IEICE TRANS. INF. & SYST., VOL.E91–D, NO.5 MAY 2008

LETTER

Automatic Acronym Dictionary Construction Based on Acronym Generation Types

Yeo-Chan YOON[†], *Nonmember*, **So-Young PARK**[††], **Young-In SONG**[†††], **Hae-Chang RIM**[†††a)], *Members*, *and* **Dae-Woong RHEE**[††], *Nonmember*

SUMMARY In this paper, we propose a new model of automatically constructing an acronym dictionary. The proposed model generates possible acronym candidates from a definition, and then verifies each acronym-definition pair with a Naive Bayes classifier based on web documents. In order to achieve high dictionary quality, the proposed model utilizes the characteristics of acronym generation types: a syllable-based generation type, a word-based generation type, and a mixed generation type. Compared with a previous model recognizing an acronym-definition pair in a document, the proposed model verifying a pair in web documents improves approximately 50% recall on obtaining acronym-definition pairs from 314 Korean definitions. Also, the proposed model improves 7.25% F-measure on verifying acronym-definition candidate pairs by utilizing specialized classifiers with the characteristics of acronym generation types.

key words: *acronym, automatic dictionary construction*

1. Introduction

Acronym identification is regarded as the problem of finding pairs of a long form(a definition such as a named entity) and a short form(an acronym). These pairs are known to be useful for an information retrieval system or a question answering system. For example, given a query with a definition *Carnegie Mellon University*, an information retrieval system can retrieve some documents including its acronym *CMU* by using acronym-definition pairs. However, it's hard to construct an acronym dictionary manually because too many acronyms exist and also new acronyms are continuously generated. Consequently, it is required to develop a method of automatically constructing the dictionary.

Recently, some approaches have been proposed to automatically find an acronym-definition pair in an English document [1]–[5]. These approaches recognize an acronym consisting of capital letters, and then search its definition in a document by using some heuristics or statistical classifiers. These approaches show a good performance for finding co-reference relation between an acronym and a definition in the document. Nevertheless, it is impossible to recognize an acronym such as *radar*, which has no explicit evidence to distinguish an acronym from a non-acronym. Moreover, these approaches are not guaranteed to construct a large-scale dictionary because the approach, which can find a single acronym-definition pair in a given document, cannot obtain other correct acronym-definition pairs which do not appear in the document.

On the other hand, other approaches have been proposed to verify a pair of two representations in web documents [6], [7]. These approaches are based on the assumption that a representation much more frequently occurs with its correct correspondent than its incorrect correspondent in a very large corpus. However, these approaches show limitation on verification performance because these approaches depend on only co-occurrence information without considering relative characteristics between a representation and its correspondent.

In this paper, we propose a new model of automatically constructing an acronym dictionary. In order to reduce the difficulty of recognizing a Korean acronym, which has no explicit evidence to distinguish an acronym from a non-acronym, the proposed model generates acronym candidates from a given definition. For the purpose of achieving high dictionary quality, the proposed model utilizes the characteristics of acronym generation types as well as the co-occurrence information in web documents

2. Acronym Generation Types

An acronym is generated by aligning and extracting some pieces from a definition consisting of a few words. As described in Fig. 1, acronym generation types can be classified into four types: a character-based generation type, a syllable-based generation type, a word-based generation type, and a mixed generation type.

First, the character-based generation type(CGT) de-

		Character-based Generation Type(CGT)	Syllable-based Generation Type(SGT)	Word-based Generation Type(WGT)	Mixed Generation Type(MGT)
English	Acronym	MOST	UPenn	Oxford Alumni Association	Radar
	Definition	**M**useum **o**f **S**cience and **T**echnology	**U**niversity of **Penn**sylvania	**Oxford** University **Alumni Association**	**R**adio **D**etecting **A**nd **R**anging
Korean	Acronym	애 (a child)	고대 (KU)	교육부 (the Ministry of Education)	대우자판 (DWMS)
	Definition	**아이** (a child)	고려대학교 (Korea University)	**교육**인적자원**부** (the Ministry of Education and Human Resources Development)	**대우자**동차판**매** (Daewoo Motors Sales Corporation)

Fig. 1 Examples according to acronym generation types.

