(제 10회 한글 및 한국어 정보처리 학술대회)

확장된 변이 규칙을 이용한 용어의 점진적 획득

정한민*
전자통신연구원
지식처리연구팀
jhm@mail.etri.re.kr

김영길
전자통신연구원
지식처리연구팀
ykkim@ mail.etri.re.kr

최승권
전자통신연구원
지식처리연구팀
choisk@ mail.etri.re.kr

An Incremental Acquisition of Terms Using Extended Variation Rules

Hanmin Jung*
Knowledge Processing Team
ETRI

Young-Kil Kim
Knowledge Processing Team
ETRI

Sung-Kwon Choi
Knowledge Processing Team
ETRI

요약

자연어 처리 응용 분야에서 다양하게 이용할 수 있는 용어들을 자동적으로 획득하고, 나아가 이 과정을 점진적으로 반복하여 수행함으로써 획득할 수 있는 용어의 수를 증가 시키고 그 용어들 간의 의미적 관계도 얻을 수 있다. 점진적인 용어 획득을 위하여 용어의 형태에 변이 규칙을 적용하여 새로운 용어를 획득하는 과정을 반복한다. 우리는 변이의 종류를 단어 간의 변이 뿐만 아니라 단어 내의 변이 그리고 이 둘을 결합한 복합 변이로까지 확장하여 새로운 용어 획득 과정을 더욱 다양화하는 기법을 제시한다. 실험은 확장된 변이 규칙으로부터 얻은 용어들 중에서 기존의 단어 간의 변이로부터 획득한 용어들의 비율이 전체의 38.6%라는 사실로부터 변이의 종류 및 규칙의 확장이 획득할 수 있는 용어들의 수를 증가 시킬 수 있다는 것을 보여준다.

1 서론

용어들은 자연어 처리 응용 분야에서 다양하게 사용될 수 있다. 획득된 용어들은 단어 사전이나 패턴 사전의 구축을 통한 기계 번역에 이용할 수 있으며, 전문 분야로부터 추출하여 해당 전문 분야를 인식할 수 있게 하고, 이를 그 분야에 대한 문맥적 지식으로 활용하여 예측을 이용한 자연어 처리를 가능하게 하며, 자동 인덱싱과 같은 분야에서도 활용이 가능하다. 시간적, 인력적인 비효율성을 제고하고자 자동적인 용어 획득에 관한 연구가 많이 이루어졌다. [Church & Hanks 1989] [Ikehara et al. 1996]의 통계학적인 방법이나 [Bourigault 1993]의 기호적인 방법들을 제시한 연구들이 있었으며, [Haruno et al. 1996]의 양 언어의 말뭉치를 위한 정렬된 패턴 추출에 관한 연구도 있었다. 그렇지만, 이들은 모두 선험적인 지식 (미리 정의되거나 얻어진 용어를 포함하는)을 이용하지 않고 한번의 처리 과정 (once-and-all process)을 통해 말뭉치로부터 용어들을 얻고자 하는 방법들이다. 결국, 이들은 말뭉치로부터 얻을 수 있는 용어들의 수를 극대화하지 못하고, 얻어진 용어들 간의 의미 관계를 파악할 수 없게 만드는 문제점들을 가진다. Jacquemin은 이들을 극복하고자 점진적인 방법으로 용어들을 획득할 수 있는 방법을 제시하였다 [Jacquemin 1995]. 이 방법은 미리 선별된 용어들 (참조 용어라고 정의)을 이용하여 이들이 가지는 단어들의 순서나 다른 단어들과의 관계를 이용하여 새로운 용어들을 획득하는 것으로, 이 획득 과정이 더 이상의 새로운 용어를 얻을 수 없을 때까지 반복하여 일어난다. 그렇지만, 이 알고리즘은 어떻게 참조 용어를 얻을 수 있을 것인가 하는 문제와 단어

간의 변이 만을 고려한 제한된 형태로부터의 한정된 용어 획득이라는 단점을 가진다.

