Original Article

Healthc Inform Res. 2014 October;20(4):272-279.
http://dx.doi.org/10.4258/hir.2014.20.4.272
pISSN 2093-3681 • eISSN 2093-369X

HIR
Healthcare Informatics Research

Korean Anaphora Recognition System to Develop Healthcare Dialogue-Type Agent

Junggi Yang, BS[1], Youngho Lee, PhD[2]

[1]Department of IT Convergence Engineering and [2]IT Department, Gachon University, Seongnam, Korea

Objectives: Anaphora recognition is a process to identify exactly which noun has been used previously and relates to a pronoun that is included in a specific sentence later. Therefore, anaphora recognition is an essential element of a dialogue agent system. In the current study, all the merits of rule-based, machine learning-based, semantic-based anaphora recognition systems were combined to design and realize a new hybrid-type anaphora recognition system with an optimum capacity. **Methods:** Anaphora recognition rules were encoded on the basis of the internal traits of referred expressions and adjacent contexts to realize a rule-based system and to serve as a baseline. A semantic database, related to predicate instances of sentences including referred expressions, was constructed to identify semantic co-relationships between the referent candidates (to which semantic tags were attached) and the semantic information of predicates. This approach would upgrade the anaphora recognition system by reducing the number of referent candidates. Additionally, to realize a machine learning-based system, an anaphora recognition model was developed on the basis of training data, which indicated referred expressions and referents. The three methods were further combined to develop a new single hybrid-based anaphora recognition system. **Results:** The precision rate of the rule-based systems was 54.9%. However, the precision rate of the hybrid-based system was 63.7%, proving it to be the most efficient method. **Conclusions:** The hybrid-based method, developed by the combination of rule-based and machine learning-based methods, represents a new system with enhanced functional capabilities as compared to other pre-existing individual methods.

Keywords: Anaphora Resolution, Anaphora Recognition, Reference Resolution, Dialogue Analysis, Natural Language Processing

Submitted: February 26, 2014
Revised: August 15, 2014
Accepted: August 29, 2014

Corresponding Author
Youngho Lee, PhD
IT Department, Gachon University, 1342 Seongnamdaero, Sujeong-gu, Seongnam 461-701, Korea. Tel: +82-32-820-4506, Fax: +82-32-820-4504, E-mail: lyh@gachon.ac.kr

I. Introduction

To avoid the repetition of any noun, appearing in a sentence in the course of a dialogue, pronouns are often used in subsequent sentences to replace that particular noun. This economy of articulations is a phenomenon commonly observed in human languages. The Anaphora resolution is essentially required in any natural language dealing-based system for a variety of functions, such as information extraction, question and answer, machine translation, document classification, document summarization, dialogue agent systems and the like [1-3]. To correctly analyze the meaning that a text

recalls, the co-relationship and reference relationship of the objects included in the text should be made clear.

The Tapper algorithm, which can extract the step-wise information of documents and the accumulated data of medical organizations by applying the Bernoulli theorem and the Bayesian theorem, was developed for classification of the documents [4]. Thus, it became possible to automatically generate medical terms and abbreviation dictionaries from medical data by using a statistical learning method and logistic regression analysis [5]. A prior study also used filtering and searching methods of information to look up medical records similar to the symptoms of a certain patient [6].

Computerization of laboratory medicines has been used to report, check, and manage all the test results rapidly, which increased the overall efficiency of the examination and treatment processes [7]. In this regard, it is worth mentioning that, though the use of computer-based expert systems has increased in decoding test results and interpretive reports, such processes are still in their rudimentary stages [8,9]. The primary reason that computerized analysis has gained immense popularity is because it enables clinicians to provide more elaborate information on treatments, diagnoses, and healing progress than simple briefings on the resulting test values [10]. In other words, the purpose of making interpretive reports has been not only to reduce any possible error in the interpretation of the test results, but also to be able to select additional tests and treatments, if required [11].

With the introduction of Electronic Medical Records, where all the medical records are saved as computer-based documents instead of the conventional paper-based ones, medical organizations started processing all data that has been accumulated over a long period to extract meaningful information and knowledge from them. The processing of these accumulated data collected through clinical tests and medical studies has enabled the reutilization of old data to share information and has also improved the quality of medical services through the gaining of new knowledge and the development of a rapid decision-making support system [12].

