말뭉치에 기반한 상호정보를 이용한 연어의 자동 추출

이 호 석†

요 약

본 논문은 말뭉치에 기반한 연어의 자동 추출에 관한 연구이다. 연어는 말뭉치로부터 단어의 동시발생빈도(cooccurrence frequency)와 상호정보(mutual information)를 이용하여 추출하였다. 영어에는 5가지 종류의 연어가 정의되어 있다. 이들은 타동사와 목적어, 자동사와 주어, 형용사와 명사, 동사와 부사 그리고 형용사와 부사이다. 여기에 동사와 전치사의 단어쌍을 새롭게 연어로 파악하여 6가지 종류의 연어를 추출하였다.

Automatic Extraction of Collocations based on Corpus using mutual information

Ho Suk Lee†

ABSTRACT

This paper describes the automatic extraction of collocations based on corpus. The collocations are extracted from corpus using cooccurrence frequency and mutual information between words. In English, 5 types of collocations are defined. These collocations are transitive verb and object, intransitive verb and subject, adjective and noun, verb and adverb, and adverb and adjective. In this paper another type of collocation is recognized and extracted, which consists of verb and preposition. So 6 types of collocations are extracted based on corpus.

1. 서 론

현대 언어 연구의 한가지 주류에는 말뭉치[7, 9]를 기반으로 한 방법이 있다. 말뭉치를 기반으로 한 언어의 연구는 확률적 방법을 이용하여 말뭉치에 존재하는 단어들 사이의 관계를 조사함으로써 행하여진다. 확률적 조사에 이용되는 정보에는 말뭉치의 크기, 단어의 빈도, 단어와 단어사이의 동시발생빈도가 있다. 확률적 조사 방법은 주관적인 단어의 의미적 정보 대신 말뭉치로부터 단어의 빈도를 구하여 조사하는 것이기 때문에 말뭉치에 기반한 객관적인 결과를 얻을 수가 있다.

본 논문은 말뭉치를 기반으로 하여 현대 언어 연구에서 중요한 비중을 차지하는 연어의 자동 추출에 대하여 제시하고자 한다. 영어에는 일반적으로 5가지 종류의 연어가 정의되어 있다[11]. 이 5가지 종류의 연어는 타동사와 명사(목적어), 자동사와 명사(주어), 형용사와 명사, 동사와 부사 그리고 부사와 형용사이다. 본 논문은 이외에도 동사와 전치사의 연어를 추가하여 모두 6가지의 연어를 말뭉치로부터 자동으로 추출하는 방법에 대하여 제시하고자 한다. 본 논문에서 제시하는 연어는 영어-한국어 기계번역의 관점에서 정확한 목표언어 표현을 얻고자 할 때 반드시 고려되어야 하는 단어들의 쌍이다. 이들 추출된 연어들은 기계번역용 사전에 수록되어 정확한 목표언어 표현의 생성에 활용되고 있다.

말뭉치로부터 연어를 추출하려고 하는 시도는 [6, 8, 10]에도 보고되었다. 이 연구에서는 연어 추출 과정은 3단계로 이루어진다. 첫 번째 단계

† 정 회 원: 호서대학교 컴퓨터공학과 교수
논문접수: 1994년 7월 8일, 심사완료: 1994년 11월 2일

는 통계적인 방법만을 이용하여 연어로서의 가능성이 있는 bigram(2-word pair)들을 추출한다. 두 번째와 세 번째 단계는 모두 첫 번째 단계의 출력을 이용하며 두 번째 단계는 bigram으로부터 n-gram을 만들어 낸다. 세 번째 단계에서는 구문분석기를 사용하여 추출된 연어들의 문장내에서의 역할을 파악한다. 그리고 이 단계에서 문장내에서의 역할이 파악되지 않아 연어로서의 가능성이 없는 후보들은 제외된다. 그러나 본 논문에서는 구문 분석기를 먼저 사용하여 단어 쌍들을 모두 추출한 후 이들 단어 쌍에 통계적인 방법을 적용하여 연어를 파악하는 방법을 택하고 있다.