우리는 Jacquemin의 방법을 개선하여 참조 용어를 자동적으로 획득할 수 있도록 하고, 단어 간의 변이 뿐만 아니라 다양한 형태의 변이를 이용하여 용어 획득을 극대화하는 기법을 제시한다. 참조 용어를 자동적으로 얻을 수 있도록 기존의 한번의 처리 과정을 갖는 통계학적인 용어 획득 방법을 이용하며, 이 결과를 자동적으로 필터링하여 전문가에 의해 잘 선정된 참조 용어와 유사한 효과를 낼 수 있도록 한다. Ikehara의 통계학적인 용어 획득 알고리즘을 개선하여 일차적으로 참조 용어 후보들 (두 단어 이상의 연속 패턴)을 얻으며, 불용어 사전과 패턴 제거 규칙을 이용하여 참조 용어를 자동적으로 선정한다 (실험적으로 약 98.14%의 부적당한 참조 용어 후보들을 제거). 변이의 형태를 다양하게 하여 쉽게 발견할 수 없는 형태의 용어들을 획득하는 것 또한 점진적인 방법의 성능을 향상시키는 좋은 방법이다. 우리는 기존의 단어 간의 변이 이외에도 단어 자체의 변형을 고려하는 단어 내의 변이와, 단어 간의 변이와 단어 내의 변이가 결합한 형태인 복합 변이까지도 고려한다. 실험에서는 기존의 단어 간의 변이만을 고려한 시스템보다 많은 새로운 용어들을 획득할 수 있다는 사실을 보여준다. 본 논문은 자동 획득한 참조 용어에 다양한 변이들을 적용하는 것에 초점을 맞추어 기술한다.

2 참조 용어의 자동 획득

참조 용어는 변이 규칙을 적용하여 새로운 용어들을 얻기 위한 기본 용어를 의미한다. 참조 용어의 올바른 획득과 변이 규칙의 적절한 기술은 새롭게 획득한 용어들의 신뢰성을 증가 시킨다. 참조 용어의 올바른 선택을 위하여 해당 분야 전문가들에 의한 수동적인 용어 선택이 가장 바람직하지만, 다량의 말뭉치를 대상으로 할 경우에는 현실적으로 이 작업이 불가능해진다. 이를 대체하기 위한 방안으로 자동적인 참조 용어의 획득이 필요하다. 참조 용어의 자동 획득을 위하여 우리는 Ikehara의 통계학적인 방법에 근간하여 연속 패턴 추출을 수행한다 [Ikehara et al. 1996] [김영길 외 1998]. 비록, Ikehara의 방법이 [Nagao & Mori 1994]의 N-gram 방식을 개선하여 불필요한 패턴들의 생성을 줄였지만, 그의 방법은 여전히 시간적으로 비효율적인 면을 가지고 있다. 이에 우리는 비효율적인 패턴 추출 시간을 줄이기 위하여 Reduced PT & SPT 테이블을 이용한다 [정한민 외 1998]. 이 테이블들은 유효하다고 인정되는 패턴 후보들만을 저장하고 정렬하여 전체 테이블을 대상으로 하는 것보다 효율적인 작업을 할 수 있게 한다.

추출된 연속 패턴들을 점진적 용어 획득을 위한 참조 용어로 이용하기 위한 과정을 자동화하기 위하여 우리는 불용어 사전과 패턴 제거 규칙을 이용하여 참조 용어로 부적당한 패턴들을 제거한다. 불용어 사전은 우리가 개발한 영한 기계번역 시스템인 FromTo/EK의 해석 사전 중 관사, 전치사, 조동사, 부사, 접속사, 대명사 사전 등으로부터 추출한 단어들과 웹 문서 3,600 여 페이지 (25 개 분야)로부터 실험적으로 추출한 단어들로 이루어져 있으며, 현재 약 580 여 엔트리로 구성되어있다 [Sim et al. 1998] [Jung et al. 1998]. 부적당한 패턴들을 제거하기 위한 패턴 제거 규칙은 영문자 검사와 특수 문자 및 문자열 포함 여부를 검사하는 30 여 개의 규칙으로 구성된다.