To extract the right meaning of a pronoun for medical information out of a natural language dialogue, it is important to consider the features of the pronoun, to determine anaphora. To understand the meaning of a natural language dialogue, a rule-based method is needed to form a system that can replace the personal and demonstrative pronouns of a Korean language dialogue with personal and demonstrative words. Many researchers have used the machine learning-based method for natural language dealings to check its classification function in a number of studies [13-18]. Machine learning-based methods have also been reported to translate anaphora in the past [19]. However, the machine learning-based method has a limitation that it depends on a small number of data. It also has a difficulty in annotating a variety of information to translate anaphora. On the other hand, an information searching and processing system using meaningful information was also used in the semantic-based method [20-24]. The semantic-based method involves encoding of cultural memories or facts through a graph data structure. One past studies used the semantic-based method to check the possibility of making a meaningful combination between predicates and referent candidates [25,26].

The current study proposes a method of translating anaphora in a dialogue unit of language processing for a dialogue processing system. The proposed method was designed to solve several types of anaphora phenomena. Interestingly, anaphora recognition can appear in a variety of forms, depending upon the various expressions made in the course of a dialogue; however, they are not considered as different phenomena. That is why various types of anaphora, instead of a single type, were considered together at the same time in past studies related to the Korean anaphora recognition systems. In terms of language processing, a better system can be constructed in which a variety of language forms and anaphora phenomena can be analyzed efficiently. However, straightforward application of the abovementioned methods to resolve anaphora has several limitations. Therefore, in the current study, a new system was developed, a so-called hybrid-type method, which combined the rule-based, semantic-based, and machine learning-based methods to achieve better functionality. This new system with a hybrid-based method has the potential to be used actively as a translation engine or dialogue agent in medical information systems.

II. Methods

1. Bases to Anaphora Resolution

People reveal things through verbal expressions. An object can be expressed as 'A', 'it', or 'the previously mentioned one' after an elapse of time. A difference is made in using an explicit representation, such as 'A' and an indirect or generic representation, 'it' and 'the one'. Unlike a verbal expression, which has an explicit representation, an anaphora, a reference, or a substitute, is generally expressed as an indirect representation, and is not understood directly. Therefore, an anaphora should be carefully resolved from expressions, which may have several possibile meanings. Table 1 shows some definite noun phrases and pronouns processed as the scope of anaphora resolution.

Table 1. Definite nouns and pronouns

No	Explicit form	Type	Analytical method
1	It, the one, that one, the aforementioned one	- In case that a referent is inanimate. Yesterday I lost my smartphone in a department store. Can I get it back? - It is not guaranteed that the referent must be a noun (or a noun phrase). Obama is president of the USA. Is there anybody who doesn't know it? *The referent is not a name, but it is expressed in a single sentence.	Machine learning-based
2	The one, the lady, the man, the woman, the boy, the girl, the person, the human-being	- In case a referent is classified as a person.	Rule-based method Semantic method Machine learning method
3	The car, the computer, the thing, the cellphone The laptop computer, the PC, the school	- In case that a referent is not classified as a person. - The word line indicating what a referent is exists in a referred expression.	Rule-based method Semantic method

Table 2. Examples of semantic classes of referents

Indicating semantic scopes of referents.
A: Yesterday I bought [a Louis Vuitton bag] in the [Lotte department store].
(The Louis Vuitton bag: a product, the Lotte department store: a company)
B: Good for you.
A: Do you know the price of it?
B: Maybe, as expensive as about one million won?
A: Wrong. Five hundred thousand.

Table 3. Examples of a rule-based anaphora recognition system

IF (REFERRED_EXPRESSION == a company && REFERENT_CANDIDATE == COMPANY)
THEN INCREASE_MAPPING_SCORE BY 10
IF (REFERRED_EXPRESSION == a film && REFERENT_CANDIDATE == FILM)
THEN INCREASE_MAPPING_SCORE BY 10

2. Structures of Rule/Semantic-Based Anaphora Recognition Systems

The methods of selecting referents by using rule-based and semantic-based analyses were utilized in the study. After priority referent candidates were selected by a noun phrase extractor, the semantic information of individual noun phrases was determined by an entity name recognizer. Table 2 shows an example of semantic classes of referents.

