Computing and Informatics, Vol. 33, 2014, 503–517

ACRONYM-EXPANSION DISAMBIGUATION FOR INTELLIGENT PROCESSING OF ENTERPRISE INFORMATION

Myunggwon Hwang, Do-Heon Jeong[*], Jinhyung Kim
Sa-Kwang Song, Hanmin Jung

Korea Institute of Science and Technology Information (KISTI)
245 Daehak-ro, Yuseong-gu
305-806 Daejeon, Korea
e-mail: {mgh, heon, jinhyung, esmallj, jhm}@kisti.re.kr

Sajjad Mazhar

Korea University of Science and Technology (UST)
217 Gajung-ro, Yuseong-gu
305-350 Daejeon, Korea
e-mail: ms@kisti.re.kr

Abstract. An acronym is an abbreviation of several words in such a way that the abbreviation itself forms a pronounceable word. Acronyms occur frequently throughout various documents, especially those of a technical nature, for example, research papers and patents. While these acronyms can enhance document readability, in a variety of fields, they have a negative effect on business intelligence. To resolve this problem, we propose a method of acronym-expansion disambiguation to collect high-quality enterprise information. In experimental evaluations, we demonstrate its efficiency through the use of objective comparisons.

Keywords: Acronym-expansion disambiguation, business intelligence, enterprise information, text processing

[*] Corresponding author

1 INTRODUCTION

The ultimate aim of information processing is changing, from providing retrieval convenience to helping firms earn actual profits. Information processing has been accomplished by analyzing documents that deal with technical information, while keeping this new aim in mind. These changes have led to the development of business intelligence as a new business information-processing field, by which firms can garner competitive market advantages and new opportunities. Various organizations have sought to undertake business intelligence for scholarly and commercial reasons. Business intelligence systems that provide such services as decision support systems [1] do not comprise entirely new fields; rather they extend and adopt existing techniques – such as those relating to reporting, online analytical processing, analytics, data mining, processing mining, complex event processing, business performance management, benchmarking, text mining, and predictive analytics – by advancing and optimizing the key features therein. Many studies address the methods by which specific features can be advanced [5, 6, 17, 19, 21, 10]; ultimately, the fundamental of these studies is to extract reliable technical terms by processing documents that include business information. These documents consist mainly of papers, patents, and news articles.

This study addresses business intelligence [1, 20, 18] while focusing on understanding several expression types of technical terms. Technical terms are expressed generally as noun phrases, and these mainly have their own acronyms. Documents often contain acronyms; each is typically introduced once in its full, expanded form, and throughout the balance of the document, the acronym is used. We developed AcroDic 1.0[1] in previous work [3], and so we know that the dictionary contains a maximum of 466 individual expansions for the acronym "SC", and that WordNet[2] has 13 meanings for the word "twist" [4]. This suggests that acronyms can be much more ambiguous than general terms. Indeed, ambiguity is a considerable obstacle when constructing reliable terms for use in business intelligence, but most research focuses on term extraction and general word-sense disambiguation. Figure 1 shows the research trends pertaining to the terms "acronym" and "terminology", between 1991 and 2009; these results were acquired from IEEE Xplore.[3]

As shown in Figure 1, there has been little research on acronyms. If it were possible to pinpoint appropriate expansions for acronyms that have appeared in documents, business intelligence initiatives could generate highly refined, quality information for users. Taking this as our motivation, we propose in this study a method for acronym-expansion (AE) disambiguation. To determine the appropriate expansion for a given acronym, we employ a learning method based on Naive Bayesian classifiers and analyze the results that use single nouns (NN), noun phrases

[1] AcroDic (Acronym Dictionary): http://steak.kisti.re.kr/acrodic/, http://johnnie.kisti.re.kr/