Manuscript received July 4, 2007.
Manuscript revised October 26, 2007.

[†]The author is with Speech/Language Information Research Center, ETRI, Gajeong-dong, Yuseong-gu, Daejon, Korea.
[††]The authors are with Division of Digital Media Technology, SangMyung University, 7 Hongji-dong, Jongro-gu, Seoul, Korea.
[†††]The authors are with the Department of Computer Science Engineering, Korea University, Anam-dong, Seongbuk-gu, Seoul, Korea.
a) E-mail: rim@nlp.korea.ac.kr

DOI: 10.1093/ietisy/e91–d.5.1584

scribes that an acronym is generated by aligning and extracting some characters from a definition. Most English acronyms are applicable to this type. For example, a definition *museum of science and technology* can be represented as an acronym *MOST*.

Second, the syllable-based generation type(SGT) represents that an acronym is generated by aligning and extracting some syllables from a definition. For example, a definition *University of Pennsylvania* is denoted as an acronym *UPenn* based on syllable units.

Third, the word-based generation type(WGT) describes that an acronym is generated by aligning and extracting some key words from a definition. As shown in Fig. 1, an acronym *the Ministry of Education* represents its definition without some non-key words.

Fourth, the mixed generation type(MGT) describes that an acronym is generated based on more than two acronym generation types. As shown in Fig. 1, an acronym *radar* is generated by aligning and extracting the first syllable *ra* from *radio* and the first characters *d*, *a*, and *r* from *detecting and ranging* respectively. Next section will describe a model of constructing a acronym dictionary based on these acronym generation types.

3. Korean Acronym Dictionary Construction

In order to construct a Korean acronym dictionary for a given definition list, the proposed model generates possible acronym candidates and verifies each acronym-definition candidate pair as shown in Fig. 2. For dictionary quality, the proposed model considers three acronym generation types: the syllable-based generation type, the word-based generation type, and the mixed generation type. For the purpose of reducing the model complexity, the proposed model excludes the character-based generation type because this type is applicable to an extremely low ratio in the whole Korean acronyms while this type generates too many spurious acronym candidates.

Section 3.1 describes how to generate each candidate pair of a definition and an acronym according to alignment rules. And then, Sect. 3.2 represents how to yield final dictionary entries by using each type's own verifier as shown in Fig. 2. In this paper, a final dictionary entry is represented as *(acr,def)*, a pair of an acronym and a definition, while a candidate pair is represented as *type(acr,def)* with an acronym generation type in order to utilize the characteristics of each acronym generation type during verification.

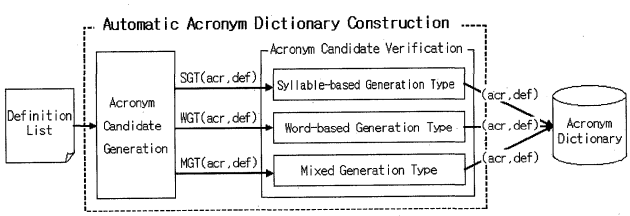

Fig. 2 Acronym dictionary construction model.

3.1 Acronym Candidate Generation

Given a definition represented as a sequence of words, the proposed model aligns some nodes from each word according to the following three alignment rules. And then, the proposed model connects two immediate nodes with transition arcs as shown in Fig. 3. In order to reduce the number of candidate acronyms, the syllable alignment rule does not utilize the middle syllables in a word. Compared with a simple rule generating $2^s - 2$ acronym candidates by combining some syllables, these three rules generates $4^w - 2$ acronym candidates where w(the number of words) is much smaller than s(the number of syllables), and -2 represents both a definition itself and a null candidate ϕ.