If the semantic scope of a referent candidate is a 'company (a name of a firm or corporation)' and it is identical to the semantic scope of the referent shown by a referred expression, the rule shown in Table 3 means that its mapping score should be increased by 10 or more than other candidates, who are not.

The merit of the rule/semantic-based anaphora recognition method is that it can readily identify referents if the referred expressions have bases that can restrict the semantic scope of referents themselves. On the other hand, a disadvantage is that it is impossible to make correct anaphora recognition if the referred expression is a pronoun, such as 'it'; if the predicate containing 'it' has an object, then 'it' is not in a meaningful combination with the semantic information of the referent candidate.

Figure 1 represents the overall process through which the system detects a given referring expression as an anaphora and a corresponding noun phrase as a referent. For this, it it necessary for each noun or noun phrase to be tagged with a semantic class label using the named entity recognizer and semantic role tagger to check whether a given candidate referent may be used as an object of the predicate co-occurring with the anaphoric expression in which the semantic role attached to a candidate is used to find whether a given referent satisfies a set of semantic constraints. In the last step, the ref-

 http://dx.doi.org/10.4258/hir.2014.20.4.272

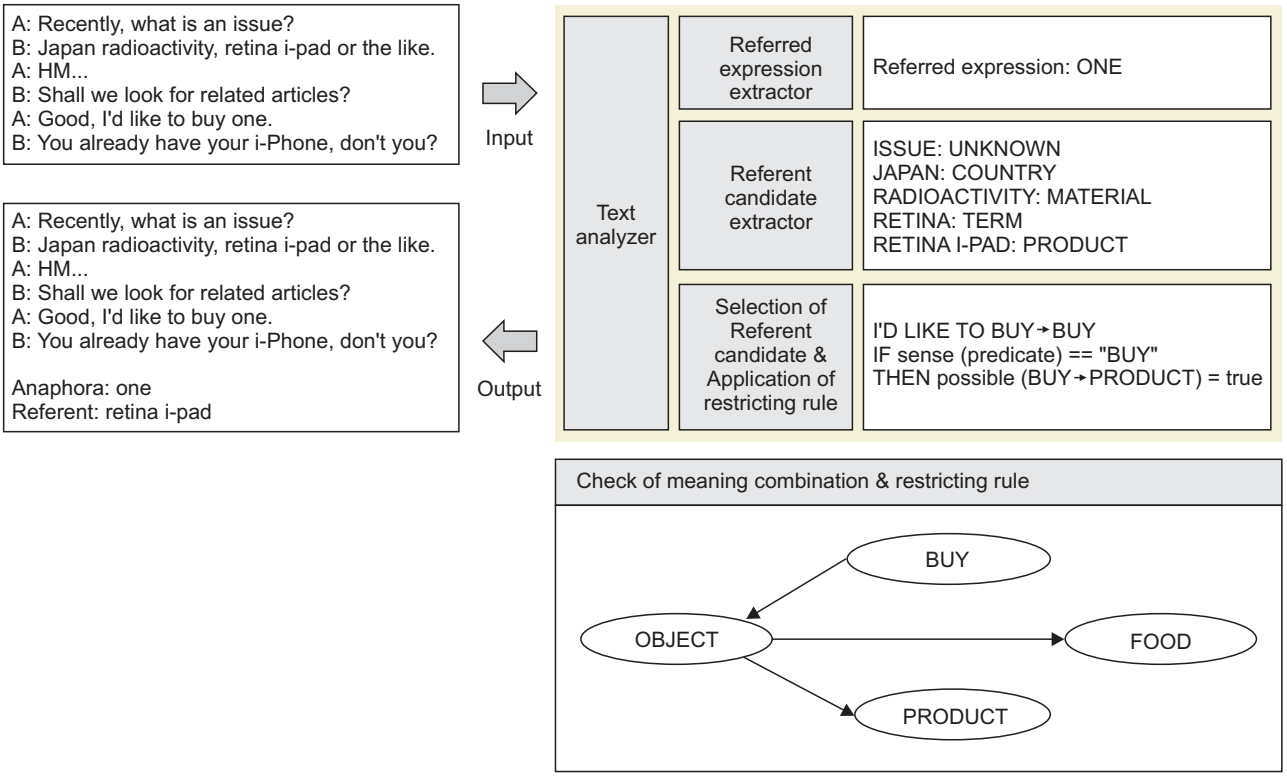

Figure 1. Anaphora resolution system architecture based on rule and semantic anaphora resolution strategies.