3 확장된 변이 규칙

변이 (variation)는 참조 용어로부터 새로운 용어를 얻어내기 위하여 참조 용어를 변형시키는 것을 의미한다. 변형의 대상이 되는 참조 용어는 자신을 구성하는 한 단어 또는 두 단어 이상을 다른 단어와 교체하거나, 다른 단어를 삽입하는 등의 외부적인 변형 (우리는 이를 단어 간의 변이로 정의)을 통해서, 또는 접사나 단어를 이용하여 특정 단어를 파생시키는 내부적인 변형 (우리는 이를 단어 내의 변이로 정의)을 통해서, 또는 내부적 변형과 외부적 변형이 결합된 복합적인 변형 (우리는 이를 복합 변이로 정의)을 통해서 새로운 참조 용어로 재생산된다. 변이를 통하여 얻어진 용어들은 그 용어들의 근원이 된 참조 용어와 개념적인 관계로 연결된다. 예를 들어, 용어 "A B"가 말뭉치 내에서

나타난 "A B and C" (머리어인 B와 C가 대등 접속된 형태)로부터 "A C"라는 용어를 새롭게 얻었다면, 이 두 용어들의 머리어는 같은 의미 범주에 속한다는 사실을 알 수 있다 (예. "A B and C" -> "surgical exploration and closure"). Jacquemin은 변이의 적용 범위를 단어 간의 변이 (inter variation)로 한정하고, 이 변이의 세부 범주를 크게 coordination, insertion, 그리고 permutation으로 나누었다 (실제로 Jacquemin은 단어 간의 변이라는 용어를 사용하지 않았지만, 그가 사용한 모든 변이들의 종류가 이 단어 간의 변이에 속한다).

변이의 기술은 메타 규칙으로 하며, 참조 용어를 기술한 규칙은 이 메타 규칙에 의해 새로운 용어를 기술한 규칙으로 바뀌게 된다. 메타 규칙은 변이의 종류에 따라, 변이를 적용하는 용어의 단어 수에 따라 구분할 수 있다. 그렇지만, 우리는 이와 같은 단어 간의 변이 이외에도 단어 내의 변이 (intra variation)와 복합 변이 (inter-intra variation) 모두를 일관성 있는 기술 방식으로 기술하여, 기술 방식의 효율성을 높이고 확장성을 용이하게 한다.

3.1 단어 간의 변이

단어 간의 변이는 크게 세 종류로 나누어진다.

Coordination: 참조 용어와 대등 접속된 정보를 이용하여 새로운 용어를 추출
Insertion: 참조 용어와 삽입된 형태의 단어로부터 새로운 용어를 추출
Permutation: 참조 용어의 치환된 형태와 매개 정보를 이용하여 새로운 용어를 추출

변이의 기술을 위한 메타 규칙은 변이의 종류와 용어의 단어 수에 따라 나누어지는데, 4 단어 이상으로 구성된 용어를 위한 메타 규칙은 그들의 출현 빈도가 너무 낮아 생략한다 (전체 변이의 약 1%) [Jacquemin 1995]. 결국, 단어 간의 변이를 위한 메타 규칙의 범주는 변이의 종류에 따라 세 가지로, 다시 용어의 단어 수에 따라 두 가지로 하여 6 가지가 된다. 메타 규칙의 기술에서 X는 임의의 단어를, C는 임의의 대등 접속사를, N은 참조 용어에 속한 단어를 의미한다. 다음은 두 단어의 coordination을 위한 메타 규칙과 용어를 기술한 규칙의 예를 보여준다.

참조 용어를 기술한 규칙
N1 -> N2 N3
두 단어의 coordination을 위한 메타 규칙
Coor(X1 -> X2 X3) ≡ X1 -> X2 C4 X5 X3
새로운 용어를 기술한 규칙
N1 -> N2 C4 X5 N3
(C4는 임의의 대등 접속사)
획득된 새로운 용어
X5 N3 (X5는 임의의 단어)

다음은 참조 용어인 business visas로부터 위의 메타 규칙을 적용하여 새로운 용어인 student visas를 얻어내는 예를 보여준다.