[2] WordNet (A lexical database for English): http://wordnet.princeton.edu/

[3] IEEE Xplore: http://ieeexplore.ieee.org/search/advsearch.jsp?tag=1

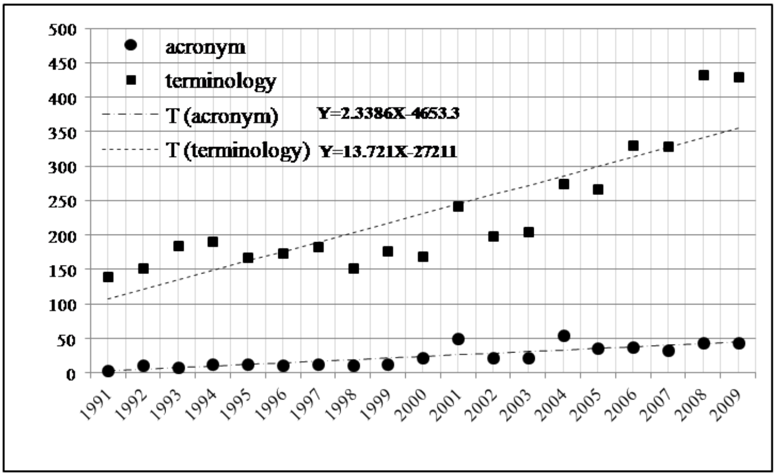


Figure 1. Research trends (study counts) regarding "acronym" and "terminology"

(NP), and both noun types together (NN + NP) as clues. Moreover, the results are compared to those within a baseline, to derive an objective performance evaluation.

This paper is organized as follows. Section 2 describes several previous studies that relate to this area of research. In Section 3, we explain the specific learning processes inherent in AE disambiguation, and provide examples thereof. Furthermore, in Section 4, we present our experimental results, as well as the results of our performance evaluation. Finally, in Section 5, we summarize our research and outline possible directions for future research.

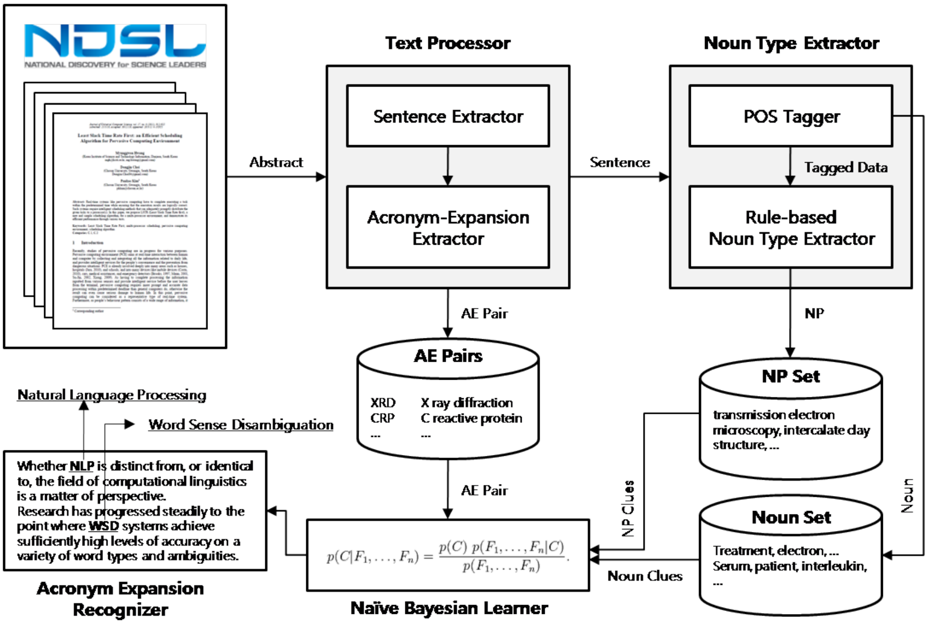
2 LITERATURE REVIEW

Compared to other fields in natural language processing (NLP), limited research has been done on acronyms. In addition, most of the previous studies limit themselves to the extraction of acronyms and their expansions from raw text documents [3, 9, 11, 12, 13, 14, 15, 16]; they define heuristic rules on how to efficiently extract AE pairs by utilizing various schemes, for example, ranking models [11, 14], Hidden Markov Model (HMMs) [16], and Support Vector Machines (SVMs) [9, 14, 15], and so on. In particular, Hwang et al. attempted to build a state-of-the-art AcroDic (Acronym Dictionary) that consists of reliable AE pairs, their definitions (glossaries), semantically related terms, and Wikipedia links [3]. The AcroDic contains 108 237 AE pairs (i.e., 108 237 meanings), and its data and search browser are provided online for public use.

Similarly, almost all previous work has concentrated on the extraction of AE pairs. Although it is important to construct a body of fundamental knowledge, what is essential to better NLP performance is how that knowledge is applied to AE disambiguation. The current study contains a partial AE pair-extraction method from raw text, but concentrates more heavily on a disambiguation method based on Naive Bayesian classifiers; ultimately, the objective is to determine the best expansions for acronyms found in documents.

