolution.

## 5 Evaluation

This paper devised the method to extend heuristics iteratively for coreference resolution. In figure 3, workbench for iterative heuristic acquisition is described. Coreference Heuristic Extractor has a role to derive heuristics empirically. The context buffer is used for thedata structure in coreference module. It maintains the intermediate results of coreference resolution such as NE-SET and antecedents candidates.

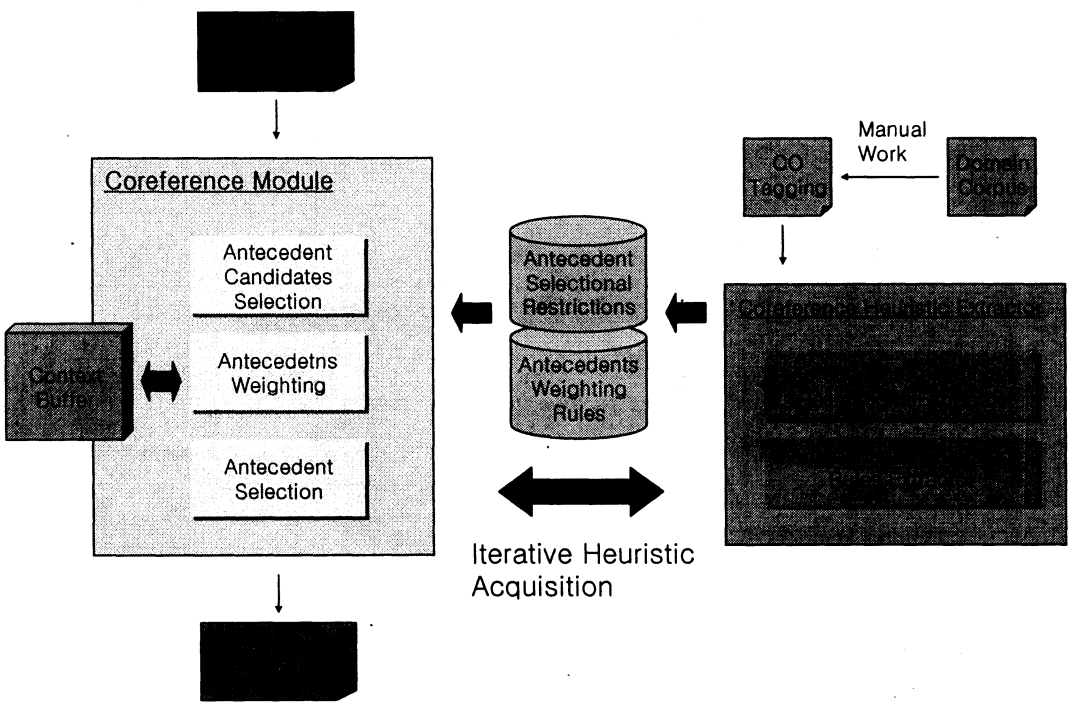


Figure 3. Workbench for Iterative Heuristic Acquisition

In the experiment, we trained our heuristics with 78 articles having 138 anaphoric lexical items. The domains of the articles are economy and performance. We tagged name entities and coreferences in the articles. In this evaluation, we excluded the temporal anaphora and non-referential anaphora. The average of antecedents per the anaphor is 12.45. We tried to test the coreference resolution with only heuristics. Thus, the tagged article has already name entity group and anaphora selection. Therefore, the test depends on the salience factors' weight and conditional rule usage. First test use the conditional rule that is category restriction such as type matching between antecedent and anaphor. Second test don't use it. The result of both tests is described in table 4. We can resolve the coreference at 80% precision even if we use only simple heuristics. However, we assumed the name entity type matching. If we developed a coreference module, the result would lower a little since the semantic compatibility could not be implemented completely.

From the evaluation, we found that the heuristics is not general. In table 4, the weights of test1 and test2 are different. Thus, the best result of one weight is not in the another test. We should change the weight to get the best precision. The source of the trouble is the small set of the training data. However, the coreference phenomenon is very sparseness. It is the enormous work to construct appropriate test set.

In addition, we could not find the reflexive pronoun and syntactic pattern matched coreference. This is also the problem of data sparseness, and the feature of Korean would be the source.