본 논문의 방법은 [14]의 방법과 비슷한 면이 있으나 [14]의 연구에서는 동사와 전치사의 관계는 다루어지지 않았다. 참고문헌 [6]은 연어가 언어의 생성에 활용되는 방법을 제시하였다.

참고문헌 [4, 5]에서는 동사와 전치사의 관계를 동사 관용어로 간주하여 조사되어 있다. 이 연구는 말뭉치를 태깅[3]하여 어휘 모호성을 제거한 후 동사 관용어의 조사를 시작한다.

이 밖에도 연어에 대한 연구는 [15]에서 소개되었다. 이들 연구는 사전을 구축하는 관점에서 연어를 조사하였으며 일반사전(human-readable dictionary)로부터 연어를 추출하여 사전을 구축하였다. 이들 방법은 말뭉치로부터 연어를 자동으로 추출하는 방법에 대하여서는 제시하지 못하였다.

연어의 가능성이 있는 단어쌍은 말뭉치의 문장들을 구문분석[12]한 결과로부터 추출한다. 이 결과에 동시발생빈도(cooccurrence frequency)와 상호정보(mutual information)[1, 2]를 적용하여 연어를 추출한다.

영어에는 5종류의 연어가 정의되어 있다. 이들 연어는 타동사와 명사(목적어), 자동사와 명사(주어), 형용사와 명사, 동사와 부사 그리고 부사와 형용사의 쌍이다. 문장을 구문 분석하면 문장을 구성하는 단어의 품사 및 문장 내에서의 역할을 알 수 있으며 이로부터 연어를 추출한다. 동사 연어는 동사와 전치사의 쌍으로 구성된 표현으로 정의하며 구문분석의 결과를 이용하여 추출한다.

2. 상호정보

상호정보는 확률변수 사이에 관한 것으로 확률변수 Y에 의하여 제공되는 확률변수 X의 정보의 양을, 다시 말하면 확률변수 X와 Y사이의 의존관계를 정량적으로 나타낸 것이다. X와 Y대신에 조사하고자 하는 단어들을 대입시키면 단어와 단어 사이의 의존관계를 정량적으로 나타내게 된다. 따라서 이러한 상호정보의 성질은 말뭉치에서 단어와 단어 사이의 상대적 의존 관계를 조사하는데 이용할 수 있다. 상호정보는 다음의 식[2, 4, 5]으로 표현된다.

$$MI(x, y) = \mathrm{Log}\ \frac{P(x, y)}{P(x) \times P(y)}$$

(식 1) 확률을 이용한 상호정보
(Exp. 1) mutual information using probability

이 식에서 P(x)와 P(y)는 단어 x와 y가 전체 말뭉치에서 나타난 확률을 의미하며 P(x, y)는 단어 x와 y가 한 문장에서 함께 나타난 확률을 의미한다. 만약 단어 x와 y사이에 의존관계가 있다면 P(x, y) 값은 P(x)×P(y) 값보다 클 것이며 MI 값은 0보다 큰 숫자가 될 것이다. 만약 단어 x와 y사이에 관계가 크지 않다면 P(x, y) 값과 P(x)×P(y) 값은 큰 차이가 없을 것이며 MI 값은 0에 가까운 값이 될 것이다. 그리고 단어 x와 y사이에 아무런 관계가 없다면 P(x, y) 값은 P(x)×P(y) 값보다 작을 것이며 MI 값은 0보다 작을 것이다.