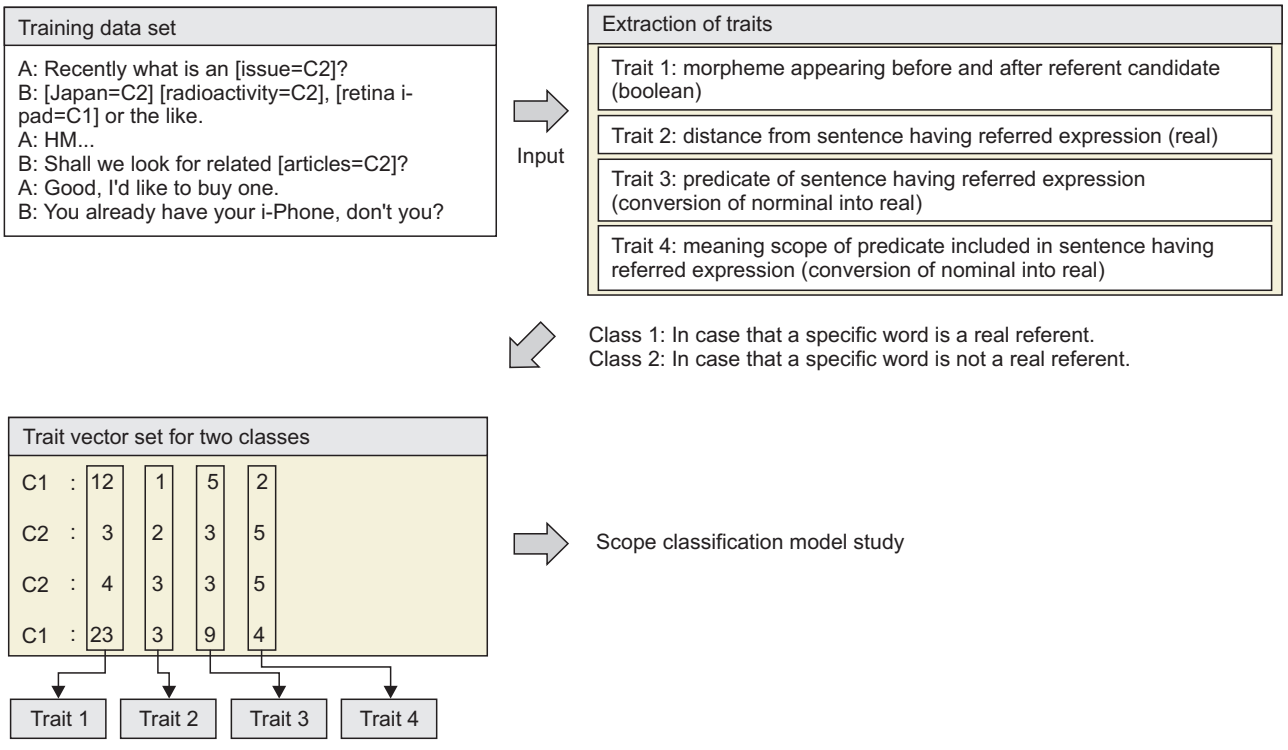

Figure 2. Structure of a machine leaning-based anaphora recognition system.

erent candidate with the highest score in the priority queue is selected as a referernt of the anaphora.

3. Structure of a Machine Learning–Based Anaphora Recognition System

Figure 2 shows the structure of a machine learning-based anaphora recognition system, where the semantic recognition part of the referred expressions attach semantic class labels to the referent candidates, which are selected by a noun phrase extractor. For example, a semantic label 'country' is attached to the word 'Japan'. At the same time, the scope classifier of referent candidates trained by means of machine learning is used to determine the scope of the referent candidates. The scope classifier is used to attach class labels {C1=referent (REF), C2=not referent (NONREF)} to specific referent candidates.

