1998 International Computer Symposium
Workshop on Artificial Intelligence
December 17-19, 1998, N.C.K.U., Tainan, Taiwan, R.O.C.

An Incremental Acquisition of Terms Using Automatic Selected Seeds and Extended Variation Rules

Hanmin Jung, Young-Kil Kim, Sung-Kwon Choi, and Sangkyu Park

Knowledge Processing Research Team
Department of Natural Language Processing
Electronics and Telecommunications Research Institute
Gajeong-Dong 161, Youseong
Taejeon, Korea 305-350
Email: {jhm, kimyk, choisk}@etri.re.kr, skpark@computer.etri.re.kr

ABSTRACT

An automatic pattern extraction from very large corpora can be easily applied to corpus-based NLP (Natural Language Processing) research using the domain characteristics. This paper focuses on a fully automatic acquisition of terms and its incremental processing. For the increment, we repeat the process that applies variations to original terms and uses the acquired terms as the original terms of the next iteration. We extend the variations from one to three types: inter variation, intra variation, and combined variation that simultaneously uses both variations. This extension helps to produce more terms incrementally. Our experiment for Web documents shows that the number of newly discovered terms increases by 61.4%, when compared with original variation.

1. INTRODUCTION

One of recent major NLP approaches is corpus-based processing using information from very large corpora [Armstrong 1994]. Terms acquisition can be widely used as a basic tool for machine translation by lexical dictionary and pattern database, domain identification using domain-specific keywords, automatic indexing, and so on. However, the more the size of corpora, the more the need of automatic acquisition. A number of researches have focused on an automation with statistical approach like [Church & Hanks 1989] [Ikehara *et al.* 1996], symbolic approach like [Bourigault 1993], and bilingual pattern alignment like [Haruno *et al.* 1996]. However, all of them have same mechanism with once-and-all process without considering any priori knowledge, so they have many problems including the deficiency of the number of acquired terms and the blindness of the semantic relation between the terms.

Jacquemin introduced an incremental acquisition of terms to overcome these weak points [Jacquemin 1995]. He manually selected reference terms as the seeds and repeated the iteration of the acquisition with inter variation rules (coordination, insertion, and permutation) until no more new terms are discovered. Nevertheless, the approach has two problems: coping with very large corpora using cost-high manual seeds selection and the narrow scope of variation rules.

We firstly combine statistical approach and Jacquemin's symbolic approach to automatically select the first reference terms, and secondly extend the variation rules to three types. We modify Ikehara's statistical uninterrupted pattern extraction by using the reduced PT (Pattern Table) and the reduced SPT (Sorted Pattern Table) [Jung *et al.* 1998b]. For the automation of terms' selection, we filter the uninterrupted patterns by a stop-word dictionary and tens of pattern removal rules. Experimental results show this pattern filter can prune 98.14% of initial patterns extracted from Web documents.

The extension of variation categories is a good way to discover new terms from corpora. Our categories of variation rules are extended to intra- and combined variation as well as inter variation. When compared with the Jacquemin's single category, the experiment also shows the extension makes us acquire more terms by 61.4%. This paper focuses on the topics from the automatic seeds (reference terms) determination to the application of the extended variations with the point of hybrid view.

2. THE AUTOMATIC ACQUISITION OF THE FIRST REFERENCE TERMS

Reference terms are basic terms to acquire new terms with incremental approach. The proper selection of the terms and description of variation rules are crucial to increase the reliability of newly discovered terms. Although it is desirable that specialists carefully select the terms, this manual selection becomes practically impossible in the case of very large corpora. Jacquemin tried to verify his incremental approach with manually selected reference terms, but the size of the corpus is only 1.56 Mbytes. To cope with the largeness, automatic construction of reference terms needs to be introduced. Especially, our interest is how to apply the automation to incremental terms' acquisition and reduce meaningless and improper terms.

We first use modified Ikehara's statistical approach, which is typical once-and-all process, to extract uninterrupted patterns from corpora [Ikehara *et al.* 1996] [Jung *et al.*

1998b]. Then, we use a stop-word dictionary and pattern removal rules to filter unnecessary patterns and pick out the first reference terms. The stop-word dictionary was constructed, based on the lexical dictionary of an English-Korean machine translation system "FromTo/EK" [Sim *et al.* 1998] and words from 3,600 Web documents of 25 domains [Jung *et al.* 1998a]. The entries mainly consist of determinants, prepositions, auxiliaries, conjunctions, and pronouns. The dictionary has about 580 entries. The pattern removal rules consist of about 30 rules that filter patterns by English and special character/string check.

