

통계 기법을 이용한 한국어 연어 추출 모형 연구

안성민
충남대학교
smahn@cnu.ac.kr

The Study on the Model of Extracting Collocations from Corpus in Korean Using the Statistical Tools

Sung-Min Ahn
Chung-Nam University

요약

공기하여 나타나는 구 정보 중에서 연어에 대한 연구는 응용 언어학에 발전에 기여할 수 있는 부분이 크다. 연어란 어휘들 간의 제한된 결합 관계를 갖는, 공기 확률이 높은 구 구성이다. 이러한 연어 구성에 대한 연구는 특히 기계 번역이나 사전 편찬 등의 분야에서 관심이 높아지고 있다. 본 연구에서는 연어를 추출하기 위해 T-test와 상호 정보, 조건 확률 등의 여러 통계 기법의 사용을 제시한다. 각 기법을 적용하였을 때 연어 추출에 어떠한 변화를 보이는지 조사하였고, 가장 적절한 기법의 적용도 모색함으로써 향후 연어 추출의 방향을 제시하고자 한다.

주제어: 연어, T-test, 상호 정보

1. 서론

하나의 텍스트는 문장들이 모여 이루어지며, 이러한 문장들은 구가 모여 만들어진다. 마찬가지로, 구는 단어들이 모여서 형성이 된다. 단어들이 모여 구를 형성할 때에는 언어의 창조성에 의해 얼마든지 새로운 구를 형성해낼 수 있다. 이렇게 자유롭게 형성된 구와는 달리 자주 결합하는 구들이 있다. 이러한 구들은 모국어 화자들에게 하나의 단위로 인식되어 쓰이곤 하는데, 이러한 구를 연어라 할 수 있다. 연어는 이론 언어학에서 뿐만 아니라 전산 언어학에서도 많은 연구가 이루어지고 있는 분야이다. 신효필(2007)[1]에서는 연어를 어휘들 간의 제한된 결합 관계를 중시하여 연어 구성을 정의하는 것을 이론적 정의로 보았고, 빈도의 관점에서 공기 확률을 높이는 구성을 연어 구성으로 파악하는 것을 통계적 분석으로 보았다. 즉, 연어에 대한 이론적인 접근과 전산적인 접근이 다름을 보여주고 있다. 이론 언어학에서는 언어의 범주를 한정하는 데에 관심을 기울이고 있다면, 전산 언어학에서는 연어를 추출해내는 기술에 주목하고 있기 때문에 그 접근법이 다를 수밖에 없다. 그러나 연어를 추출해내는 데에는 언어의 정의가 명확히 되어야 한다. 본 논문에서는 신효필(2007)[1]이 제시한 이론적 관점과 전산적 관점을 동시에 수렴하는 언어의 정의를 제시하고, 상호 정보와 T-test, 조건 확률 등을 이용하여 연어를 보다 손쉽게 추출해 낼 수 있는 방법을 연구하고자 한다.

2. 관련연구

연어를 처음 도입한 학자는 Firth(1957)[2]이다. Firth는 “언어의 의미란 결합적 층위에서 추출된 것이지,

단어 의미에 대한 개념적 접근이나 추상적 접근과는 직 접적으로 관련된 것이 아니다”라고 말함으로써 연어라는 단어를 사용하였다. 그 이후 많은 학자들에 의해 연어에 대한 연구가 진행되었는데, 한국어에 관련한 전산 언어학적 연구를 간략히 살펴보자면 이공주 외(1995)[3]는 다양한 통계 기법을 여섯 단계에 거쳐 적용하여 연어를 추출해내는 작업을 하였고, 홍종선 외(2000)[4]에서는 인접하고 있는 어절과 논항 자리에 오는 어휘들의 언어 성에 대한 연구를 하였으며, 박경미 외(2002)[5]는 엔트로피를 이용하여 한국어 연어를 추출하였고, 임근석(2002)[6]은 분포 제약과 t-score, 상대비율 등을 반영하여 어휘적 연어를 추출하였다. 본 연구에서는 연어의 정의를 통계적인 기술에만 국한시킨 것이 아니라 이론 언어학적 측면도 최대한 도입시키고자 노력하였다. 즉, 연어란 공기 확률이 높은 구성일 뿐만 아니라 어휘들 간의 제한된 결합 관계를 갖는다는 관점 하에 연구를 진행하였다. 그리고 선행연구에서 연어를 추출하기 위해 사용된 많은 통계 기법들 중 한국어 말뭉치에서 연어를 추출해내는 데에 조금 더 나은 성능을 보여주는 기법은 무엇인지를 실험하여 향후 한국어 연어 추출의 연구 방향을 모색하고자 하였다.