Training data has comprises a dialogue event representation set whose components are short spoken sentences in which the actual referent is tagged C1, and other non-referents are tagged C2. Then, each referent candidate is analyzed to extract a feature vector, forming an actual training data set represented in a quantitative real or boolean value array for each referent candidate. Here, 90% of the training data set was used for parameter estimation of our classifier using a sequential minimal optimization algorithm, and 10% of the total data set was used for system evaluation. Table 4 illustrates an example of training data.

One hundred training data items were used to learn the scope classifier. An example of the training data is shown in Table 4.

The training data indicate the ranges of referents and referred expressions. Labels C1 and C2 are automatically attached to the column of the referent letters and that of the other referent candidates, respectively, to extract the feature vectors of the individual referents. The set of the extracted training data is used to assume an optimum model parameter.

The feature vector has four dimensions, starting with how far a given referent candidate is located from an anaphoric expression, namely, 'recency'. The second feature is called 'semantic distance' between an anaphora and a given referent candidate. Semantic distance quantifies how semantically close a given referent candidate is to an anaphora in question by checking the semantic class of the referent candidate. For example, if a candidate is tagged 'FOOD' and the anaphoric expression is categorized into the food-referrring class, the semantic distance is assumed to be short. The third feature is the possibility of a referent candidate being used as an argument of the predicate co-occurring with the target anaphora, thus being a boolean data type. The last feature is whether a referent candidate is equivalent to the semantic class of the noun within the anaphoric expression. If the semantic class is identical, the referernt candidate is assumed to be semantically equivalent to the noun within the anaphora, thus being 1.

Table 5 shows an example of a distance between sentences having referents and referred expressions, where there are two referent candidates, <cellphone> and <bag>.

The feature vector of the above-mentioned individual instances is four-dimensional. The first feature is the distance from the sentence of the referred expression to that of the referent candidate. The second feature is to convert the semantic class of the referents into a type of real number. The third feature is to convert the semantic type of the predicate of the sentence containing the referred expression into a type of real number. The fourth feature is to convert the semantic type of the predicate of the sentence containing the referent, into a type of real number.

Several algorithms, such as naïve Bayesian, sequential minimal optimization, and maximum entropy, were utilized to assume the parameters, and the Java library of WEKA (Waikato Environment for Knowledge Analysis), an open-source machine learning platform, was also utilized to study

Table 4. Example of training data

<data>
dialogue segment-5
SEN=Have you been to Busan?
SEN=No, not yet.
SEN=Why haven't you been there?
SEN=It snowed too much there.
#Meta Information
REFERENT=Busan
ANAPHORA=the place
Extra Information
USE=TRAINING
</data>

Table 5. Distance from referred expressions

A: [My cellphone] is lost. (My cellphone : distance=4)
B: Where did you lose? (The sentence where the object, "it" is omitted.)
A: I even forgot where I lost it.
B: Did you check the inside of your bag? (Your bag: distance=1)
A: Yes, [it] is gone, so I feel bad. (The statement where the referred expression "it" exists: the standard position to measure a distance)

http://dx.doi.org/10.4258/hir.2014.20.4.272

the new system in this research.

4. The Hybrid Anaphora Recognition System

Figure 3 shows the structure of the hybrid anaphora recognition system developed in this study. In the proposed system, a dialogue text is input to select three or four previous sentences having a referred expression, such as 'it' or 'the person'.

A list of referent candidates, having the possibility of indicating the referred expression at the selected dialogue segment is extracted first. A noun phrase extractor (NP chunker) is used in the course of this extraction process, and the NLP-HUB Parser of Korea Advanced Institute of Science and Technology (KAIST) is included in the noun phrase extractor. The NLP-Hub is a Korean parser developed by the KAIST Semantic Web Research Center, and it is used to identify information in the predicate part of a sentence. The NP chunker is a language analyzing tool that searches for noun phrases in a sentence chunker. In this study an NE recognizer was built to analyze noun phrases like the NP chunker. Noun phrases are found because the substrings of a sentence consist of a noun phrase or more than a noun, and referent candidates could be greater than the actual number due to misrecognition of computer programs, although a noun phrase refers to a single object. For this reason, noun phrases need to be pre-processed for the designed system to recognize a noun phrase as a single identity by performing NP chunking.