3 ACRONYM-EXPANSION DISAMBIGUATION

This study proposes an AE disambiguation method that uses Naive Bayesian classifiers to determine the full expansions of certain acronyms that occur frequently in documents. Figure 2 shows the architecture of the proposed method. Our methods process consists of three main steps.


NDSL
NATIONAL DISCOVERY for SCIENCE LEADERS
Abstract
Text Processor
Sentence Extractor
Acronym-Expansion Extractor
Sentence
Noun Type Extractor
POS Tagger
Tagged Data
Rule-based Noun Type Extractor
AE Pair
NP
AE Pairs
XRD X ray diffraction
CRP C reactive protein
... ...
NP Set
transmission electron microscopy, intercalate day structure, ...
Natural Language Processing
Word Sense Disambiguation
AE Pair
NP Clues
Noun
Noun Set
Treatment, electron, ...
Serum, patient, interleukin, ...
Noun Clues
$p(C|F_1,\dots,F_n) = \frac{p(C)\ p(F_1,\dots,F_n|C)}{p(F_1,\dots,F_n)}.$
Naïve Bayesian Learner
Acronym Expansion Recognizer

Whether NLP is distinct from, or identical to, the field of computational linguistics is a matter of perspective. Research has progressed steadily to the point where WSD systems achieve sufficiently high levels of accuracy on a variety of word types and ambiguities.

Figure 2. Architecture for acronym-expansion disambiguation

Step 1. The Text Processor extracts sentences from raw text and checks whether each sentence contains acronyms. Sentences that contain acronyms are transferred to step 2, which involves the use of the Noun Type Extractor. If the acronyms are accompanied by expansions, the AE pairs are stored in the first database, AE Pairs.

Step 2. The Noun Type Extractor processes the sentences that are transferred from the Text Processor. The extractor uses a part of speech (POS) Tagger and a Rule-based Noun Type Extractor, step by step, to extract NNs and NPs separately. The extracted nouns are inserted into a NP Set or Noun Set database, according to noun type.

Step 3. The Naive Bayesian Learner measures the similarity between an AE pair and the clues (NNs, NPs, and NN + NPs). The similarity measured in step 3 is used to determine appropriate expansions for the given acronyms.

3.1 Text Processor

The Text Processor extracts sentences that contain acronyms or AE pairs and transfers the sentences to the Noun Type Extractor. If an AE pair exists, the processor saves the pair to the AE Pairs database. As this study deals only with the abstracts of research papers, we follow the rules below to extract AE pairs and acronyms.

Rule 1. The main acronym occurs at least once with its expansion, and the acronym or its expansion is contained within rounded brackets. To extract the pair, the following rule is used:

i. $ABC : A_{word}(p|c)\, B_{word}(p|c)\, C_{word}$, where A_{word} is a word and A is the initial letter; p is prepositions, such as of, for, in, at, under, into, with; and c is a coordination, such as and, or. Here, A, B and C are simply symbols used to express patterns; the preposition or coordination can be ignored. This pattern can be applied to the extraction of pairs like AA, average age; BEREC, Body of European Regulators of Electronic Communications; and BERTS, Bangkok Elevated Road and Train System.

ii. $ABC : {}_{word}A(p|c)\, {}_{word}B(p|c)\, {}_{word}C$, where ${}_{word}A$ is a word that includes A in capitalized form to make pairs like XML, eXtensible Markup Language; and XHTML, eXtensible HyperText Markup Language.

Rule 2. If an acronym is accompanied once by its expansion, we consider the same acronym that occurs in the same abstract section as having the same expansion.

Table 1 shows an example of the Text Processor.

3.2 Noun Type Extractor (Construction of Contextual Information)

The meaning of a word in some given text can often be determined by examining its context words within that piece of text; this is also true of acronyms. To pinpoint appropriate expansions, contextual words that co-occur with acronyms are collected. This study considers only noun types and extracts NNs and NPs individually for various analyses. It is easy to find NNs through the use of the Stanford POS Tagger,[4] which returns tagged results for each word. To construct an NP clue set, we employ a Rule-based Noun Type Extractor (RUN) [2] while following two additional rules:

[4] The Stanford Natural Language Processing Group: `http://nlp.stanford.edu/`

Abstract	. . . The **head-mounted display (HMD)** eliminates the negative effects of yaw, roll, and pitch - each of which is detrimental to the performance of complex operative procedures. There has been no visual strain or ocular fatigue observed. In contrast, the HMD allowed increased concentration without subjective muscle strain for as long as 640 mins. The authors conclude that the HMD improves efficiency in complex procedures, increases safety, diminishes cost, . . .
AE pair	HMD – head mounted display
Sentence(s)	1. The head-mounted display (HMD) eliminates the negative effects . . . procedures. 2. In contrast, the HMD allowed increased concentration . . . 640 mins. 3. The authors conclude that the HMD improves efficiency . . . diminishes cost, . . .