3. TERMS ACQUISITION BY INCREMENTAL PROCESS

For incremental acquisition, new candidate terms should be used as the reference terms of the next iteration. Automatically extracted and filtered patterns become the first reference terms. A metarule interpreter applies three kinds of variation rules (inter-, intra-, and combined) to the corpus that is used to extract the first reference terms. It produces the first candidate terms that would become the second reference terms after being filtered by a term filter. This process repeats until no more candidate terms are found. The repetition is finite because the corpus for terms acquisition is finite, the number of variation rules also is finite, and the rules do not create any new candidate term that is not included in the corpus. The following equation shows that the process reaches a fixed point after a finite number of iteration.

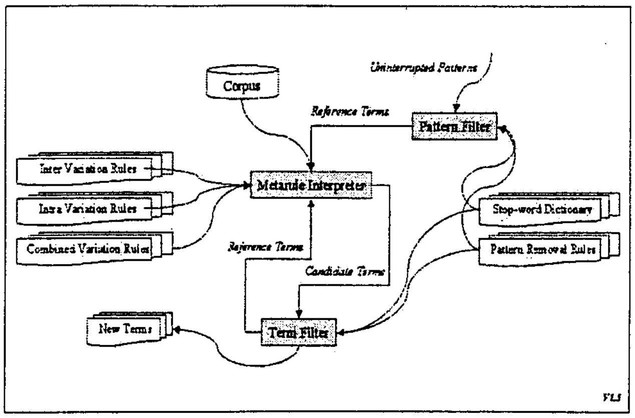

Figure 1. The process for terms acquisition with incremental mechanism.

$n(T_i) = n(T_{i-1}) + n(R_i)$, repeat until $n(R_i) = 0$, where $0 \leq n(R_j) < n(R_{j-1})$ and $1 \leq j \leq i$

$n(T) \geq n(T_0) + n(T_1) + \ldots + n(T_i)$ ($\because$ $S(R_j) \subset S(T)$)

($n(T)$: The number of all possible terms in a finite corpus, $n(T_i)$: The number of all accumulated terms acquired until the *i*th iteration, $n(R_i)$: the number of terms newly acquired by the *i*th iteration, $S(R_j)$: The set of terms newly acquired by the *j*th iteration, $S(T)$: The set of all possible terms in a finite corpus. $n(T)$ is equal to $n(T_0) + n(T_1) + \ldots + n(T_i)$ if recall is 100%. However, $n(T)$ is larger because the rules are not perfect to cover the corpus.)

4. EXTENDED VARIATION RULES

A variation means the deletion of one or more words in a term, the substitution with, and insertion of other word(s) to discover a new term. The original term is called a reference term, and a candidate term is a variant of it. Term variant has the semantic relation with its original term. For example, from the fact that term "A C" was obtained from "A B" using "A B and C" (head "B" coordinately conjuncts with head "C") in a corpus, we can know that the two heads "B" and "C" share the same semantic category (e.g. "A B and C" -> "surgical exploration and closure"). Jacquemin used this characteristic of term variants and defined an inter variation that consists of three sub-variations: coordination, insertion, and permutation. Metarules and rules describe variations for reference terms. The rules are converted into new rules for candidate terms by the metarules.

In this chapter, we introduce two additional variations: intra- and combined variation. Intra variation is for the variation of a word itself. We define combined variation as the combination of inter- and intra-variation.

4.1 Inter Variation

An inter variation consists of three kinds of sub-variations as follows.

> *Coordination*: The candidate term is the term coordinated with the original one.
>
> *Insertion*: The candidate term is the term that has replaced the head of the original term through substitution.
>
> *Permutation*: The argument of the original term is shifted from the left of the head to its right and is transformed into a prepositional phrase in the case of two-word term.