The hybrid-based system contains three methods, namely, machine learning-based, rule-based and semantic-based methods, to process the referred expressions classified as types 1, 2, and 3 as defined in Table 1.

III. Results

Functional comparisons were made depending on the algorithms applied to machine learning and the size of training

Table 6. Comparison depending on machine learning algorithms

Algorithm category	Precision rate (%)
Sequential Minimal Optimization	66.8
Maximum entropy	64.7
Naïve Bayesian	61.3

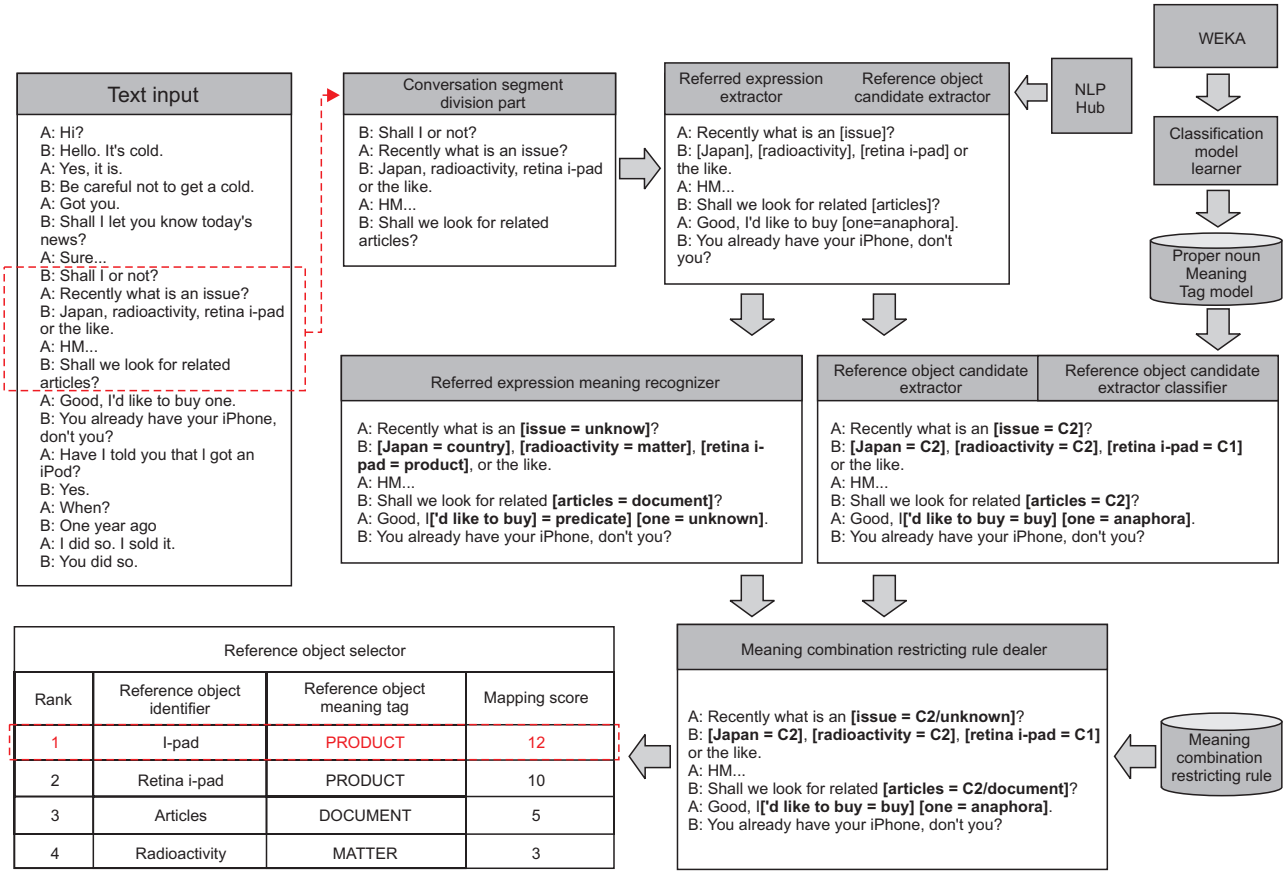

Figure 3. Structure of the hybrid anaphora recognition system. NLP: Natural Language Processing, WEKA: Waikato Environment for Knowledge Analysis.