P(x)와 P(y)는 전체 말뭉치에서 단어 x와 단어 y가 나타낸 빈도를 구하여 이 값을 전체 말뭉치의 크기 N으로 나누어 얻으며 P(x, y)는 단어 x와 단어 y가 한 문장에서 함께 나타난 빈도를 전체 말뭉치의 크기 N으로 나누어 얻는다. 따라서 위의 식은 다음과 같이 말뭉치의 크기와 단어

의 빈도를 이용하여 다시 표현할 수가 있다. F(x)는 단어 x의 발생빈도를 F(y)는 단어 y의 발생빈도를 그리고 F(x, y)는 단어 x와 단어 y가 한 문장에서 동시에 발생한 빈도를 의미한다. 말뭉치의 크기 N은 말뭉치 전체의 단어의 갯수이다.

$$MI(x, y) = \text{Log}\ \frac{N \times F(x, y)}{F(x) \times F(y)}$$

(식 2) 빈도를 이용한 상호정보
(Exp. 2) mutual information using frequency

3. 연어의 정의

연어는 일반적으로는 함께 자주 사용되는 단어들의 쌍으로 정의된다[11, 13]. 이 정의는 언어학의 관점에서 정의된 것이다. 그러나 본 연구에서는 연어를 말뭉치에 기반한 객관적인 데이타를 이용하여 새롭게 정의한다. 객관적인 정보에는 단어의 발생빈도, 단어쌍의 동시발생빈도 그리고 상호정보가 있다. 동시발생빈도는 단어와 단어가 한 문장에서 함께 사용된 빈도를 의미한다. 따라서 동시발생빈도는 말뭉치 전체에서 단어 사용의 절대적인 수치이다. 상호정보는 단어의 발생빈도 그리고 단어쌍의 동시발생빈도를 이용하여 계산하며 단어와 단어가 서로 의존되어 있는 정도를 나타낸다. 이는 상대적인 수치로서 말뭉치 전체에서 단어와 단어의 상대적인 의존관계를 나타낸다. 동시발생빈도와 상호정보는 언어학에서 연어의 정의인 "함께 자주 사용되는"이라는 구절을 정량적으로 계산한 것으로 간주될 수 있다.

4. 연어 추출

영어에는 기존의 5종류의 연어가 있으며 연어 추출은 구문분석의 결과를 이용한다. 구문분석의 결과는 술어논항구조로 표현된다. 술어논항구조에서는 문장을 술어와 술어의 인자로 표현하며 구절도 술어와 술어의 수식부로 표현한다. 예를 들어 전체 문장의 술어는 동사이고 주어, 목적어는 동사의 인자로 간주하며 명사구의 술어는 명사이고 명사를 수식하는 형용사는 명사의 인자로 간주한다.

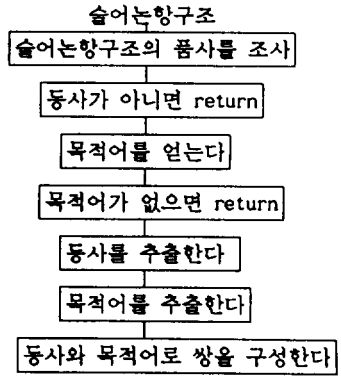
다음은 예제 문장과 이 문장의 술어논항구조이다.

(문장) Users carelessly update very important data.

(술어논항구조)

(그림 1) 술어논항구조
(Fig. 1) predicate-argument structure

문장 전체의 술어는 "update"이고 "update"는 "subj", "obj" 두개의 인자와 수식어 "adv_mods"를 가진다. 이들은 포인터로서 이들이 가리키는 곳에 실제 단어 "user", "data", 그리고 "carelessly"가 있다. 명사구 "very important data"의 술어는 "data"이고 형용사 "important"는 수식어가 되며 부사 "very"는 형용사의 수식어가 된다. 이와 같이 술어논항구조는 술어와 술어의 인자 혹은 술어와 술어의 수식어로 구성되기 때문에 연어를 추출하기에 매우 유리하다. 앞의 술어논항구조로부터 타동사와 명사의 연어(update,



(그림 2) 타동사와 목적어 연어 추출 순서도
(Fig. 2) flowchart for transitive verb + object collocation extraction

data), 형용사와 명사의 연어(important, data), 동사와 부사의 연어(update, carelessly), 그리고 부사와 형용사의 연어(very, important)를 추출할 수가 있다.