The metarules for the description of variations are distinguished each other by the types of variations and the number of words in a term. We do not consider the metarules for more than four-word terms because the frequency of them in corpora is too low (about 1% of whole inter variation) [Jacquemin 1995]. We define "X" as a word, "C" as a coordinate conjunction, and N as a word in a reference term. The following notation explains how a metarule applies to a rule and yields a new candidate term.

```
[Rule for reference term]
N1 -> N2 N3
[Metarule for the coordination]
Coor(X1 -> X2 X3)  ≡  X1 -> X2 C4 X5 X3
[Rule for candidate term]
N1 -> N2 C4 X5 N3 (C4 is a coordinate conjunction)
[Acquired term (candidate term)]
        X5 N3 (X5 is a word)
```

The following example shows how a reference term "business visas" is applied to the above metarule for a coordination and a new candidate term "student visas" is acquired.

[In corpora]
CHARGE: BUSINESS & STUDENT VISAS US$50
REQUIREMENTS: BUSINESS & STUDENT VISAS
[Rule for reference term]
N1 -> business visas
[Metarule for the coordination]
Coor(X1 -> X2 X3) ≡ X1 -> X2 C4 X5 X3
[Rule for candidate term]
N1 -> business [and or & ...] X5 visas (X5 is "student")
[Acquired term (candidate term)]
student visas

Table 1. Three inter sub-variations for two-word terms and their candidate terms.

Variation	Metarule	Candidate term
Coordination	Coor(X1 -> X2 X3) ≡ X2 X4 C5 X3	X2 X4
	Coor(X1 -> X2 X3) ≡ X2 C4 X5 X6 X3	X5 X6 X3
	Coor(X1 -> X2 X3) ≡ X2 X5 C4 X2 X3	X2 X5
Insertion	Ins(X1 -> X2 X3) ≡ X2 X4 X3	X4 X3
	Ins(X1 -> X2 X3) ≡ X2 X3 X4 X3	X4 X3
Permutation	Per(X1 -> X2 X3) ≡ X2 P4 X5 X3 (P4 = for)	X5 X3
	Per(X1 -> X2 X3) ≡ X2 X5 P4 X6 X3 (P4 = for, from)	X2 X5

4.2 Intra Variation

An intra variation is defined as the variation that occurs in a specific word in a term, and it is caused by the combination of the word with a suffix or with other word. We classify the variation into two sub-variations according to the combined position of them: pre-intra variation (e.g. "Affirmative Action" -> "Anti-Affirmative Action", "American women" -> "African-American women") and post-intra variation (e.g. "women business" -> "women-owned business", "Puerto Rican" -> "Puerto Ricans"). It is distinguished from inter variation in that a word in a reference term is replaced by a word without any removal of the other words in the term.

Pre-intra variation: The candidate term is the term that has prefix or a word in front of a word in the original one.

Post-intra variation: The candidate term is the term that has postfix or a word at the back of a word in the original one.

The description of the metarule for intra variation is the same as that of inter one. We define "R" as a prefix or a front word and "O" as a postfix or a backward word. "{ }" indicates a new word to be discovered by the variation.

[Rule for reference term]
N1 -> N2 N3
[Metarule for the pre-intra variation]
Pre(X1 -> X2 X3) ≡ X1 -> {R4X2} X3
[Rule for candidate term]
N1 -> {R4N2} N3 (R4 is a prefix or a front word)
[Acquired term (candidate term)]
{R4N2} N3

The following example shows a metarule for post-intra variation and a rule for candidate term.

[Rule for reference term]
N1 -> N2 N3
[Metarule for the post-intra variation]
Post(X1 -> X2 X3) ≡ X1 -> X2 {X3O4}
[Rule for candidate term]
N1 -> N2 {N3O4} (O4 is a postfix or a backward word)
[Acquired term (candidate term)]
N2 {N3O4}

Table 2. Two intra sub-variations for two- and three-word terms and their candidate terms.

Variation	Metarule	Candidate Term
Pre-intra variation	Pre(X1 -> X2 X3) ≡ X2 {R4X3}	X2 {R4X3}
	Pre(X1 -> X2 X3 X4) ≡ {R5X2} X3 X4	{R5X2} X3 X4
post-intra variation	Post(X1 -> X2 X3) ≡ {X2O4} X3	{X2O4} X3
	Post(X1 -> X2 X3 X4) ≡ X2 {X3O5} X4	X2 {X3O5}X3

4.3 Combined Variation

We define a combined variation as the variation that simultaneously occurs both inter- and intra variation in a reference term. It has six kinds of sub-variations: the multiplication of three inter variations (coordination, insertion, and permutation) and two intra variations (pre-intra and post-intra variation). In this variation, one or more words in a reference term are removed or substituted with other word(s) because it has the characteristics of both inter- and intra variation. It also helps to reduce the number of iterations of terms' acquisition. The following shows a metarule for the combined variation and its result.