말뭉치 내의 문장들
CHARGE: BUSINESS & STUDENT VISAS US$50
REQUIREMENTS: BUSINESS & STUDENT VISAS
참조 용어를 기술한 규칙
N1 -> business visas
두 단어의 coordination을 위한 메타 규칙
Coor(X1 -> X2 X3) ≡ X1 -> X2 C4 X5 X3
새로운 용어를 기술한 규칙
N1 -> business [and or & ...] X5 visas
(X5는 student)
획득된 새로운 용어
student visas

표 1. 두 단어로 구성된 참조 용어를 위한 단어 간의 변이에 따른 메타 규칙들의 예.

변이	메타 규칙	획득된 새로운 용어
Coordination	Coor(X1 -> X2 X3) ≡ X2 X4 C5 X3 Coor(X1 -> X2 X3) ≡ X2 C4 X5 X6 X3 Coor(X1 -> X2 X3) ≡ X2 X5 C4 X2 X3	X2 X4 X5 X6 X3 X2 X5
Insertion	Ins(X1 -> X2 X3) ≡ X2 X4 X3 Ins(X1 -> X2 X3) ≡ X2 X3 X4 X3	X4 X3 X4 X3
Permutation	Per(X1 -> X2 X3) ≡ X2 P4 X5 X3 (P4 = for) Per(X1 -> X2 X3) ≡ X2 X5 P4 X6 X3 (P4 = for, from)	X5 X3 X2 X5

3.2 단어 내의 변이

단어 내의 변이는 용어의 특정한 단어에서 발생하는 변이를 의미하며, 해당 단어와 접사 또는 단어가 결합되면서 발생한다. 결합되는 위치에 따라 pre-intra variation (예. Affirmative Action -> Anti-Affirmative Action, American women -> African-American women)과 post-intra variation (예. women business -> women-owned

business, Puerto Rican -> Puerto Ricans)으로 나눈다. 단어 간의 변이와는 달리 단어 내의 변이는 기존의 참조 용어 내의 단어가 삭제되는 경우는 없으며, 참조 용어 내의 단어가 새로운 단어로 대치된다.

Pre-intra variation: 접두사 또는 단어가 참조 용어의 특정한 단어의 앞에서 결합한 새로운 용어를 추출
Post-intra variation: 접미사 또는 단어가 참조 용어의 특정한 단어의 뒤에서 결합한 새로운 용어를 추출

단어 내의 변이를 위한 메타 규칙의 기술은 이전의 단어 간의 변이를 위한 메타 규칙의 기술 방식을 따른다. 다음은 두 단어의 pre-intra variation을 위한 메타 규칙과 용어를 기술한 규칙의 예를 보여준다. 접두사 또는 단어와 결합하여 만들어진 새로운 단어는 { } 내에 기술한다.

참조 용어를 기술한 규칙
N1 -> N2 N3
두 단어의 pre-intra variation을 위한 메타 규칙
Pre(X1 -> X2 X3) ≡ X1 -> {R4X2} X3
새로운 용어를 기술한 규칙
N1 -> {R4N2} N3
(R4는 임의의 접두사 또는 단어)
획득된 새로운 용어
{R4N2} N3

다음은 두 단어의 post-intra variation을 위한 메타 규칙과 용어를 기술한 규칙의 예를 보여준다.

참조 용어를 기술한 규칙
N1 -> N2 N3
두 단어의 post-intra variation을 위한 메타 규칙
Post(X1 -> X2 X3) ≡ X1 -> X2 {X3O4}
새로운 용어를 기술한 규칙
N1 -> N2 {N3O4}
(O4는 임의의 접미사 또는 단어)
획득된 새로운 용어
N2 {N3O4}

표 2. 둘, 세 단어로 구성된 참조 용어를 위한 단어 내의 변이에 따른 메타 규칙들의 예.