- **Non-Alignment Rule:** Align no character, no syllable, and no word from a given word.
- **Syllable Alignment Rule:** Align either the first syllable or the last syllable from a given word.
- **Word Alignment Rule:** Align the whole word from a given word.

In order to assign each candidate pair to the generation type, the propose model analyzes the rules applied for the acronym candidate. Specifically, the generation type SGT is assigned when an acronym candidate is generated by the syllable alignment rule and the non-alignment rule. Also, the generation type WGT is assigned when an acronym candidate is generated by the word alignment rule and the non-alignment rule. Especially, the generation type MGT is assigned when both the syllable alignment rule and the word alignment rule must be used as shown in an acronym candidate with bold lines of Fig. 3.

3.2 Acronym Candidate Verification

In order to select final dictionary entries from candidate pairs, the proposed model utilizes a Naive Bayes classifier as shown in the following equation. For the purpose of improving verification performance, a training set is divided into three training sets according to the acronym generation types, and then each classifier learns from its own training set with its own suitable features as shown in Fig. 2.

$$\underset{pair\in\{0,1\}}{argmax}\ P(\ pair \mid acr,\ def,\ web\ docs\ ) \tag{1}$$

$$\approx \underset{pair\in\{0,1\}}{argmax}\ P(\ pair \mid f_1, f_2, \ldots f_n\ ) \tag{2}$$

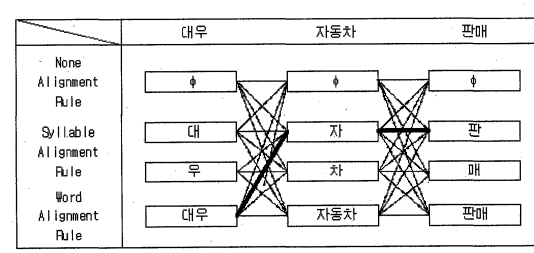

Fig. 3 Acronyms candidate generation example.

$$= \underset{pair\in\{0,1\}}{argmax} \frac{P(\ pair\ ) \times P(\ f_1,\ f_2,\ ...,\ f_n \mid pair\ )}{P(\ f_1,f_2,...f_n\ )} \quad (3)$$

$$= \underset{pair\in\{0,1\}}{argmax} P(\ pair\ ) \times P(\ f_1,\ f_2,\ \ldots,\ f_n \mid pair\ ) \quad (4)$$

$$\approx \underset{pair\in\{0,1\}}{argmax} P(\ pair\ ) \times \prod_{i=1}^{n} P(\ f_i \mid pair\ ) \quad (5)$$

Given an acronym candidate, a definition, and web documents, the proposed model estimates the probability of appearing the given acronym-definition pair, and selects whether the *pair* takes either 1 to accept the candidate pair as a final dictionary entry or 0 to reject it according to the highest probability as described in the equation (1). The equation (2) describes that the given events are replaced by some features, and is rewritten as the equation (3) based on Bayes' theorem. In the equation (4), the denominator is removed because the denominator is not related to the goal value *pair*. Finally, the equation (5) generalizes multiple events by the chain rule and the assumption that a feature is independent on the other features.

As shown in Table 1, the proposed model uses eight verification features to represent the characteristics between an acronym candidate and a definition. The generation features, f_{num_1}, f_{num_2}, and f_{num_3}, express the numeric characteristics related to generate an acronym from a definition. Also, the difference features, f_{diff_1}, f_{diff_2}, and f_{diff_3}, describe the difference between a definition and an acronym. Besides, the frequency features, f_{freq_1} and f_{freq_2}, describe the tendency that a definition more often occurs with its correct acronym than its incorrect acronym in web documents.

4. Experimentation

For the purpose of examining the coverage of the proposed model, we have applied three alignment rules to a set of 815 correct acronym-definition pairs based on 314 definitions. The definitions written by Korean consist of 161 company names, 85 government agency names, and 68 university names. As a result, the model covers 99.51% correct pairs while generating 158.68 candidate pairs per definition on average. Finally, the model yields 2.62 dictionary entries per definition on average after verification.