Table 1. An example of the text processor

Tagged nouns	1. effects, yaw, roll, pitch, performance, operative, procedures 2. concentration, muscle, strain 3. authors, efficiency, procedures, increases, cost, visualization, field, surgeon, assistants, operating-room, environments
Stemmed nouns	1. effect, yaw, roll, pitch, perform, oper, procedur 2. concentr, muscl, strain 3. author, effici, procedur, increas, cost, visual, field, surgeon, assist, operating-room, environ
Stemmed noun phrases	1. neg effect, complex oper, complex procedur, complex oper procedur, oper procedur 2. subject muscl, subject strain, subject muscl strain, muscl strain 3. complex procedur, optimum visual, oper field, operating-room environ

Table 2. Examples of POS tagged nouns, stemmed nouns, and stemmed noun phrases; number indicates sentence identifier for an AE pair

RUN 1. NPs consist of multiple words that contain two or more nouns (e.g., “health sciences” and “health sciences librarianship”) or contain an adjective(s) and a noun(s) (e.g., “electronic structure” and “apparent catalytic rate”). To compare influences, the NN is not involved in this NP extraction.

RUN 2. If the NP contains more than two words, this step constructs the possible subsets from the NP. In this step, all subsets should follow the NP conditions described in RUN 1. For example, from “health sciences librarianship”, which consists only of nouns, four subsets can be derived: “health sciences”, “sciences librarianship”, “health librarianship”, and “health sciences librarianship”. This is not the case with NPs that contain adjectives; from the NP containing two adjectives and one noun, “apparent catalytic rate”, three NP subsets are created: “apparent rate”, “catalytic rate”, and “apparent catalytic rate”.

Acr.	Acr. Freq. *freq(acr)*	Exp.	Exp. Freq. *freq(exp)*	Clue (frequency)
PM	88	Post Mortem	56	mrna (3), sampl (5), surgeri (1), hippocamp (1), sclerosi (1), patient (2), subiculum (1), neuron (1), gyru (1), cell (6), pancreas (2), interv (1), sodium (2), potassium (2), cyanid (2), treatment (1), cistern (1), ...
PM	88	Penman Monteith	32	priestleytaylor (3), pt (6), evapotranspir (5), crop (7), growth (1), water (9), penman (5), allen (2), surface (5), resist (3), summertim (2), evapor (3), dam (2), et0 (2), fao (7), nsc (2), ...

Table 3. Learning examples for Naive Bayesian classifier

Based on the tagger results and the two aforementioned rules, three kinds of contextual information containing NNs, NPs, or NN + NPs are prepared. In calculating relatedness, one should note that term variations can affect relatedness in unexpected ways; this means that determining the original forms of words is essential. Thus, we use a Porter stemmer[5] [8]. Table 2 shows examples of the Noun Type Extractor (Construction of Contextual Information) section, including POS tagged nouns, stemmed nouns, and stemmed NPs. The example is the AE pair (HMD, head mounted display) taken from Section 3.1.

Through these steps, we construct contextual information for each AE pair; this information is inserted into the Noun Set or NP Set database, depending on the noun type. The sentences shown in Table 2 are used to prepare a learning set and test set (explained in Section 3.3).

3.3 Learning Based on Naive Bayesian Classifiers

We prepared AE pairs and their contextual information, using the processes outlined above. Equations (1) and (2) express the expansions (*Exp_Set*) of an acronym and the contextual information (*Context*) of an expansion, respectively.

$$\begin{aligned} \mathit{Exp\_Set}(acr_i) &= \{exp_j, 0 \le j \le k\} && (1) \\ \mathit{Context}(exp_j) &= \{clue_m, 0 \le m \le n\} && (2) \end{aligned}$$

where exp_j is an element of *Exp_Set*, $clue_m$ is a word, and k and n are the length of each set. The contextual information for an expansion includes three kinds of set: NNs, NPs, and NN+NPs. We describe here the method by which the Naive Bayesian classifier determines an appropriate expansion from a given context. The classifier, which is based on neighboring words, is used widely to disambiguate meaning [7].

