Int. J. Advance Soft Compu. Appl, Vol. 7, No. 1, March 2015
ISSN 2074-8523

Incremental Discriminating Method for Acronyms in Heterogeneous Resources

Do-Heon Jeong, Jangwon Gim* and Hanmin Jung

Department of Computer Intelligence Research, KISTI
Daejeon, South Korea
e-mail: {heon, jangwon, jhm} @kisti.re.kr
*Corresponding Author: Jangwon Gim

Abstract

In this paper, a method is proposed to discriminate acronyms and their full names or expansions in scientific and technical literature abstracts by learning Wikipedia definition statements. Through this study, we aim to verify the effective utilization of an open knowledge base in the knowledge processing of domain-specific fields. Experimental results confirm that a noun phrase (NP)-type feature has better performance than a noun (NN) type feature in terms of precision rate. On the contrary, the results of measuring query response rate indicate that a single NN-type feature has better performance than an NP-type feature. We also verify that additional collocation information can contribute to improve the response rate. This study is mainly divided into three parts: 1) a process of sense discrimination is classified into many steps according to feature types; 2) the measured results are combined and processed; and 3) a data fusion-based incremental approach is proposed for sense discrimination. Through the method, we can adjust a precision rate to a certain level while considering classifier response rate.

Keywords: *acronym, Naïve Bayesian, sense disambiguation, text mining, classification, data fusion, response rate, Wikipedia.*

1 Introduction

Open knowledge bases such as Wikipedia and Freebase are developed based on the collaborative activities of Internet users. Common knowledge repositories that contain reliable high-quality information have provided many opportunities to numerous researchers to exchange and utilize knowledge. Moreover, knowledge bases are increasingly applied in knowledge engineering, text mining, linked data-based data sharing, and data analysis based on big data, which has recently been emphasized [1-3].

The purpose of this study is to analyze word senses of domain-specific scientific articles based on common knowledge built on Wikipedia, which is a typical open knowledge base. Accordingly, we determine the possibility of practical application in the area of knowledge bases. In particular, after extracting acronyms widely used to represent terminologies in scientific and technical literatures, a discrimination test is performed to estimate the expansion of acronyms practically by learning definitions in Wikipedia. Through word discrimination performance evaluation, we determine whether the sense discrimination process of heterogeneous resources using open knowledge bases is effective in various application studies, such as text mining and analysis.

In Section 2, previous studies related to our study are described; data construction and an experiment environment are explained in Section 3. In Section 4, performance is evaluated through a variety of experiments, and relevant opinions are presented. Finally, the contributions of this study are briefly described, as is future research in Section 5.

2 Motivation

Jeong et al. [4, 5] presented a complex relationship between acronyms and their expansions extracted from scientific and technical literatures and Wikipedia, and discussed a sense discrimination issue in heterogeneous data environments (Fig. 1). They proposed combined performance in terms of the response rate and precision of queried input for word discrimination through several preliminary tests. This study also attempts to thoroughly measure performance from these aspects. Jeong et al. [6] performed a comparative study on the factor characteristics that can improve precision and reproducibility, and measured performance among stemmized nouns, noun phrases, and a fusion method that combines both, thereby proposing that the performance of F1 measure and break-even point (BEP) is stable when a simple noun (NN)-type feature is used. Their study results indicated that the precision rate was about 90% with regard to semantic interpretation over heterogeneous resources with two different characteristics, thus leading to significantly high word sense discrimination performance.

After reviewing previous studies, this study focuses on the response rate that indicates classifier response. In Wikipedia, response rate is defined as follows [7]. The response rate in survey research refers to the number of people who answered the survey divided by the number of people in the sample. In this paper, it means the number of acronym-expansion pairs which responded to the whole test document.