3 연구 방법

3.1 코퍼스 선택

본 연구에서는 품사 태깅이 되어있는 코퍼스를 이용하였다. 사용한 코퍼스의 크기는 103,105문장으로 총 1,939,349어절이었으며, 한 문장 평균 어절 수는 약 19어절이었다[7].

3.2 Bigram 추출

연어는 서로 인접하고 있기 때문에[8], bigram을 추출하여 연구를 하였다. 이때 Bigram은 형태소 단위가 아닌 구 단위이다. 추출된 Bigram 중 단 한 번 공기한 것으로 나타난 것은 자유 결합에 의한 구 구성으로 간주하여 연구에서 배제시켰다. 빈도수가 많은 상위의 구들 중 일부는 다음과 같다.

표1 : 고빈도 공기 단어열

출현횟수	해당 공기 단어열
2148	수/NNG 있/VV+/는/ETM
1662	수/NNG 있/VV+/다/EF+./SF
951	수/NNG 없/VV+/다/EF+./SF
798	할/NNG 것/NNG+이/VV+/다/EF+./SF
750	수/NNG 없/VV+/는/ETM
720	이/NNG 같/VV+/은/ETM
647	수/NNG 있/VV+/도록/EC
646	할/NNG 수/NNG
532	있/VV+/다/EF+./SF 그러나/MAJ
528	수/NNG 있/VV+/을/ETM

3.3 상호 정보

3.2에서 제시한 표에 보면, 아홉 번째의 “있다. 그러나”는 그 공기 횟수가 532회나 됨에도 불구하고 연어로 볼 수는 없는 조합이다. 공기 횟수가 많다면 연어일 가능성은 물론 그만큼 높아지겠지만, 모두 연어가 될 수 있는 것은 아니다. 따라서 추출한 bigram에 추가적인 작업을 수행해야 하는데, 본 연구에서는 상호 정보(Pointwise Mutual Information[9])를 사용한다.

(1)

$$I(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$$

상호 정보는 이벤트 x와 y의 상호 의존성을 측정하는 통계 기법이다. 그러나, 수치는 각각의 단어 빈도수에 따라 의존적으로 나타나기 때문에 상호 정보는 데이터가 적을 경우 문제가 된다[10]. 실제 데이터에서도 이러한 결과가 그대로 나타난다.

표2 : 상호 정보 수치 및 발생 빈도

MI	$C(w^1)$	$C(w^2)$	$C(w^1, w^2)$	해당 단어열
19.89	2	2	2	외 형 / N N G + 위 주 /NNG+의/JKG 성장 /N N G + 가 도 / N N G + 를 /JKO
19.30	3	3	3	흔인/NNG 빙자/NNG
15.83	4	25	3	잔뼈/NNG+가/JKS 굽

14.20	25	33	8	/VV+은/ETM 아쉬움/NNG+이/JKS 남/VV+는다/EF+./SF
13.86	13	50	5	시일/NNG+이/JKS 결 리/VV+ㄹ/ETM

따라서, 이러한 저빈도의 구 구성이 높은 수치를 갖는 문제를 해결하기 위하여 상호 정보 값에 빈도수를 곱하였다. 그 식은 다음과 같다[11].

(2)

$$C(w^1w^2)I(w^1w^2)$$

식(2)를 적용시켜, 표2의 수치를 보완하였는데 이는 평가에 있어서 가장 최적의 기법을 찾기 위함이다. 얻어낸 값은 각각 39.78, 57.90, 47.49, 113.6, 69.3이다.

3.4 T-test

T-test는 어떤 배열의 발생이 믿을 만한지, 그렇지 않은지를 검증하는 방법이다[10]. T-test 값 역시 공기 횟수가 같다고 하여 같은 값을 갖는 것이 아니라, 표본의 신뢰도가 얼마나 있는지를 검증하는 것이다. 식은 다음과 같다.

(3)

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{S^2}{N}}}$$

\bar{x} 는 표본 평균을, μ 는 평균을 나타내며, S^2 은 표본의 분산을, 그리고 N 은 골라낸 샘플 수를 의미한다. 이렇게 얻어진 값의 일부는 다음과 같다.