[5] The Porter Stemming Algorithm: `http://tartarus.org/~martin/PorterStemmer/`

The current study applies this method to AE disambiguation, and Equation (3) determines an expansion.

$$Decide(exp) = arg_{exp_j \in Exp\_Set(acr_i)} \max \left[\log(1 + P(exp_j)) + \sum_{m=1}^{n} (1 + P(clue_m)) \right] \quad (3)$$

where *exp* is the final expansion, which has the maximum value as measured by the method. Probability P is calculated by Equations (4) and (5).

$$P(exp_j) = \frac{freq(exp_j)}{freq(acr_i)} \quad (4)$$

$$P(clue_m) = \frac{freq(clue_m)}{freq(exp_j)} \quad (5)$$

Table 3 shows learning examples of the Naive Bayesian classifier for AE disambiguation. The AE pairs used in the table are part of our real experimental data.

We assume that the acronym "PM" has two types of expansion (i.e., "post mortem" and "penman monteith") and that the expansions accompany their contexts individually, as shown in Table 3. If another "PM", accompanied by clues, is encountered in a new document, the classifier compares relatedness and chooses one expansion from among them. Table 4 shows the calculation processes.

The classifier calculates the weights of possible expansions based on Equation (3) and chooses that which has the greatest weight. In the examples shown in Table 4, the classifier selects "post mortem" as an appropriate expansion, given the clues at hand.

			Post Mortem		Penman Monteith	
Index	Acr.	Clues	Matching	Clue weight	Matching	Clue weight
93261	PM	analys	X	0	X	0
		increas	X	0	X	0
		total	X	0	X	0
		volum	O	0.0077	X	0
		increas	X	0	X	0
		total	X	0	X	0
		membran	X	0	X	0
		surfac	X	0	O	0.0631
		golgi	X	0	X	0
		cisterna	O	0.0077	X	0
		period	O	0.0077	O	0.0134
$Sigma^{n}_{m=1}(1 + P(clue_m))$				0.0231		0.0765
$\log(1 + P(exp_j))$				0.2139		0.1347
Summarization				**0.2370**		0.2112

Table 4. Test examples for Naive Bayesian classifier

4 EXPERIMENTAL EVALUATION

In this section, we evaluate the efficiency of the proposed method used in AE disambiguation; we also look at which contexts in which method performs especially well. To provide an objective evaluation, the results are compared to a baseline. In this sense, we have constructed a "gold standard" for our evaluation and we provide comparative results.

4.1 Gold Standard

Constructing a gold standard for AE disambiguation is easier than in other NLP fields, because documents contain an AE pair at least once, if the acronym is important. Thus, questions and their correct answers with respect to the gold standard on AE pairs can be collected simultaneously. The process is described in detail in Section 3.1. Throughout that process, we prepared a total of 1 253 335 AE pairs as the gold standard, as sourced from NDSL[6]. The gold standard is divided into two sets: a learning set and a test set. We use n-fold validation (cross-validation) for the evaluation and set n to 5. This means that 80 percent of the data is used in learning, and 20 percent in testing. In addition, we need to know the influence of the frequency of AE pairs; thus, we assign a threshold value (tv) to the frequencies. Table 5 summarizes each statistic, by tv.

tv	cnt_{AE}	cnt_{occ}	Avg
≥ 1	90 687	1 253 335	13.820
≥ 5	28 280	1 121 938	39.672
≥ 10	12 501	1 021 280	81.696
≥ 15	8 273	972 034	117.495
≥ 20	6 218	937 602	150.788
≥ 25	5 004	911 130	182.080
≥ 30	4 198	889 492	211.885

cnt_{AE} = total count of AE pairs; cnt_{occ} = total count of frequencies of AE pairs. Avg refers to average frequency of AE pairs (high frequency indicates numerous and wide-ranging clues).

Table 5. Statistics according to threshold values

For this evaluation, we used AE pairs that occur five or more times, on account of our use of five-fold validation. The evaluation results are presented in the following section.