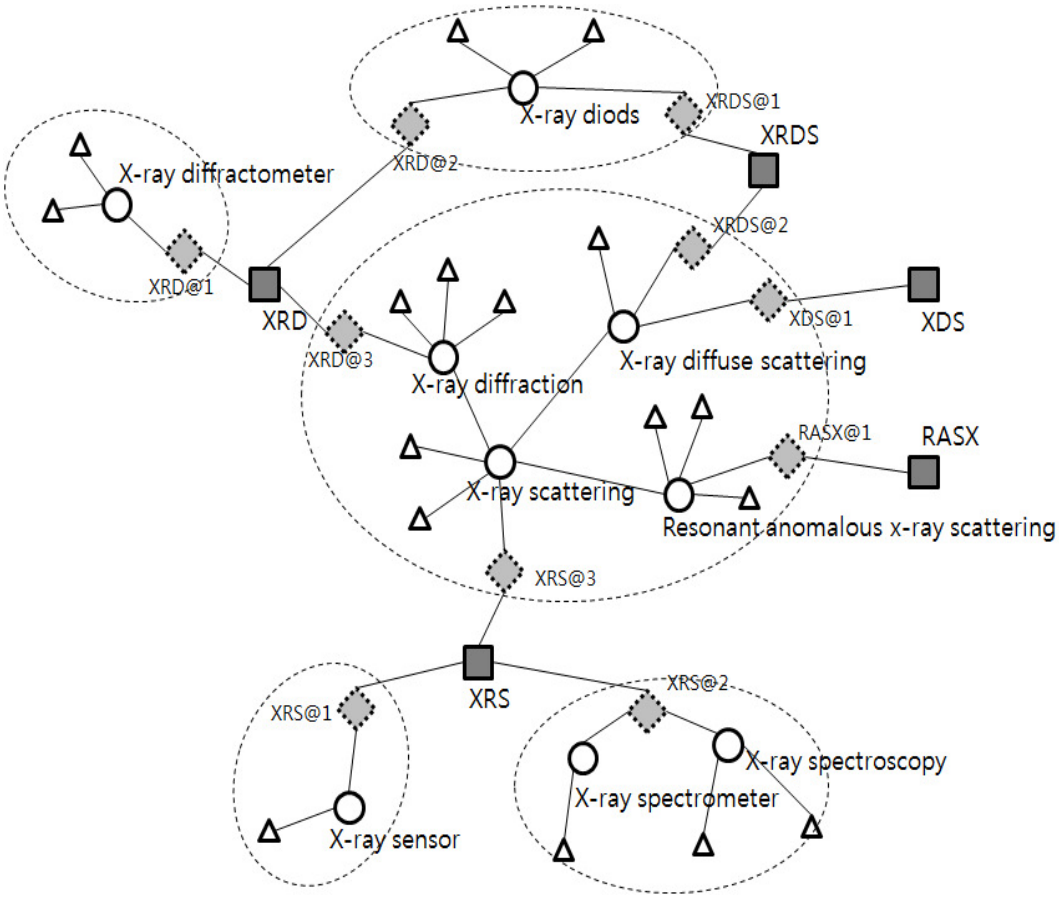


Fig. 1 Semantic network of acronyms with multiple expansions [4, 5]

$$\text{Precision} = \frac{acronym - expansion\ pairs\ found\ and\ correct}{acronym - expansion\ pairs\ found} \quad (1)$$

$$\text{Response} = \frac{acronym - expansion\ pairs\ answered}{total\ acronym - expansion\ pairs\ tested} \quad (2)$$

Hence, the important factor for classifier performance in the sense discrimination environment is determining a method to reduce non-response. In other words, it is important to measure the areas excluded from the first performance measurement because of sparsity, which is generated by lack of clues in response to a requested query. In previous studies, precision rate achieved a maximum of 90% in the discriminable data area, whereas it decreased significantly when indiscriminable areas were considered. In terms of practical perspective, increasing the processing performance of indiscriminable information because of the data sparsity problem

is crucial. Hence, managing user queries through latent semantics rather than exact string values in terms of information search perspective is consistent with the purpose of using a latent semantic index (LSI) based on singular value decomposition (SVD) [8].

3 Experimental Setting

3.1 Data collection

First, keywords associated with acronym information were acquired from the abstract parts of Wikipedia and NDSL (http://www.ndsl.kr/), S&T scholarly service. Then data collection is divided into a training set from Wikipedia only, and a test set from NDSL only. We eliminated acronyms out of the range of Wikipedia in advance (see Fig.2).