Table 7. Functional comparison of systems depending on the size of the entity name recognizer (unit: %)

Method	50 instances	100 instances	150 instances	200 instances
Rule-based	53.9	54.9	59.8	56.8
Machine learning-based (by SMO)	65.6	66.8	66.8	66.8
Hybrid-based	69.3	71.2	74.5	74.5

SMO: Sequential Minimal Optimization.

data of an entity name recognizer.

To estimate its functional difference from that of the hybrid-based system, the machine learning-based method was employed solely using the learning algorithms, which are related only to machine learning-based anaphora recognition. As shown in Table 6, no significant difference was observed in the precision rate, depending on the learning algorithms. However, it is assumed that the feature vector failed to reflect the whole structure of the targeted anaphora recognition domain.

Due to dependence on the semantic scope classifying capacity of the nouns/noun phrases of the entity name recognizer, an experiment was conducted in which the size of the training data used in the NER model was treated as an independent variable. The results shown in Table 7 indicate that the precision rate of the anaphora recognition model increased log-linearly in proportion to the size of the training data of the entity name recognition model. At the training time of the entity name recognition model, its internal features were disregarded to utilize only the feature vector of its neighboring morphemes, so that the function of the entity name recognition model can be proved unfavorable.

The system evaluation results show that the rule-based method achieves 54.9% precision, whereas our suggested system, which is a hybridization of rule and machine learning, achieves 63.7% precision as shown in Table 8. In terms of sensitivity and specificity, the results are consistent.

The purpose of the current research was to check a hypothesis that the rule-based and machine learning-based algorithms could impart positive effects to the functionality of the system. As predicted, it was confirmed that anaphora recognition precision rate could be improved as compared to systems using only a rule or classification model singly or that calculated the mapping table score in the hybrid-based method.

Table 1 also showed that the rule-based system achieves superior function of types 2 and 3, and the machine learning-based system achieves superior function of type 1. Since the hybrid system is designed to take a wider context into consideration than rule-based system, it is natural that it shows better performance in anaphora resolution.

Table 8. Correct proportion of hybrid-based system

	Sensitivity (%)	Specificity (%)
Correct proportion	70.6	91.9

IV. Discussion

In the current study, a pronoun-type anaphora recognition system was developed. To optimize the mapping scoring model for referent candidates and referred expressions, a hybrid-based method was generated. This hybrid-based method, developed by the combination of rule-based and machine learning-based methods, represents a new system with enhanced functional capabilities as compared to other pre-existing individual methods.

The anaphora resolution system has great potential for use in the medical domain in that it is an essential component in semantic analysis systems; it can be used either for deep semantic analysis of medical free-text to extract information and knowledge in an automatic manner or to design a dialogue system which can provide a healthcare consulting service.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

This work was supported by the Gachon University research fund of 2014 (No. GCU-2014-R028).

References

1. Peral J, Rodriguez AF. Translation of pronominal anaphora between English and Spanish: Discrepancies and evaluation. J Artif Intell Res 2003;18:117-47.