[Rule for reference term]
N1 -> N2 N3
[Metarule for the combination of insertion and pre-intra variation]
InsPre(X1 -> X2 X3) ≡ X1 -> {R4X2} X5 X3
[Rule for candidate term]
N1 -> {R4N2} X5 N3 (R4 is a postfix or a word)
[Acquired term (candidate term)]
{R4N2} N3

Table 3. Combined variations for two-word terms and their candidate terms.

Variation	Metarule	Candidate rule
Coordination & Pre-intra variation	CoorPre(X1 -> X2 X3) ≡ {R6X2} X4 C5 X3	{R6X2} X4 {R6X2} X3
Coordination & Post-intra variation	CoorPost(X1 -> X2 X3) ≡ {X2O6} X4 C5 X3	{X2O6} X4 {X2O6} X3

Insertion & Pre-intra variation	InsPre(X1 -> X2 X3) ≡ X2 X4 {R5X3}	X4 {R5X3}
		X2 {R5X3}

Insertion & Post-intra variation	InsPost(X1 -> X2 X3) ≡ X2 X4 {X3O5}	X4 {X3O5}
		X2 {X3O5}

Permutation & Pre-intra variation	PerPre(X1 -> X2 X3) ≡ X2 P4 X5 {R6X3}	X5 {R6X3}
Permutation & Post-intra variation	PerPost(X1 -> X2 X3) ≡ X2 P4 X5 {X3O6}	X5 {X3O6}

We can discover more than one candidate term with a metarule as the first four rows in the above table (e.g. "X4 {R5X3}"and "X2 {R5X3}" from "InsPost(X1 -> X2 X3) ≡ X2 X4 {R5X3}"). This frequently occurs in the case of combined variations because several inter- and intra rules can be condensed into a metarule for the combined variation. It helps to increase the efficiency between the size of rules and candidate terms.

5. EXPERIMENTAL RESULTS

For our incremental approach with extended variation rules, we experiment with 2,950 Web documents of 7 domains. The scope of our experiment is from the extraction of uninterrupted patterns by statistical approach to the incremental acquisition of terms. We used stop-word dictionary and pattern removal rules to filter the patterns.

Table 4 is the result of iterations to acquire new reference terms (these are the candidate terms of a previous iteration). The ratio of reference terms-0 is only 1.86%, when compared with original uninterrupted patterns. This shows an efficiency of our pattern filtering. The number of iteration is from 1 ("home office" domain) to 6 ("arts" domain), which increases as the size of reference terms-1 grows up.

Table 4. Incremental terms acquisition for 7 domains.
(Use only two-word terms, reference terms-0: basic terms after pattern filtering, reference terms-*i*: acquired new terms after *i*th iteration)

Domain	# of documents	Before filtering	After filtering	Reference terms-1	Reference terms-2	Reference terms-3	More than reference terms-4
Arts	562	12723	310	75	7	2	4
Commerce	119	2910	82	19	1	0	0
Home office	184	3941	68	14	0	0	0
News	1251	44386	495	74	4	0	0
Reference	487	11315	386	92	6	0	0
Society	253	5368	131	25	6	2	1
Travel	94	1613	55	17	1	0	0

We used three kinds of extended variations (inter-, intra-, and combined variation) which consist of 27 rules for two-word terms and 9 for three-word. This size is small than that of Jacquemin's (73 metarules only for inter variation), but the add-up of rules is simple and easy.

Table 5 shows the number of acquired terms according to the types of variations. The ratio is 38.6% for inter variation (coordination: 12.4%, insertion: 19.8%, permutation: 6.4%), 55.7% for intra variation (pre-intra variation: 9.8%, post-intra variation: 45.9%), and 5.7% for combined variation (pre-intra variation + inter variation: 1.8%, post-intra variation + inter variation: 3.9%).