대표적으로 타동사와 명사(목적어)의 연어를 추출하는 과정의 순서도를 살펴보자. 술어논항구조를 입력으로 하여 술어논항구조가 동사를 포함하지 않거나 술어논항구조에 목적어가 존재하지 않으면 타동사와 목적어의 연어를 추출할 수 없으므로 상위 모듈로 되돌아 간다.

다음 〈표 1〉을 이용하여 타동사와 명사(목적어) 간의 동시발생빈도와 상호정보 사이의 관계를 알아보자. 타동사 빈도는 말뭉치 전체에 나타난 타동사의 빈도이며 명사(목적어)의 빈도도 말뭉치 전체에 나타난 명사의 빈도이다.

〈표 1〉 타동사와 목적어의 동시발생빈도와 상호정보
〈Table 1〉 co-occurrence frequency and mutual information between transitive verb and object

타동사	명사	타동사빈도	명사빈도	동시발생빈도	상호정보
"accept"	"data"	4	321	2	7.254006
"accept"	"input"	4	83	1	10.205396
"access"	"RLOB"	89	9	1	17.886244
"access"	"body"	89	1	1	21.056170
"access"	"data"	89	321	15	8.822848
"access"	"database"	89	83	8	11.681129
"access"	"file"	89	312	1	12.770766
"access"	"resource"	89	9	3	16.301281
"access"	"table"	89	278	5	10.615299
"access"	"value"	89	199	1	13.419544

〈표 1〉에서 첫 번째 행의 "access + data"의 경우는 동시발생빈도는 2 그리고 상호정보는 7.254006으로 나타났다. 두 번째 행을 보면 "access + input"의 경우에 동시발생빈도는 1 그리고 상호정보는 10.205396으로 나타났다. 두 번째 경우가 상호정보의 값은 더 크게 나타났으나 동시발생빈도가 1 밖에 되지 않기 때문에 절대적인 사용의 측면에서 "access + data"의 경우보다 적게 쓰였음을 알 수 있으며 연어로 취급하기에는 곤란하다. 세 번째와 네 번째의 경우도 동시발생빈도는 모두 1이므로 상호정보의 값은 비록 크다고 할지라도 연어로 취급하기에는 곤란하다. 특히 네 번째의 경우는 "body"라는 명사가 말뭉치 전체에서 사용된 빈도가 1밖에 되지 않기 때문에 "access + body"의 경우를 연어로 취급하기에는 단어 사용의 충분한 사례가 부족하다고 할 수 있다. 다섯 번째 "access + data"의 경우는 동시발생빈도가 15 그리고 상호정보는 8.822848이므로 연어로 취급할 수 있다. 앞에서도 설명했듯이 동시발생빈도는 단어 쌍의 절대적 사용의 경우를 나타내며 상호정보는 상대적 사용의 경우를 나타내므로 이 두 수치가 모두 일정한 값 이상이 되어야 단어 쌍을 연어로서 파악할 수 있다. 동시발생빈도가 낮은 단어 쌍은 절대적인 사용이 빈번치 않은 단어 쌍이므로 연어로 취급하기에 곤란하며 상호정보가 낮은 단어 쌍은 상대적인 사용이 빈번치 않은 단어이므로 역시 연어로 취급하기에 곤란하다. 표의 다음의 예에서 "access + database", "access + resource", "access + table"의 경우는 연어로 취급할 수 있으며 "access + file", "access + value"는 연어로 취급하기에 곤란하다.

다음 〈표 2〉를 이용하여 형용사와 명사 간의 동시발생빈도와 상호정보 사이의 관계를 알아보자. 형용사 빈도는 말뭉치 전체에 나타난 형용사의 빈도이며 명사의 빈도도 말뭉치 전체에 나타난 명사의 빈도이다.