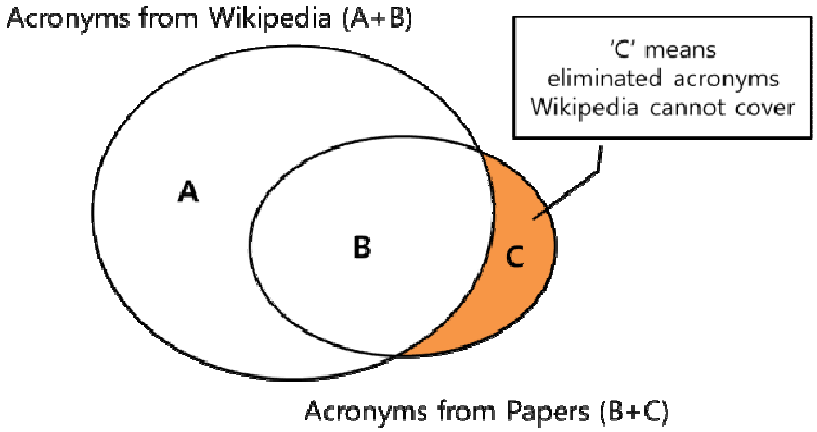


Fig. 2 Specific acronym relationships between the representative resources

For the processing of feature selection, we extracted two types of terms, which were single noun type (NN) and noun phrase type (NP) and we used the Porter stemmer [9] as well as the Stanford POS tagger [10]. Collocation information (CL) was also built from the resources. The statistics are given in Table 1.

Table 1: Data sets and statistics

	Wikipedia (training set)	**NDSL (test set)**
Range (see Fig.2)	A+B	B
Documents	108,236	102,108
Acronyms	33,101	11,889

3.2 Classification method

We use a Naïve Bayesian classifier based on a word sense disambiguation technique (WSD) as a baseline classification referring to the previous study [5]. $P(exp_j)$ is a probabilistic value of the frequency number of all expansions with a specific acronym to the total frequency number of the acronym and $P(clue_m)$ is a probabilistic value of the frequency number of a clue with a specific expansion to the total frequency number of the expansion. Finally, a Naïve Bayesian classifier gives a correct expansion to the specific acronym by using the final theorem, *Decide(exp)*.

$$\mathrm{P}(exp_j) = \frac{freq(exp_j)}{freq(acr_i)} \tag{3}$$

$$\mathrm{P}(clue_m) = \frac{freq(clue_m)}{freq(freq_j)} \tag{4}$$

$$\mathrm{Decide(exp)} = arg_{exp_j \in Exp_{Set(acr_i)}} \max \left[\log\left(1 + p(exp_j)\right) + \sum_{m=1}^{n} \log\left(1 + P(clue_m)\right) \right] \tag{5}$$

4 Experiments

4.1 Trend of precision according to the number of expansions

Prior to the main experiment, because the number of expansions per acronym varies greatly, changes in the overall micro-averaged precision according to this number were measured. As listed in Table 2 and shown in Fig. 3, performance decreased because the number of cases to be found became large as the number of categories (or expansions) to be discriminated increased. A maximum of 410 expansions can be generated because the A+B area of the entire Wikipedia was applied for learning, as shown in Fig. 2.

Table 2: Comparison of precision according to the number of expansions per acronym

# of Expansion (per Acronym)		2	3~4	5~11	12~23	24~39	40~410
NP	precision	98.39	95.99	94.61	88.79	84.22	73.39
	# of feature	6,212	6,211	6,306	6,701	4,004	6,110
NN	precision	94.72	85.01	76.87	69.46	39.17	49.22
	# of feature	14,830	13,648	18,300	24,782	13,752	14,702
CL	precision	93.83	79.33	66.63	52.93	24.77	37.62
	# of feature	15,403	14,254	18,586	25,035	13,997	14,823

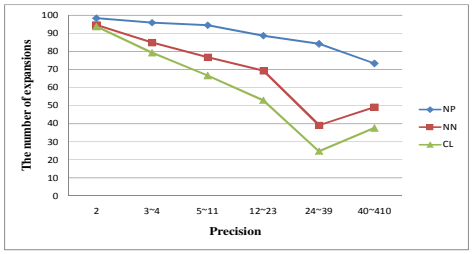


Fig. 3 Performance graph of precision according to the number of expansions per acronym

4.2 Performance trend of precision and response rate

In Fig. 4, the categories “Precision (incl. 1S)” and "Response (incl. 1S)” have only one expansion per acronym so that data of 100% precision are included in the measurement during the sense discrimination process. These data are included for statistical reference because they are frequently found in practical data. The number of test documents for the experiment is 102,108. When acronyms with only one expansion are included, the number will be 140,239.