Table 1 Verification features.

Feature	Description
f_{num_1}	the number of syllables in the acronym candidate.
f_{num_2}	the number of the last syllables of words used in the acronym candidate.
f_{num_3}	the number of using *Syllable Alignment Rule* to generate the acronym candidate.
f_{diff_1}	the difference between the number of total syllables in the definition and the number of syllables in the acronym candidate.
f_{diff_2}	the difference between the number of total words in the definition and the number of words used in the acronym candidate.
f_{diff_3}	the difference between the number of words used in the acronym candidate and the number of words not used in the acronym candidate.
f_{freq_1}	the co-occurrence frequency in the Google's top 100 snippets retrieved by using both the definition and the acronym candidate as a query.
f_{freq_2}	the frequency of the definition in the Google's top 100 snippets retrieved by using the acronym candidate as a query.

In order to prove the verification validity of the proposed model, we have tested the model with 5-fold cross validation on a set which is divided into 80% for the training set and 20% for the test set. Clearly, the set consists of 50,139 pairs of an acronym candidate and a definition, which are manually classified into 815 correct pairs and 49,324 incorrect pairs based on three alignment rules. Besides, we utilize the following performance measures. *Precision* indicates the average ratio of correct dictionary entries from final dictionary entries generated and verified by the proposed model. *Recall* indicates the average ratio of correct final dictionary entries from 815 correct pairs in the set. *F-measure* indicates their harmonic mean.

4.1 Performance on Feature Combination

For the purpose of evaluating the verification performance according to feature combinations, we select some useful features, and evaluate these features on *F-measure* as shown in each row of Table 2. In order to clearly show the characteristics of each acronym generation type, we also divide both the training data set and the test data set by acronym generation types.

The experimental results show that the best single feature indicates the feature f_{freq1} because the co-occurrence in web documents provides decisive information to classify plausible acronyms and implausible acronyms. Also, the feature f_{num2} is so useful since the last syllables tend to be strongly disallowed in most acronyms. Besides, the length difference feature f_{diff1} is more effective than the acronym length feature f_{num1}. As a result, the synergy effect of these features shows the best performance at 68.17% F-measure of ALL as shown in Table 2.

However, the useful features are different according to each acronym generation type. Particularly, SGT obviously depends on the frequency features, f_{freq1} and f_{freq2}, because *Syllable Alignment Rule* allows to generate implausible acronyms. Unlike other two types, WGT prefers the word feature f_{diff3} because WGT is closely related to words rather than syllables. On the other hand, MGT prefers the feature f_{num3} indicating the number of using *Syllable Align-*

Table 2 Performance on feature combination.

Features	ALL	SGT	WGT	MGT
f_{freq1}	62.87	62.15	66.65	59.22
f_{freq1}, f_{num1}	63.67	62.15	68.88	60.69
f_{freq1}, f_{num2}	**66.25**	65.18	66.65	**68.99**
f_{freq1}, f_{num3}	63.05	62.91	66.39	67.92
f_{freq1}, f_{diff1}	64.18	62.91	69.43	63.44
f_{freq1}, f_{diff2}	63.98	61.54	69.72	59.88
f_{freq1}, f_{diff3}	63.98	61.41	**70.67**	58.63
f_{freq1}, f_{freq2}	62.34	**67.66**	66.60	58.28
$f_{freq1}, f_{num2}, f_{diff1}$	**68.17**	66.39	69.43	69.63
$f_{freq1}, f_{num2}, f_{diff1}, f_{freq2}$	66.63	**69.74**	68.14	64.17
$f_{freq1}, f_{num2}, f_{diff1}, f_{num3}$	66.86	66.96	64.17	**70.69**

Table 3 Comparison with previous model.