〈표 2〉 형용사와 명사의 동시발생빈도와 상호정보
〈Table 2〉 co-occurrence frequency and mutual information between adjective and noun

형용사	명사	형용사빈도	명사빈도	동시발생빈도	상호정보
"active"	"set"	7	37	5	9.856409
"actual"	"linkage"	3	2	1	15.165398
"all"	"command"	65	185	3	11.486459
"all"	"constraint"	65	29	1	15.744823
"all"	"data"	65	321	7	9.469019

〈표 2〉에서 두 번째와 네 번째 행의 "actual + linkage"와 "all + constraint"는 동시발생빈도가 너무 낮아서 연어로 취급하기에 곤란하다. 첫 번째와 세 번째 그리고 다섯 번째의 경우는 동시발생빈도 그리고 상호정보 모두가 일정한 값 이상이므로 연어로 취급할 수 있다.

자동사와 명사(주어), 동사와 부사, 그리고 부사와 형용사 연어의 추출도 타동사와 명사(목적

어), 형용사와 명사의 연어를 추출하는 방법과 비슷한 방법에 의하여 추출한다.

5. 동사와 전치사 연어

동사 연어는 주어진 말뭉치에서 서로 의존관계에 있는 동사와 전치사의 쌍으로 정의한다. 동사와 전치사의 쌍을 동사 연어로 간주하는 이유는 영어에서는 연어는 앞에서 제시한 5가지의 종류로 정의되어 있기 때문이며, 동사와 전치사의 쌍으로 정의된 관용어는 구문분석의 효율을 크게 향상시키는 반면 연어는 구문 분석의 효율 향상과는 관계가 없기 때문에 동사와 전치사의 쌍을 동사 연어로 간주하였다. 동사와 전치사 쌍의 의존관계는 의미적인 것이다. 그러나 본 연구에서는 의미적 정보를 사용하지 않고 말뭉치로부터 객관적인 정보를 추출하여 동사 연어를 판단한다. 말뭉치로부터 추출할 수 있는 객관적인 정보에는 동사의 빈도, 전치사의 빈도 그리고 동사와 전치사의 동시발생빈도가 있다. 그리고 이들을 이용하여 계산할 수 있는 동사와 전치사의 상호정보가 있다. 동시발생빈도는 동사와 전치사의 절대적인 의존관계를 나타내고 상호정보는 상대적인 의존관계를 나타낸다. 동사 연어는 이들 동시발생빈도와 상호정보를 이용하여 추출한다.

5.1 동사와 전치사 연어 정의

"IBM SQL/DS 지침서" 말뭉치를 살펴보면, 동사와 전치사로 구성된 다음과 같은 표현들을 자주 발견할 수 있다.

(1) You can *add* text *to* the report.
(2) The program *associate* a cursor *with* the described query.
(3) INSERT *places* one or more rows *in* a table.
(4) Other users and applications are *working with* other data.

(예 1) 동사 연어의 예
(Example 1) verb collocation example

동사 연어는 주어진 말뭉치에서 서로 의존관계에 있는 동사와 전치사의 쌍으로 구성된 표현으로 정의하며 타동사와 자동사에 대하여 2종류 생각할 수 있다.

타동사의 경우 : 타동사 + 목적어 + 전치사 + 전치사의 목적어
자동사의 경우 : 자동사 + 전치사 + 전치사의 목적어

타동사는 목적어를 필요로 하므로 타동사와 전치사 사이에는 목적어가 존재하며 자동사는 목적어를 필요치 않으므로 자동사 다음에 전치사가 바로 존재한다.

위 예의 (1)번 문장을 다시 살펴보자.

(1) You can *add* text *to* the report.

이 문장에는 타동사 "add"를 다른 타동사 예를 들면 "associate"로 대체하면 문장은

(1.1) You can *associate* text *to* the report.

로 되며, 또 다른 타동사 "place"로 대체하면 다음의 문장이 된다.

(1.2) You can place text to the report.