 http://dx.doi.org/10.4258/hir.2014.20.4.272

2. Zhou X, Han H, Chankai I, Prestrud A, Brooks A. Approaches to text mining for clinical medical records. Proceedings of the 2006 ACM Symposium on Applied Computing; 2006 Apr 23-27; Dijon, France. p. 235-9.
3. Song MH, Kim SH, Park DK, Lee YH. A multi-classifier based guideline sentence classification system. Healthc Inform Res 2011;17(4):224-31.
4. Chakrabarti S, Dom B, Agrawal R, Raghavan P. Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. VLDB J 1998;7(3):163-78.
5. Chang JT, Schütze H, Altman RB. Creating an online dictionary of abbreviations from MEDLINE. J Am Med Inform Assoc 2002;9(6):612-20.
6. Hsu WH. Activities of the KSU Bioinformatics and Medical Informatics (BMI) Working Group, 2001-2002. Manhattan (KS): Department of Computing and Information Science, Kansas State University; 2001.
7. Wycoff DA, Wagner JR. Distributed laboratory computing: integration of a laboratory computer into a hospital information system. Am J Clin Pathol 1978;70(3):390-9.
8. Burke MD. Clinical laboratory consultation: appropriateness to laboratory medicine. Clin Chim Acta 2003;333(2):125-9.
9. Kratz A, Soderberg BL, Szczepiorkowski ZM, Dighe AS, Versalovic J, Laposata M. The generation of narrative interpretations in laboratory medicine: a description of service-specific sign-out rounds. Am J Clin Pathol 2001;116 Suppl:S133-40.
10. Lim EM, Sikaris KA, Gill J, Calleja J, Hickman PE, Beilby J, et al. Quality assessment of interpretative commenting in clinical chemistry. Clin Chem 2004;50(3):632-7.
11. Kratz A, Laposata M. Enhanced clinical consulting: moving toward the core competencies of laboratory professionals. Clin Chim Acta 2002;319(2):117-25.
12. Hripcsak G, Friedman C, Alderson PO, DuMouchel W, Johnson SB, Clayton PD. Unlocking clinical data from narrative reports: a study of natural language processing. Ann Intern Med. 1995;122(9):681-8.
13. Lee SK, Kang BY, Kim HG, Son YJ. Predictors of medication adherence in elderly patients with chronic diseases using support vector machine models. Healthc Inform Res 2013;19(1):33-41.
14. Kim S, Kim W, Park RW. A comparison of intensive care unit mortality prediction models through the use of data mining techniques. Healthc Inform Res 2011;17(4):232-43.
15. Wang S, Zhu W, Liang ZP. Shape deformation: SVM regression and application to medical image segmentation. Proceedings of the 8th IEEE International Conference on Computer Vision; 2001 Jul 7-14; Vancouver, Canada. p. 209-16.
16. Do TN, Poulet F. Incremental SVM and visualization tools for bio-medical data mining. Proceedings of the European Workshop on Data Mining and Text Mining for Bioinformatics; 2003 Sep 23; Dubrovnik, Croatia. p. 14-9.
17. Li S, Fevens T, Krzyzak A. A SVM-based framework for autonomous volumetric medical image segmentation using hierarchical and coupled level sets. Int Congr Ser 2004;1268:207-12.
18. Amaral IF, Coelho F, da Costa JF, Cardoso JS. Hierarchical medical image annotation using SVM-based approaches. Proceedings of the 10th IEEE International Conference on Information Technology and Applications in Biomedicine; 2010 Nov 3-5; Corfu, Greece. p. 1-5.
19. Evans R. Applying machine learning toward an automatic classification of it. Lit Linguist Comput 2001;16(1):45-57.
20. Mojsilovic A, Gomes J. Semantic based categorization, browsing and retrieval in medical image databases. Proceedings of the 2002 International Conference on Image Processing; 2002 Sep 22-25; Rochester, NY. p. 145-8.
21. Lee Y, Patel C, Chun S, Geller J. Compositional knowledge management for medical services on semantic web. Proceedings of the 13th international World Wide Web Conference on Alternate Track Papers & Posters; 2004 May 17-20; New York, NY. p. 498-9.
22. Cimino JJ. Auditing the Unified Medical Language System with semantic methods. J Am Med Inform Assoc 1998;5(1):41-51.
23. Yun JH, Ahn SJ, Kim Y. Development of clinical contents model markup language for electronic health records. Healthc Inform Res 2012;18(3):171-7.
24. Hwang KH, Lee H, Choi D. Medical image retrieval: past and present. Healthc Inform Res 2012;18(1):3-9.
25. Arch-int N, Arch-int S. Semantic ontology mapping for interoperability of learning resource systems using a rule-based reasoning approach. Expert Syst Appl 2003;40(18):7428-43.
26. Choi S, Choi J, Yoo S, Kim H, Lee Y. Semantic concept-enriched dependence model for medical information retrieval. J Biomed Inform 2014;47:18-27.