Features	Precision	Recall	F–Measure
(Yeates99)	67.04	21.47	32.52
(Qu04)	58.72	67.65	62.87
Single Classifier	66.67	69.84	68.17
Specialized Classifiers	69.36	70.90	70.12

ment Rule since the mixed rate between SGT and WGT is important in MGT. These results show that each acronym generation type has its own characteristics.

4.2 Comparison with Previous Model

For comparison with previous models on the same test environment, we have applied (*Yeates99*) [3] and (*Qu04*) [6] to a Korean acronym generation topic. Furthermore, we have implemented both the proposed model using a *Single Classifier* and the proposed model using *Specialized Classifiers* indicating that each acronym generation type has its own classifier.

Table 3 shows that (*Qu04*) improves 46.18% recall by verifying each acronym-definition pair in web documents compared with (*Yeates99*). It describes that (*Yeates99*), which focuses on finding an acronym-definition pair appeared in the same document, cannot obtain other correct acronyms which do not appear in the document while (*Qu04*) can verify these acronyms in the web documents.

(*Qu04*) using only the co-occurrence feature in the web documents prefers the more frequent words to the less frequent words although a definition is unrelated to the more frequent words in the web documents. On the other hand, a *Single Classifier* can reduce the preference by using the acronym generation features to represent the generation characteristics between an acronym candidate and a definition. Therefore, it improves 5.30% F-measure. Furthermore, *Specialized Classifiers* based on each type's own characteristics show 1.95% better F-measure than a *Single Classifier*. As a result, *Specialized Classifiers* lead to 7.25% F-measure improvement compared to (*Qu04*), and 49.43% recall improvement compared to (*Yeates99*).

5. Conclusion

In this paper, we propose a new model of automatically constructing an acronym dictionary based on three acronym generation types: the syllable-based generation type, the word-based generation type, and the mixed generation type. The proposed model has achieved following performance improvements.

First, the proposed model verifies each acronym-definition candidate pair in web documents to obtain a large-scale acronym dictionary. The experimental results show that the proposed model improves 49.43% recall compared with a previous model recognizing each acronym-definition pair in the same document.

Second, the proposed model utilizes acronym generation characteristics for verifying acronym-definition candidate pairs. The experimental results show that the proposed model improves approximately 5.30% F-measure by using the generation characteristics.

Third, the proposed model uses a different classifier for each acronym generation type. The experimental results show that specialized classifiers yield 1.95% better F-measure compared to a single classifier.

For future works, we will apply the proposed model to other languages such as Chinese and Japanese, which do not have any explicit property to distinguish an acronym from a non-acronym. Also, we will design a method which can filter out the highly frequent words unrelated to a definition. Furthermore, we will consider a more robust method against some spelling or spacing errors in web documents.

References

[1] D. Nadeau and P.D. Turney, "A supervised learning approach to acronym identification," Canadian Conference on AI, pp.319–329, 2005.

[2] K. Taghva and J. Gilbreth, "Recognizing acronyms and their definitions," Technical Report 94-07, Information Science Research Institute, 1994.

[3] S. Yeates, "Automatic extraction of acronyms from text," New Zealand Computer Science Research Students' Conference, pp.117–124, 1999.

[4] L.S. Larkey, P. Ogilvie, M.A. Price, and B. Tamilio, "Acrophile: An automated acronym extractor and server," Proc. Fifth ACM Conference on Digital Libraries, pp.205–214, 2000.

[5] Y. Park and R.J. Byrd, "Hybrid text mining for finding abbreviations and their definitions," Genetic and Evolutionary Computation Conference Late Breaking Papers, pp.317–324, 2001.

[6] Y. Qu and G. Grefenstette, "Finding ideographic representations of Japanese names written in Latin script via language identification and corpus validation," ACL 2004, pp.183–190, 2004.

[7] G. Grefenstette, Y. Qu, and D.A. Evans, "Mining the Web to create a language model for mapping between English names and phrases and Japanese," Web Intelligence 2004, pp.110–116, 2004.