[6] NDSL (National Discovery for Science Leaders): `http://scholar.ndsl.kr/index.do`

4.2 Evaluation

We have prepared learning and testing sets according to tv (i.e., 5, 10, 15, 20, 25, and 30). To provide an objective evaluation, the results are compared to a baseline that selects only expansions that bear the maximum frequency. For example, if an acronym "PM" is given and the acronym has two expansion candidates – like "post mortem" and "penman monteith", with frequencies of 56 and 32, respectively, the baseline method chooses the first candidate based on the higher frequency. Table 6 summarizes the evaluation results. Here, the improvement rate (IR) measures performance improvement in terms of tv. The IR is calculated in Equation (6):

$$improvement\_rate = \frac{\sum_{i=2}^{n} precision_i - precision_{i-1}}{n-1} \tag{6}$$

where i is the index of tv and n is the count of precision results. For the evaluation, n is fixed to 6. A high IR means that the clue type affects the disambiguation performance positively.

Methods		Precisions on threshold values						
CT	ET	5	10	15	20	25	30	*IR*
NN	Max	91.44	93.25	93.97	94.46	94.82	95.09	0.73
	Min	90.88	92.92	93.74	94.25	94.74	95.03	0.83
	Avg.	**91.26**	93.15	93.90	94.40	94.78	95.06	0.76
NP	Max	87.39	89.25	90.34	91.13	91.73	92.26	0.97
	Min	87.14	89.07	90.21	90.97	91.59	92.12	1.00
	Avg.	87.25	89.13	90.28	91.06	91.67	92.19	0.99
combination (NN+NP)	Max	91.33	93.44	94.06	94.52	94.92	95.21	0.78
	Min	90.72	93.09	93.79	94.27	94.78	95.06	0.87
	Avg.	91.16	**93.32**	**93.98**	**94.45**	**94.87**	**95.14**	0.80
Baseline	Max	84.47	87.30	88.83	89.82	90.58	91.28	1.36
	Min	84.41	87.26	88.79	89.79	90.55	91.25	1.37
	Avg.	84.45	87.29	88.82	89.81	90.57	91.27	1.36

CT: clue types; ET: evaluation types; NN = single noun; NP = noun phrase. Underline and bold-face text indicate the worst and best performance, respectively. *Max*/*Min*/*Avg* are computed by maximum/minimum/average performances on five-fold evaluations.

Table 6. Evaluation results by clue types and threshold values

As shown in Table 6, we were able to confirm that the baseline method generated a performance range of about 84-91 percent. The baseline contains sound and reasonable results, indicating that documents mainly use acronyms that are already generalized. Moreover, the results indicate that the higher the tv is, the better the performance will be. From these results, we can see that low-frequency AE pairs occur widely and do not have sufficient contextual information. Actually, there were many cases in which low-frequency AE pairs were not matched between the learning and test sets.

In all evaluations, the results with NN showed better performance than those with NP. The performance with combination (NN + NP) clues, on the other hand, generally returned the best results; this no doubt owes to there being sufficient clues to ensure the pinpointing of the correct meaning. Using only NPs leads to the worst performance, even though it performs better than the baseline; this is because the contextual information, which consists of NPs, does not contain NNs, and this influences clue matching adversely. However, the IR with NP is higher than that with the other clue types; this implies that NPs contain strong disambiguation clues. When we consider the amount of contextual information involved, the real-world use of the combination-based method becomes an issue, because performance improves only a *little bit* more compared to that based on an NN, while incurring a higher time cost. Figure 3 shows the average performance of the various methods.

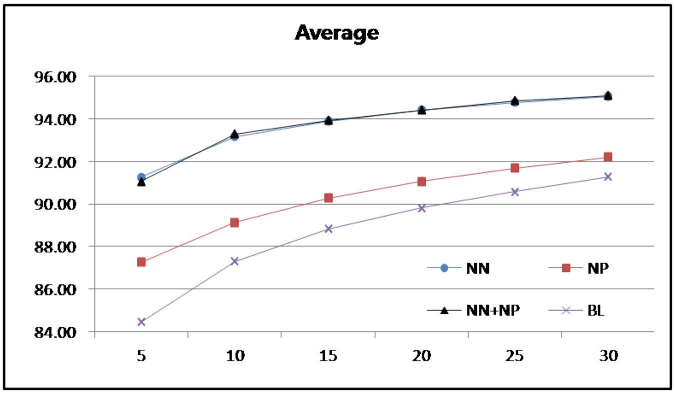


Figure 3. Average performance comparison

5 CONCLUSION

This study addresses the pinpointing of an appropriate expansion for an acronym in a given document; it also proposes a learning method that uses as clues co-occurrence words (i.e., contextual information). To refine and test the method, we prepared three kinds of contextual information that individually contained either single nouns (NN), noun phrases (NP), or a combination (NN + NP) and evaluated each through five-fold validation. Moreover, a variety of *tv*s (i.e., 5, 10, 15, 20, 25, and 30) were assigned to confirm the performance of the various methods according to AE pair frequency. The performance of each method was compared to each other and to a baseline. From these results, we could confirm that all methods performed better than the baseline, but the method involving combination clues as contextual information showed the best performance. Based on our performance evaluation, we expect that the use of our method will assist in creating high-quality, reliable term collections for use in business intelligence.