변이	메타 규칙	획득된 새로운 용어
Pre-intra variation	Pre(X1 -> X2 X3) ≡ X2 {R4X3} Pre(X1 -> X2 X3 X4) ≡ {R5X2} X3 X4	X2 {R4X3} {R5X2} X3 X4
post-intra variation	Post(X1 -> X2 X3) ≡ {X2O4} X3 Post(X1 -> X2 X3 X4) ≡ X2 {X3O5} X4	{X2O4} X3 X2 {X3O5}X3

3.3 복합 변이

복합 변이는 단어 간의 변이 뿐만 아니라 단어 내의 변이가 동시에 일어나는 변이를 의미한다. 이 변이는 단어 간의 변이인 coordination, insertion, permutation의 세 가지와 단어 내의 변이인 pre-intra variation, post-intra variation의 두 가지의 총 6 종류의 변이로 구성된다. 복합 변이에서는 용어 내의 단어가 삭제 또는 대치되는 경우가 발생한다. 이는 단어 간의 변이와 단어 내의 변이 현상 모두를 반영한 결과이기 때문이다. 다음은 단어 간의 insertion 변이와 단어 내의 pre-intra variation이 결합한 복합 변이를 위한 메타 규칙을 보여준다.

참조 용어를 기술한 규칙
N1 -> N2 N3
두 단어의 insertion과 pre-intra variation을 위한 메타 규칙
InsPre(X1 -> X2 X3) ≡ X1 -> {R4X2} X5 X3
새로운 용어를 기술한 규칙
N1 -> {R4N2} X5 N3
(R4는 임의의 접미사 또는 단어)
획득된 새로운 용어
{R4N2} N3

표 3. 두 단어로 구성된 참조 용어를 위한 복합 변이에 따른 메타 규칙들의 예.

변이	메타 규칙	획득된 새로운 용어
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}

표 3에서 보듯이, 하나의 메타 규칙이 참조 용어를 위한 규칙에 적용될 경우에 반드시 하나의 용어만을 획득할 수 있는 것은 아니다. 메타 규칙이 하나이더라도, 획득된 새로운 용어를 두 개 이상으로 정의할 수 있다. Insertion과 pre-intra variation을 가지는 복합 변이를 위한 메타 규칙 "InsPost(X1 -> X2 X3) ≡ X2 X4 {R5X3}"으로부터 "X4 {R5X3}"과 "X2 {R5X3}"의 두 개의 새로운 용어를 획득할 수 있다.

4 점진적 방법을 이용한 용어 획득

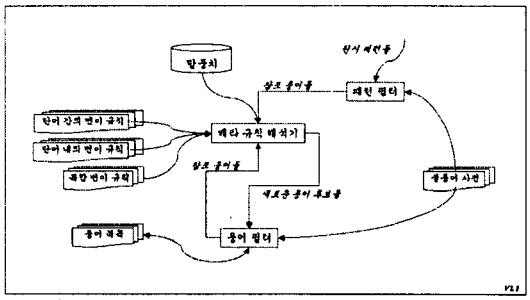

그림 1. 점진적 방법을 이용한 용어 획득 시스템의 구성도.

$n(T_i) = n(T_{i-1}) + n(R_i)$, repeat until $n(R_i) = 0$, where $0 \leq n(R_j) < n(R_{j-1})$ and $1 \leq j \leq i$
$n(T) \geq n(T_0) + n(T_1) + \ldots + n(T_i)$ ($\because S(R_x) \subset S(T)$)
(n(T): 말뭉치 내의 가능한 모든 용어들의 수, $n(T_i)$: i번째 반복까지 얻어진 전체 용어들의 수, $n(R_i)$: i번째 반복에서 새롭게 얻어진 용어들의 수, $S(R_x)$: j번째 반복에서 새롭게 얻어진 용어들의 집합, S(T): 말뭉치 내의 가능한 모든 용어들의 집합, n(T)가 $n(T_0) + n(T_1) + \ldots + n(T_i)$와 같은 경우는 변이 규칙이 완벽하여 recall이 100%인 경우이며, 일반적으로는 규칙이 완전하지 않아 n(T)가 크다.)