이제, 이 문장에서 타동사 대신 전치사의 대체를 생각해 보자. 전치사 "to"를 전치사 "with"로 대체하면, (1.3)의 문장이 되며 전치 "to"를 전치사 "in"으로 대체하면 (1.4)의 문장이 된다.

(1.3) You can add text with the report.

(1.4) You can add text in the report.

문장 (1.1), (1.2), (1.3), (1.4)는 모두 틀린 문장이다. 다시 말하면, 타동사 "add"는 전치사 "to"와 함께 사용되는 경우만이 올바른 문장이 된다. 이 것은 타동사 "add"가 전치사 "to"와 의존관계에 있다는 것을 의미한다.

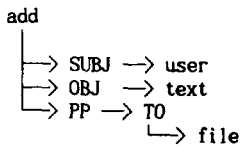
의존관계는 우선 말뭉치에서 동사와 전치사간의 동시발생빈도를 보면 알 수 있다. 즉, 동사와

전치사 간에 얼마나 빈번히 함께 사용되었는 가를 살핌으로서 동사와 전치사간의 의존관계를 알 수가 있다. 그러나 동시발생빈도만을 가지고는 충분하지 못하다. 왜냐하면, 동시발생빈도는 전체 말뭉치에서 동사와 전치사가 얼마나 빈번히 함께 사용되었는 가를 나타내는 절대적인 수치일 뿐 동사와 전치사간의 상대적인 의존관계는 나타내지 않는다. 상대적인 의존관계는 동시발생빈도와는 달리 전체 말뭉치에서 동사와 전치사가 어느 정도의 비율로 의존되어 있는가를 나타낸다. 동사와 전치사간의 상대적인 의존관계는 동사의 발생확률과 전치사의 발생확률의 곱에 대한 동사와 전치사간의 동시발생확률의 비율로서 계산할 수 있다. 이 비율에 Log를 취하면 앞에서 설명한 상호정보식과 동일한 식이 된다.

5.2 동사와 전치사 연어 추출

다음 문장을 예로 들어 동사와 전치사를 포함한 문장에 대하여 살펴보자.

(문장) Users add text to file.
(술어논항구조)



(그림 3) 술어논항구조
(Fig. 3) predicate-argument structure

문장 전체의 술어는 "add"이고 "add"는 "subj", "obj", "pp" 세 개의 인자를 가진다. "pp"는 전치사구를 나타내며 전치사 "to"와 전치사의 목적어 "file"로 구성되어 있다. 이 술어논항 구조로부터 동사 "add"와 전치사 "to"를 추출할 수가 있다.

다음 〈표 3〉을 이용하여 동사와 전치사간의 동시발생빈도와 상호정보 사이의 관계에 대하여 살펴보자.

〈표 3〉에서 타동사 "access + in"의 경우를 보면 동시발생빈도는 5이고 상호정보는 1.3768이다. 그리고 타동사 "add + to"를 보면 동시발생빈도는 3이고 상호정보는 3.7683이다. 동시발생빈도는 "access + in"의 경우가 "add + to"의 경우보다 높게 나타났으나 상호정보는 낮게 나타났다. 타동사 "use + in"의 경우에는 그 차이가 더욱 뚜렷하여 동시발생빈도는 13이나 상호정보는 1.2710에 불과하다.

〈표 3〉 동사와 전치사의 동시발생빈도와 상호정보
〈Table 3〉 co-occurrence frequency and mutual information between verb and preposition

동사	전치사	동사빈도	전치사빈도	동시발생빈도	상호정보
"access"	"in"	89	530	5	1.3768
"add"	"to"	29	186	3	3.7683
"contain"	"in"	118	530	5	0.9699
"associate"	"with"	6	94	4	7.4409
"navigate"	"through"	1	23	1	10.0569
"place"	"in"	14	530	5	4.0452
"record"	"for"	209	336	1	-1.5192
"record"	"in"	209	530	1	-2.1768
"use"	"in"	249	530	13	1.2710