AE pairs that occur five or fewer times are largely domain-specific AE pairs and semantic term variations of high-frequency AE pairs. To resolve these issues in future research, we need to examine carefully the semantics of contextual information and develop application methods for use with lexical dictionaries such as Wikipedia.

Acknowledgement

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] Jung, H.—Lee, M.—Kim, P.—Sung, W. K.: Technology-Based Decision-Making Support System. In Proceedings of Human Interface and the Management of Information, Interacting with Information, Lecture Notes in Computer Science 2011, Vol. 6772, pp. 262–267.

[2] Hwang, M.—Choi, C.—Kim, P.: Automatic Enrichment of Semantic Relation Network and Its Application to Word Sense Disambiguation. IEEE Transactions on Knowledge and Data Engineering, Vol. 23, 2011, No. 6, pp. 845–858.

[3] Hwang, M.—Jeong, D. H.—Sung, W. K.: AcroDic 1.0: Acronym Dictionary. In Proceedings of The First Conference on Terminology, Languages, and Content Resources (LaRC), Korea 2011, pp. 38–49.

[4] Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press.

[5] Jung, J. J.: Discovering Community of Lingual Practice for Matching Multilingual Tags from Folksonomies. The Computer Journal, Vol. 55, 2012, No. 3, pp. 337–346.

[6] Jung, J. J.: Semantic Annotation of Cognitive Map for Knowledge Sharing Between Heterogeneous Businesses. Expert Systems with Applications, Vol. 39, 2012, No. 5, pp. 5857–5860.

[7] Lee, Y. G.—Chung, Y. M.: An Experimental Study on an Effective Word Sense Disambiguation Model Based on Automatic Sense Tagging Using Dictionary Information. Journal of the Korean Society for Information Management, Vol. 24, 2007, No. 1, pp. 321–342.

[8] Porter, M. F.: An Algorithm for Suffix Stripping. Program, Vol. 14, 1980, No. 3, pp. 130–137.

[9] Xu, J.—Huang, Y. L.: A Machine Learning Approach to Recognizing Acronyms and their Expansion. In Proceedings of the Fourth International Conference on Machine Learning and Cybernetics 2005, pp. 18–21.

[10] Jung, J. J.: Evolutionary Approach for Semantic-Based Query Sampling in Large-Scale Information Sources. Information Sciences, Vol. 182, 2012, No. 1, pp. 30–39.

[11] Jain, A.—Cucerzan, S.—Azzam, S.: Acronym-Expansion Recognition and Ranking on the Web. In Proceedings of IEEE International Conference on Information Reuse and Integration 2007, pp. 209–214.

[12] Fox, J.—Brown, N.: Automatically Extracting Acronyms from Biomedical Text. In Proceedings of IEEE International Conference on Bioinformatics and Bioengineering 2007, pp. 1245–1248.

[13] Rafeeque, P. C.—Abdul Nazeer, K. A.: Text Mining for Finding Acronym-Definition Pairs from Biomedical Text Using Pattern Matching Method with Space Reduction Heuristics. In Proceedings of International Conference on Advanced Computing and Communications 2007, pp. 295–300.

[14] Ni, W.—Huang, Y.: Extracting and Organizing Acronyms Based on Ranking. In Proceedings of World Congress on Intelligent Control and Automation 2008, pp. 4542–4547.

[15] Gao, Y. M.—Huang, Y. L.: Using SVM with Uneven Margins to Extract Acronym-Expansions. In Proceedings of International Conference on Machine Learning and Cybernetics 2009, pp. 1286–1292.

[16] Osiek, B. A.—Xexeo, G.—Carvalho, L. A. V.: A Language-Independent Acronym Expansion from Biomedical Texts with Hidden Markov Models. IEEE Transactions on Biomedical Engineering, Vol. 57, 2010, No. 11, pp. 2677–2688.