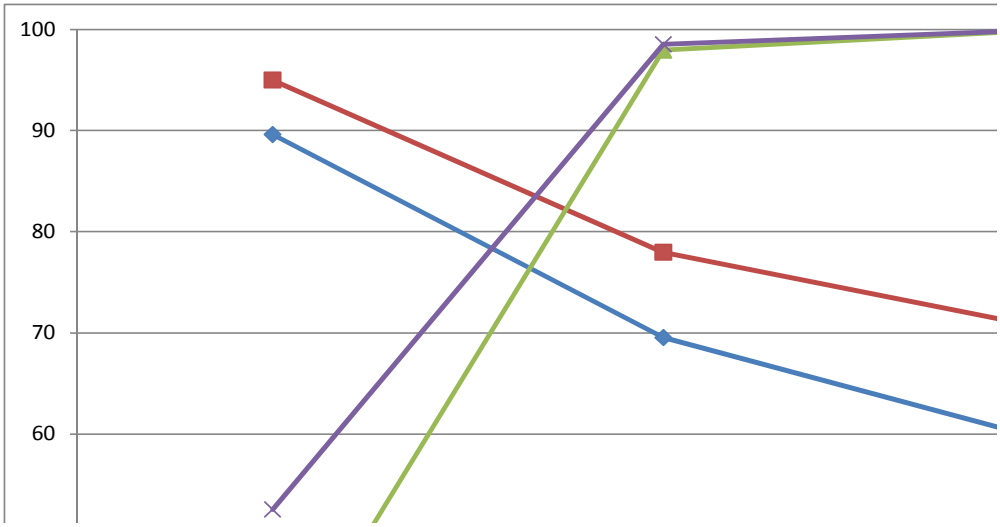
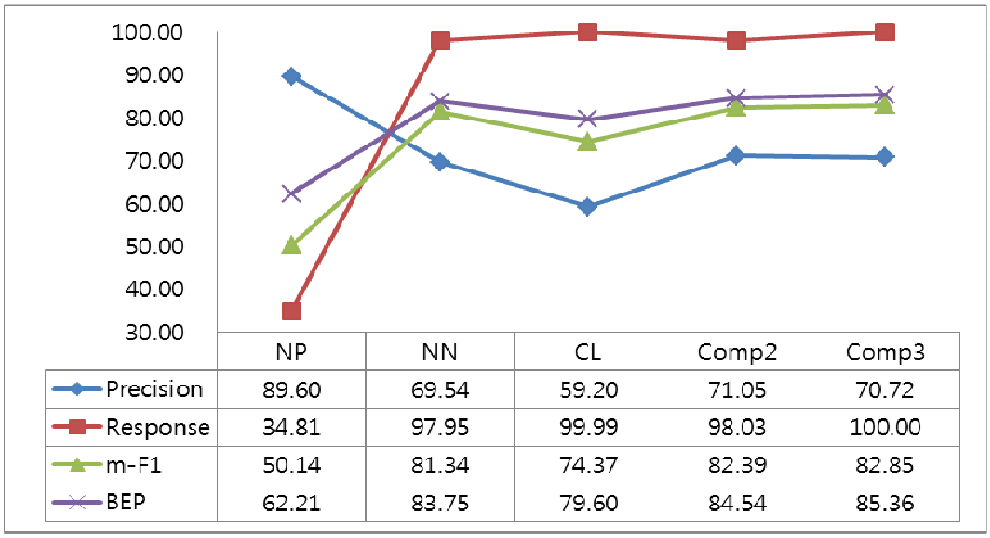


Fig. 4 Performance trend of precision and the response rate



	NP	NN	CL	Comp2	Comp3
Precision	89.60	69.54	59.20	71.05	70.72
Response	34.81	97.95	99.99	98.03	100.00
m-F1	50.14	81.34	74.37	82.39	82.85
BEP	62.21	83.75	79.60	84.54	85.36