In comparison with foreign research, 89.5% precision, we cannot achieve the better result. The reason of that is first, we use only compact heuristic rules. Second, the training data is too small to cover the general coreference phenomenon. In the future, we try to reduce the gap of the performance.

| *Heuristics* | *Similarity clues* | *check list* | *weight* | |
|---|---|---|---|---|
| | | | *Test1* | *Test2* |
| Morphological pattern rules | Case marker | match | 1 | 1 |
| | | not match | 1 | 0 |
| | Affix | prefix match | 1 | 1 |
| | | suffix match | 1 | 1 |
| | | not match | 0 | 0 |
| | Partial lexical item | lexical pattern match | 1 | 1 |
| | | not match | 0 | 0 |
| Preference rules | Subject/Object | subject antecedent | 1 | 1 |
| | | object antecedent | 1 | 1 |
| | | others | 1 | 1 |
| | Recency | most recency | 1 | 1 |
| | | same sentence | 1 | 1 |
| | | one sentence distance | 1 | 1 |
| | | others | 0 | 0 |
| | Frequency | most frequent antecedent | 1 | 1 |
| | | others | 0 | 0 |
| Conditional rules | Reflexive pronoun | not used | - | - |
| | Syntactic pattern | | | |
| | Category restriction | name entity type match | use | not used |
| Precision | | | 80% (111/138) | 54% (74/138) |

Table 4. heuristics and result in the evaluation

## 6 Conclusion

This paper presents the heuristic-based approach for coreference resolution in Korean. We organized the heuristics to be expanded gradually using the corpus and derived the salience factors of antecedents in Korean. The coreference resolution approach consists of antecedents selection and antecedents weighting. We derived three kinds of salience factors that are used to weight each antecedent of the anaphor. The experiment result shows 80% precision.

In the future work, we will consider the temporal coreference and discourse structure. These are the main causes of the coreference method now. The massive heuristic training and experiment will be processed.

## References


Aone, C., Halverson, L., Hampton, T., and Ramos-Santacruz, M. 1998. "SRA: Description of the IE2 System used for MUC-7," In Proceedings of the Seventh Message Understanding Conference (MUC-7), Columbia, MD.

Baldwin, F. B. 1995. CogNIAC: A discourse processing engine, a dissertation in computer and information science.

Fukumoto, J., Masui, F., Shimohata, M., and Sasaki, M. 1998. "Oki Electric Industry: Description of the Oki System as Used for MUC-7," In Proceedings of the Seventh Message Understanding Conference (MUC-7), Columbia, MD.

Humphreys,K., Gaizauskas, R. Azzam, S., Huyck, C., Mitchell, B., Cunningham, H. and Wilks, Y. 1998. "University of Sheffield : Description of the LaSIE-II system as used for MUC-7," In Proceedings of the Seventh Message Understanding Conference (MUC-7), Columbia, MD.

Huyck, C. 1998. "Description of the American University in Cairo's System Used for MUC-7," In Proceedings of the Seventh Message Understanding Conference (MUC-7), Columbia, MD.

Kehler, A. 1997. "Probabilistic Coreference in Information Extraction," In Proceedings of the Second Conference on Empirical Methods in Natural Language Processing (SIGDAT).

Lappin, S. and Leass, H. 1994. An algorithm for pronominal anaphora resolution. Computational Linguistics, 20(4):535-561.

Lin, D. 1998. "Using collocation statistics in information extraction," In Proceedings of the Seventh Message Understanding Conference (MUC-7), Columbia, MD.

Mitmov, R. 1998. "Robust pronoun resolution with limited knowledge," In Proceedings of the 17th International Conference on Computational Linguistics, COLING-98, Montreal.

Urbanowicz, R. A. and Nettleton, D. J. 1998. "University of Durham: Description of the LOLITA system as used in MUC-7," In Proceedings of the Seventh Message Understanding Conference (MUC-7), Columbia, MD.

Yangarber, R. and Grishman, R. 1998. "NYU: Description of the Proteus /PET system as used for MUC-7 ST," In Proceedings of the Seventh Message Understanding Conference (MUC-7), Columbia, MD.