Table 5. A result of acquired terms according to the types of variations.
(Use only two-word terms, Coor(): coordination, Ins(): insertion, Per(): permutation, Pre(): pre-intra variation, Post(): post-intra variation, Pre()+: pre-intra variation + inter variation, Post()+: post-intra variation + inter variation)

Domain	Inter variation			Intra variation		Combined variation	
	Coor()	Ins()	Per()	Pre()	Post()	Pre()+	Post()+
# of rules	5	4	6	2	2	4	4
Arts	11	24	2	4	54	1	3
Commerce	1	8	0	0	11	0	0
Home office	5	0	2	2	5	0	0
News	6	8	6	9	45	1	3
Reference	10	25	8	18	34	1	2
Society	7	6	1	2	16	0	2
Travel	3	2	0	1	11	0	1

The following is a part of candidate terms acquired from Web documents of news domain.

Candidate Term [Coordination]: biological weapons
Candidate Term [Coordination]: Deputy Minister
Candidate Term [Coordination]: public sector
Candidate Term [Insertion]: Hemisphere Securities
Candidate Term [Insertion]: Marawila Resorts
Candidate Term [Insertion]: Prince protege
Candidate Term [Permutation]: Post coverage
Candidate Term [Permutation]: Saleem Raja
Candidate Term [Post-Intra variation]: Civil War-era
Candidate Term [Post-Intra variation]: Daily News-Link
Candidate Term [Post-Intra variation]: finance ministers
Candidate Term [Post-Intra variation]:
Kemper Reinsurance
Candidate Term [Post-Intra variation]: SportsDaily News
Candidate Term [Pre-Intra variation]: ex-prime ministers
Candidate Term [Pre-Intra variation]: Indo-Sri Lanka
Candidate Term [Pre-Intra variation]:
telephone privatisation
Candidate Term [Post-Intra-Inter variation]:
former protege
Candidate Term [Post-Intra-Inter variation]:
ratchet-up-the-suspense approach

6. CONCLUSION

When compared with previous once-and-all processing algorithms to extract terms, incremental symbolic approach enables us to acquire more domain-specific terms and to get the semantic relation between reference-, and candidate terms. We extended variations for the symbolic approach to include inter-, intra-, and combined ones. It experimentally produced more than twice as many new terms that are semantically related with original ones (reference terms). We also consistently described the metarules regardless of the kind of variations, which helps to get the simplicity and extendibility.

The first future work is to extend the metarules to more kinds and number to acquire more various terms. The second is to strengthen pattern-filtering mechanisms to select only domain-specific and meaningful reference terms, third, practically apply our results to NLP systems such as domain recognizer and automatic indexing.

7. REFERENCES

[Armstrong 1994] S. Armstrong (Eds.), *Using Large Corpora*, MIT Press, 1994.

[Bourigault 1993] D. Bourigault, An Endogeneous Corpus-based Method for Structural Noun Phrase Disambiguation. *In Proceedings of the 6th European Chapter of the Association for Computational Linguistics*, 1993.

[Church & Hanks 1989] K. Church and P. Hanks, Word Association Norms, Mutual Information and Lexicography, *In Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, 1989.

[Haruno *et al.* 1996] M. Haruno, S. Ikehara, and T. Yamazaki, Learning Bilingual Collocations by Word-Level Sorting, *In Proceedings of COLING*, 1996.

[Ikehara *et al.* 1996] S. Ikehara, S. Shirai, and H. Uchino, A Statistical Method for Extracting Uninterrupted and Interrupted Collocations from Very Large Corpora, *In Proceedings of COLING*, 1996.

[Jacquemin 1995] C. Jacquemin, A symbolic and Surgical Acquisition of Terms through Variation, *In Proceedings of the Workshop "New Approaches to Learning for NLP" at the 14th International Joint Conference on Artificial Intelligence*, 1995.

[Jung *et al.* 1998a] H. Jung, Y. Kim, T. Kim, and D. Park, A Domain Identifier Using Domain Keywords from Balanced Web Documents, *In Proceedings of the First International Conference on Language Resource and Evaluation*, 1998.

[Jung *et al.* 1998b] H. Jung, Y. Kim, T. Kim, and D. Park, The Construction of English/Korean Bilingual Pattern from Morpheme-based Uninterrupted Pattern Extraction, *In Proceedings of the 9th KIPS Spring Conference (Korean)*, 1998.

[Sim *et al.* 1998] C. Sim, H. Jung, S. Yuh, T. Kim, D. Park, and H. Kwon, An Implementation of English-to-Korean Machine Translation System for HTML Documents, *In Proceedings of the IASTED International Conference on Artificial Intelligence and Soft Computing*, 1998.