이것은 타동사 "add + to"의 경우에 타동사 "add"와 전치사 "to"가 한 문장에서 함께 사용된 경우가 타동사 "add"가 독립적으로 그리고 전치사 "to"가 독립적으로 사용되는 경우보다 상대적으로 많다는 의미이다. 타동사 "access + in"의 경우에는 한 문장에서 함께 사용되는 경우가 타동사 "access"가 독립적으로 그리고 전치사 "in"이 독립적으로 사용되는 경우보다 상대적으로 적다는 의미이다. 타동사 "use + in"의 경우에는 그 차이가 더욱 뚜렷하여 타동사 "use"의 빈도는 249, 전치사 "in"의 빈도는 530인데 반하여 동시발생빈도는 13이므로 타동사 "use"와 전치사 "in"이 한 문장에서 함께 사용된 경우가 이 단어들이 독립적으로 사용된 경우보다 상대적으로 얼마나 적은지를 알 수가 있다.

"access + in", "contain + in", "use + in"의 경우에는 상호정보가 너무 낮아서 사용된 말뭉치에서는 동사와 전치사 간의 의존관계가 희박하다고 판단되어 동사 연어로 취급할 수가 없으며 "record + for", "record + in"의 경우에는 동시발생빈도와 상호정보 두 양 모두가 너무 낮아서 동사 연어로 취급할 수가 없다. "navigate + through"의

경우에는 비록 상호정보의 값은 클지라도 동시발생빈도가 너무 낮아서 동사와 전치사간의 의존관계를 동사 연어로 취급할 만큼 신뢰할 수 없다.

"associate + with", "place + in"의 경우에는 동시발생빈도 그리고 상호정보의 값이 모두 어느 수준 이상이므로 동사 연어로 파악할 수가 있다.

이러한 사실로부터 동시발생빈도 그리고 상호정보 모두가 말뭉치에서 동사와 전치사 사이의 의존관계를 파악하여 동사 연어를 추출하는데 중요한 척도가 됨을 알 수가 있다.

"IBM SQL/DS 지침서" 말뭉치에서 동사와 전치사 쌍의 표현을 동사 연어로 추출하기 위하여서는 동시발생빈도는 2이상 그리고 상호정보는 2.0이상의 값을 이용하였다. 동시발생빈도가 2미만인 표현들은 말뭉치에서 사용된 절대 빈도가 적은 표현이므로 동사 연어의 가능성을 생각할 수가 없으며 상호정보의 값이 2.0미만인 표현들도 동사 연어로 판단하기에는 동사와 전치사 사이의 의존관계가 약한 것으로 간주하였다. 다음 〈표 4〉는 말뭉치로부터 추출한 동사 연어의 예이다.

〈표 4〉 동사 연어
〈Table 4〉 verb collocation

동사	전치사	동빈	전빈	동시빈	상호정보	동사 連語
"access"	"at"	89	31	2	4.1505	"access" A "at" B
"add"	"to"	29	186	3	3.7683	"add" A "to" B
"appear"	"in"	11	530	6	4.6561	"appear" "in" A
"appear"	"on"	11	116	4	6.2630	"appear" "on" A
"append"	"to"	1	186	2	8.0413	"append" A "to" B
"assign"	"to"	12	186	6	6.0413	"assign" A "to" B
"associate"	"with"	6	94	4	7.4409	"associate" A "with" B
"begin"	"with"	4	94	2	7.0258	"begin" "with" A
"change"	"to"	76	186	17	4.8808	"change" A "to" B
"check"	"against"	28	31	3	6.4038	"check" A "against" B
"check"	"for"	28	336	2	2.3808	"check" "for" A

동빈 : 동사빈도 전빈 전치사빈도 동시빈 : 동시발생빈도

〈표 4〉의 동사 연어 열은 동사 연어의 형태이다. 대문자 A와 B는 각각 타동사의 목적어와 전치사의 목적어를 의미한다.