[17] Hwang, M.—Kim, P.: A New Similarity Measure for Automatic Construction of the Unknown Word Lexical Dictionary. International Journal on Semantic Web and Information Systems, Vol. 5, 2009, No. 1, pp. 48–64.

[18] Jung, J. J.: Attribute Selection-Based Recommendation Framework for Short-Head User Group: An Empirical Study by MovieLens and IMDB, Expert Systems with Applications, Vol. 39, 2012, No. 4, pp. 4049–4054.

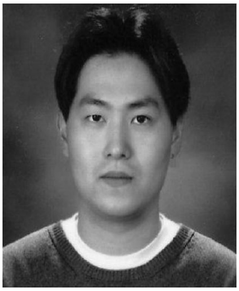
[19] Velardi, P.—Cucchiarelli, A.—Petit, M.: A Taxonomy Learning Method and Its Application to Characterize a Scientific Web Community. IEEE Transactions on Knowledge and Data Engineering. Vol. 19, 2007, No. 2, pp. 180–191.

[20] Kim, J.—Lee, M.—Song, S. K.—Sung, W. K.—Jung, H.: Toward Discovering and Predicting Technical Opportunities and Technology Trends. Advances in Information Sciences and Service Sciences (AISS), Vol. 4, 2011, No. 11, pp. 161–167.

[21] Jung, J. J.: Service Chain-Based Business Alliance Formation in Service-Oriented Architecture. Expert Systems with Applications, Vol. 38, 2011, No. 3, pp. 2206–2211.

Myunggwon Hwang received the B. Sc. Degree in Computer Engineering, the M. Sc. Degree in Computer Science, and the Ph. D. Degree in Computer Engineering from Chosun University, Korea. He is a senior researcher in Department of Software Research at Korea Institute of Science and Technology Information (KISTI). His research focuses on semantic information processing and retrieval, word sense disambiguation and business intelligence.

Do-Heon Jeong received the M. Sc. in Information Science from Yonsei University in 2003. He is a leader of analytics service development team in Software Research Center, Korea Institute of Science and Technology Information (KISTI). His research focuses on informetrics, semantic web and text mining.



Jinhyung Kim received the M. Sc. Degree and the Ph. D. Degree in Computer Science from Korea University. He is a postdoctoral researcher in the Department of Software Research at the Korea Institute of Science and Technology Information (KISTI). His research focuses on semantic web technologies, semantic information processing and retrieval, word sense disambiguation and knowledge acquisition.

Sajjad Mazhar works as student researcher in the Department of Software Research and student of UST (University of Science and Technology), Korea since September 2012. He received his M. Sc. from Dongguk University, Seoul in 2012. Previously, he was junior scientist at Information Control and Computer Complex (ICCC), Pakistan, and worked as JS and SS in ICCC, Pakistan. His current research interests include competitive technology intelligence based on the semantic web and text mining technologies, human–computer interaction (HCI).

Sa-Kwang Song received his B. Sc. Degree in Statistics in 1997 and his M. Sc. Degree in Computer Science in 1999 from Chungname National University, Korea. He received his Ph. D. Degree in Computer Science at Korea Advanced Institute of Science and Technology (KAIST), Korea. He had worked for Electronics and Telecommunications Research Institute (ETRI), Daejeon, Korea in 2005–2010. He joined Korea Institute of Science and Technology Information (KISTI), Daejeon, Korea in 2010. He is currently a senior researcher at Department of SW Research. His current research interest includes text mining, natural language processing (NLP), information retrieval, semantic web, and big data.

Hanmin Jung works as the head of the Department of Software research and chief researcher at Korea Institute of Science and Technology Information (KISTI), Korea since 2004. He received his M. Sc. and Ph. D. Degrees in Computer Science and Engineering from POSTECH, Korea in 1994 and 2003. Previously, he was senior researcher at Electronics and Telecommunications Research Institute (ETRI), Korea, and worked as CTO at DiQuest Inc, Korea, Now, he is also Adjunct Professor at University of Science & Technology (UST), Korea, Executive Director at Korea Contents Association, and committee member of ISO/IEC JTC1/SC32 and ISO/IEC JTC1/SC34. His current research interests include technology intelligence based on the semantic web and text mining technologies, human–computer interaction (HCI), and natural language processing (NLP). For these research areas, over 230 papers and 150 patents have been published and created by him.