점진적 방법을 이용한 용어 획득을 위해서는 획득된 새로운 용어들을 다시 참조 용어로 되돌려서 다음 단계의 용어 획득을 유도하는 과정이 필요하다. 자동 또는 수동으로 얻어진 원시 패턴들이 불용어 사전을 통해 1차 참조 용어들로 결정된다. 이들은 메타 규칙 해석기에 의해 새로운 용어들을 획득할 수 있도록 해주는 역할을 한다. 메타 규칙 해석기는 확장된 변이 규칙들인 단어 간의 변이 규칙, 단어 내의 변이 규칙, 복합 변이 규칙을 이용하여 새로운 용어 후보들을 제시한다. 불용어 사전에 의해 불필요한 용어 후보들이 걸러지면, 기존의 용어 목록과 비교하여 새로운 용어들만 남아 용어 목록에 저장되며, 이들은 다시 2차 참조 용어들이 되어 또 다른 용어 획득을 위하여 사용된다. 이 과정은 더 이상의 새로운 용어를 발견할 수 없을 때까지 반복된다. 용어를 획득하기 위한 말뭉치의 크기가 무한하지 않으며, 변이 규칙의 수와 종류가 제한되어 있고, 이 변이 규칙이 말뭉치에 없는 새로운 용어를 생성해 내는 것이 아니므로, 결국 점진적인 용어 획득 과정은 무한하게 반복되지 않는다.

5 실험 결과

점진적 용어 획득을 위한 참조 용어의 자동 구축을 위하여 우리는 웹으로부터 수집한 7 개 분야의 2950 개 문서를 대상으로 실험을 수행하였다. 실험은 연속 패턴 추출 다음 단계부터 점진적 용어 획득까지의 과정을 대상으로 하였으며, 연속 패턴 추출 결과의 필터링을 위하여 불용어 사전과 패턴 제거 규칙을 이용한 패턴 필터를 사용하였다.

표 4는 점진적 용어 획득을 통하여 획득되는 참조 용어들을 각 회수별로 구분한 결과를 보여준다. 패턴 필터에 의하여 불필요한 패턴들이 제거되고 남은 참조 용어0의 비율은 패턴 필터링 이전의 수와 비교할 때, 1.86%에 불과하다. 즉, 우리의 자동적인 참조 용어 획득 방법은 절대 다수의 불필요한 용어들을 제거하는 효과를 보여준다. 점진적 방법의 반복 회수는 실험에서 최소 1회 (Home Office), 최대 6회 (Arts 분야)이다. 반복 회수는 참조 용어0가 많을수록 그 회수도 증가한다는 것을 알 수 있다. 참조 용어 획득을 위한 메타 규칙이 확장되면, 회수도 그에 따라 증가할 것이다.

표 4. 분야별 말뭉치에 따른 점진적 용어 획득. (두 단어 참조 용어만 사용, 참조 용어0: 점진적 용어 획득을 위한 기본 참조 용어, 참조 용어i: i회 반복 후에 획득된 참조 용어)

분야	문서수	추출패턴수	참조용어 0	참조용어 1	참조용어 2	참조용어 3	참조용어 4 이상
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

우리는 변이를 이용한 점진적 용어 획득을 위하여 세 종류의 변이 (단어 간의 변이, 단어 내의 변이, 복합 변이)에 대하여 두 단어의 용어를 위한 27 개와 세 단어의 용어를 위한 9 개의 메타 규칙을 적용하였다. 이 수는 Jacquemin의 73 개의 메타 규칙 (단어 간의 변이만을 고려한)에 비하여 작지만, 이는 쉽게 계속적으로 확장해 나갈 수 있는 부분이다. 그렇지만, 메타 규칙의 수가 증가할수록 그로 인하여 발생할 수 있는 잘못된 용어들의 추출이라는 Trade-Off가 생길 수 있으며, 이 문제는 앞으로 해결해야 할 과제이다.