6. 결 론

본 논문은 말뭉치로부터 자동으로 연어를 추출하는 방법을 제시하였다. 연어는 구문분석의 결과로부터 단어들의 동시발생빈도와 상호정보를 이용하여 추출하였다. 영어에서 정의되어 있는 5가지의 연어에 새롭게 동사와 전치사로 구성된 연어를 정의하여 6가지의 연어를 추출하였다. 연어는 자연언어처리에서 언어 생성이나 기계번역에 활용될 수 있다. 연어의 추출은 말뭉치에 기반을 두기 때문에 말뭉치의 크기와 말뭉치의 분야별 구성분포에 따라 추출되는 연어는 큰 차이를 지닐 수 있다. 그러므로 말뭉치의 가장 바람직한 크기와 분야별 구성분포에 대한 연구가 필요하다.

참 고 문 헌

[1] Papoulis, A., Probability, Random Variables, and Stochastic Processes, McGraw-Hill Inc., 1991.

[2] McEliece, R., The theory of Information and Coding, Addison-Wesley Publishing Company, 1977.

[3] Church, K., "A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text," Proceedings of Second Conference on Applied Natural Language Processing, pp. 136-143, 1988.

[4] Church, K., Gale, W., Hanks, P., and Hindle, D, "Parsing, Word Associations anf Typical Predicate-Argument Relations," Proceesings of the International Workshop on Parsing Technologies, Pittsburgh, Carnegie Mellon University, pp. 389-398, 1989.

[5] Church, K., and Hanks P., "Word Association Norms, Mutual Information, and Lexicography," Journal of Computational Linguistics, Vol. 16, No. 1, pp. 22-29, 1990.

[6] Smadja, F., and McKeown K., "Automatically Extracting and Representing Collocations for Language Generation," Proceedings of the 28th Annual Meeting of the ACL, Pittsburgh, Pennsylvania, pp. 252-259, June, 1990.

[7] Aijmer, K., and Altenberg, B., English Corpus Linguistics, Longman Group UK Limited, 1991.

[8] Smadja, F., "From N-Grams to Collocations: An Evaluation of XTRACT," Proceedings of the 29th Annual Meeting of the ACL, Berkeley, California, pp. 279-284, June, 1991.

[9] Church, K., and Mercer, R., "Introduction to Special Issue on Computational Linguistics Using Large Corpora," Journal of Computational Linguistics, Vol. 19, No. 1, pp. 1-24, March 1993.

[10] Smadja, F., "Retrieving Collocations from Text: Xtract," Journal of Computational Linguistics, Vol. 19, No. 1, pp. 143-177, March 1993.

[11] Benson, J., Benson, E., and Robert, I, The BBI Combinatory Dictionary of English, John Benjamin Publishing Company, 1986.

[12] Jensen, K., "A Broad-coverage Natural Language Analysis System", Proc. of the International Workshop on Parsing Technologies, Carnegie-Mellon University, August, Pittsburgh, USA, pp. 425-439.

[13] Cruse, D. A. Lexical Semantics, Cambridge University Press, 1986.

[14] Hindle, D., and Rooth, M., "Structural ambiguity and lexical relations", Darpa Speech and Natural Language Workshop, Hidden Valley, PA, 1990.

[15] 옥철영, 한영 기계번역을 위한 구단위 변한 사전, 서울대학교 컴퓨터공학과, 공학박사학위논문, 1993.

이 호 석

1983년 서울대학교 공과대학 전자계산기공학과 졸업(공학사)
1985년 서울대학교 공과대학 컴퓨터공학과 졸업(공학석사)
1993년 서울대학교 공과대학 컴퓨터공학과 졸업(공학박사)
1985년~89년 한국전기통신공학 연구원
1994년~현재 호서대학교 컴퓨터공학과 교수
관심분야 : 자연언어처리, 객체지향 프로그래밍, 데이타베이스, 정보검색