Fig. 5 Comparison of overall performance according to feature types and data fusion

(※Comp2: NP+NN, Comp3: NP+NN+CL)

In this experiment, micro-averaged precision and the response rate are prioritized for measurement based on the three feature types of NN-type, NP-type, and CL-type. We used the collocation information (CL) adjacent (within window size 3) to the basic features as additional information. In addition, modified F1 and BEP are used for overall measurements. Modified F1 is a measured value in which recall is replaced with response in the existing F1 score. To comprehensively

improve the performance of measured data, the composite scores of two types are added using the data fusion technique. The Comp2 type combines the NN+NP data, whereas the Comp3 type combines the Comp2+CL (NN+NP+CL) data (refer to Fig. 5).

4.3 Discussion

In summary, the experimental results indicate that the NP-based type provides discrimination based on noun phrases and the precision is high when matching is performed successfully, whereas the response rate is significantly low because of high sparsity. On the other hand, the NN-based type has a significantly high response rate, whereas its precision is comparatively low. The CL-type has lower precision, but a higher response rate than the NN-based type after surrounding information is added. Furthermore, after data fusion, Comp2 and Comp3 have an intermediate feature of NP and NN, respectively, subsequently exhibiting stable performance. Reflecting these characteristics, classifier features can be controlled by performing various steps. To increase performance based on strict matching, only NP information of the first step is required. In addition, to increase the response rate to user queries, either Comp2 that uses NN information via the NP step or Comp3 that uses CL information as well as NN information can be selected. However, given that CL information has a relatively large data amount and long processing time, it is unlikely to be effective in improving the classifier performance.

5 Conclusion

In this paper, we have aimed to find the most efficient method to disambiguate the sense of acronyms in scholarly papers under the use of the open knowledge base, Wikipedia. The following results were derived through experiments that consider various feature types. First, a text mining technique for processing domain-specific information that utilizes open knowledge bases was found to be highly effective. It was found that performance degradation that occurs frequently when heterogeneous resources are used was insignificant when this technique was used, thereby providing valuable implication in terms of practical perspectives. Second, our study suggested that classifier features can be variably controlled through an incremental approach that uses various feature types and data fusion. In addition, further studies will be conducted constantly to increase discrimination performance and reveal and share generated data.

ACKNOWLEDGEMENTS

This work was supported by the IT R&D program of MSIP/KEIT. [2014-044-024-002, developing On-line Open Platform to Provide Local-business Strategy Analysis and User-targeting Visual Advertisement Materials for Micro-enterprise Managers]

References

[1] A. Foharolli, “Word Sense Disambiguation Based on Wikipedia Link Structure,” In the Proceedings on Semantic Computing (ICSC 2009), pp. 77-82, 2009

[2] K. Bollacker, R. Cook, and P. Tufts, "Freebase: A shared database of structured general human knowledge," In the Proceedings on Artificial Intelligence (AAAI 2007), pp. 1962-1963, 2007

[3] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "Dbpedia: A nucleus for a web of open data," The semantic web, Springer Berlin Heidelberg, pp. 722-735, 2007

[4] D.H. Jeong, M. Hwang, and W.K. Sung, "Generating Knowledge Map for Acronym-Expansion Recognition," In the Proceedings on U- and E-Service Science and Technology (UNESST 2011), pp. 287-293, 2011

[5] D.H. Jeong, M. Hwang, J, Kim, H. Jung, and W.K. Sung, "Acronym-Expansion Recognition based on Knowledge Map System," Information, An International Interdisciplinary Journal, 12(A) pp. 8403-8408, 2013

[6] D.H. Jeong, J. Gim, and H. Jung, "Comparative study on disambiguating acronyms in the scientific papers using the open knowledge base," In the Proceedings on APIC-IST 2014, pp.369-371, 2014

[7] Response Rate (Wikipedia), http://en.wikipedia.org/wiki/Response_rate/

[8] S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman, "Indexing by latent semantic analysis," Journal of the American Society of Information Science, 41(6):391–407. 1990

[9] PyStemmer, https://pypi.python.org/pypi/PyStemmer/1.0.1

[10]Stanford Log-linear Part-Of-Speech Tagger, http://nlp.stanford.edu/software/tagger.shtml