표 5는 획득된 참조 용어들을 얻는 과정에서 발생한 변이의 종류에 따른 참조 용어들의 수를 분류하여 보여준다. 발생한 변이의 비율을 살펴보면, 단어 간의 변이가 38.6% (coordination: 12.4%, insertion: 19.8%, permutation: 6.4%), 단어 내의 변이가 55.7% (pre-intra variation: 9.8%, post-intra variation: 45.9%), 복합 변이가 5.7% (pre-intra variation + 단어 간의 변이: 1.8%, post-intra variation + 단어 간의 변이: 3.9%)로 나타난다. 단어 간의 변이의 발생 빈도가 단어 내의 변이에 비해 상대적으로 적은 것은 규칙의 수가 15 개에 불과하기 때문이다. 단어 내의 변이의 경우에 4 개의 규칙에 불과하지만, 이는 두 단어의 용어를 위한 모든 경우를 고려한 것이다. 단어 간의 변이를 위한 메타 규칙의 수를 확장할 경우, 그에 따른 발생 빈도도 상승하게 된다. 복합 변이도 마찬가지인데, 이 경우에는 기술하는 메타 규칙의 형태가 다른 두 종류의 변이의 결합 형태이므로 상대적으로 더 복잡하게 나타난다. 이들에 대한 규칙을 확장하면 역시 발생 빈도가 증가한다.

표 5. 분야별 말뭉치에 따라 획득된 참조 용어에 대한 변이 종류.

(두 단어 참조 용어만 사용, Coor(): Coordination, Ins(): Insertion, Per(): Permutation, Pre(): Pre-intra variation, Post(): Post-intra variation, Pre()+: Pre-intra variation + 단어 간의 변이, Post()+: Post-intra variation + 단어 간의 변이)

분야	단어 간의 변이			단어 내의 변이		복합 변이	
	Coor	Ins	Per	Pre	Post	Pre+	Post+
규칙 수	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

다음은 News 분야에서 획득된 참조 용어들의 예를 보여준다.

Candidate Term [Coordination]: Deputy Minister
Candidate Term [Coordination]: public sector
Candidate Term [Insertion]: Hemisphere Securities
Candidate Term [Insertion]: Marawila Resorts
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]: 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 결론

기존의 한번의 처리 과정을 통해 말뭉치로부터 용어들을 얻는 지식 획득에 비하여 점진적인 방법은 의미적 연관성을 내포하는 용어들을 순차적으로 획득하고, 그 획득의 비율을 높일 수 있다는 장점이

있다. 우리는 이 방법을 더욱 발전시켜서 보다 다양한 변이들과 그에 따른 메타 규칙 기술 방식을 정의하고, 그들로부터 이전에는 얻지 못했던 새로운 용어들을 획득할 수 있도록 하였다. 단어 간의 변이 뿐만 아니라 단어 내의 변이와 이 둘이 결합된 복합 변이를 이용하는 것이 두 배 가까운 용어들을 새롭게 획득할 수 있다는 것을 실험적으로 보여주었다. 또한, 메타 규칙을 이용한 변이의 기술은 단순하면서도 명확하여 쉽게 새로운 변이를 위한 규칙을 추가할 수 있게 한다.

앞으로 우리는 새롭게 정의한 변이들에 대한 명확한 의미적 관계를 증명하고, 메타 규칙을 확장하여 말뭉치로부터의 용어 획득의 효율성을 최대로 높일 예정이다. 또한, 참조 용어들과 획득한 용어들의 검증 및 필터링을 위한 불용어 사전과 패턴 제거 규칙을 강화할 것이다.

참고문헌

[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. 1998] 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.

[Nagao & Mori 1994] M. Nagao and S. Mori, *A Comparative Study of Automatic Extraction of Collocations from Corpora: Mutual Information vs. Cost Criteria*, Journal of Natural Language Processing, Vol. 1, No. 1, 1994.

[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.

[김영길 외 1998] 김영길, 정한민, 여상화, 장원, 김태완, 박동인, 한영 기계번역 시스템을 위한 숙어 및 격 틀의 자동 추출, *정보처리학회 춘계 학술대회*, 1998.

[정한민 외 1998] 정한민, 김영길, 김태완, 박동인, 형태소 기반 연속 패턴 추출을 이용한 영한 대역 패턴 생성, *정보처리학회 춘계 학술대회*, 1998.