표3 : T-test 값

t	$C(w^1)$	$C(w^2)$	$C(w^1, w^2)$	해당 단어열
2.866	736	1058	9	문제/NNG+가/JKS 되 /VV+ㄹ/ETM
2.889	612	1058	9	일/NNG+이/JKS 되/VV+ ㄹ/ETM
2.999	50	83	9	느낌/NNG+이/JKS 들 /VV+ ㄴ다/EF+./SF
3.000	74	21	9	불/NNG+이/JKS 나/VV+ 자/EC

3.5 조건 확률식

김진해(2000)[12]에서는 “연어 구성은 두 구성 요소 중에 어느 한 쪽이 다른 요소를 요구하기 때문에 일정한 방향성을 가진다.”라고 명시하고 있다. 따라서 조건 확률식(Conditional Probability)[9]을 통하여 선택 제약을 추

출하여 실제 데이터에 적용시켜 결과 값을 비교해보고자 한다.

(4)

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

조건 확률식은 Bigram 양쪽 모두에 적용하였고, 그 결과의 일부는 다음과 같다.

표4: 조건 확률

$CP(w^2 w^1)$	$CP(w^1 w^2)$	해당 단어열	
0.182	0.020	문/NNG+이/JKS /ETM	열리/VV+ㄹ
0.043	0.222	합의/NNG+가/JKS /EC+지/VX+었/EP+다고/EC	이루/VV+어
0.010	0.333	사건/NNG+이/JKS	터지/VV+자/EC

조건 확률식으로 얻어진 값은 그 값대로 평가에 쓰였으며, 보완을 위하여 MI*CP의 값도 추가되었다.

4. 실험 및 평가

4.1 실험 환경

본 실험은 품사가 태깅된 1,939,349어절의 코퍼스를 대상으로 하였다. 코퍼스에서 Bigram을 추출하여 공기 횟수 1을 제외한 나머지에 대해 자료에 대해 수행되었다.

표5: 결과 분석

type	연어 수	MI	T-Score	$CP(w^2 w^1)$	$CP(w^1 w^2)$	C*I	$I^* CP(w^2 w^1)$	$I^* CP(w^1 w^2)$	분포율	연어 순실 수	1%당 연어 순실 수
685	173		o						25.26	46	9.76
581	159	o							27.37	60	8.79
999	211			o					21.12	8	13.86
764	196				o				25.65	23	4.50
315	114							o	36.19	105	6.71
456	93						o		20.39	126	-843.64
502	148	o	o						29.48	71	7.94
300	112		o					o	37.33	107	6.37
295	111	o	o					o	37.63	108	6.32
572	162		o		o				28.32	57	7.33
467	144	o	o		o				30.84	75	7.29
751	183					o			24.37	36	9.42
311	113					o		o	36.33	106	6.71
300	112		o			o		o	37.33	107	6.37

전체 자료에 대해 앞서 소개한 통계 기법을 적용시켜 평가하였다.

4.2 평가 방법

코퍼스에서 공기 횟수 1을 제외한 Bigram의 type 수는 129,440이었다. 여기에 네 가지 식을 적용하여 각각의 수치를 얻어냈다. 각 수치의 공정한 평가를 위해 코퍼스의 일부를 추출하여 결과 값을 비교하였다. 평가를 위해 추출된 것은 주술관계를 가진 코퍼스로서 총 type 수는 1066개였다.

이 코퍼스는 미리 준비된 연어사전[13]을 이용하여 연어가 얼마나 포함되어 있는지를 평가하였는데, 연어는 전체 20.54%를 차지하는 219개였다.

4.3 실험 결과 및 분석

준비된 코퍼스는 각각의 식을 차례로 적용시켜 가면서 그 정확도를 측정해나갔다.

각각의 기준치는 MI <10, T-Score <1.414, CP<0.01, C*I <15, I*CP <1.0이었다. 식에서 보이는 바와 같이 각 식을 따로 적용하기도 하고 여러 개를 동시에 적용시켜보기도 하면서 각각의 결과 값을 분석했다. 여러 제약을 동시에 가해 연어 추출 수를 증가 시킬 수는 있지만 그럴 경우에 많은 연어 수를 손실하는 결과를 초래하기도 한다. 반대로 제약이 너무 느슨할 경우, 연어 이외의 구성들이 뽑혀 나올 수가 있다. 따라서 어떤 식을 적용할 때, 연어 손실 수를 최소로 줄이면서 좋은 결과 수치를 가져올 수 있는지를 분석하였다. 분석 결과 MI와 T-score, I*CP($w^1|w^2$)값을 적용했을 때 가장 좋은 값을 얻을 수 있음을 알 수 있었다. 본 데이터에서는 37.63%의 결과 값을 얻을 수 있었는데, 이는 임근석(2002)[6]에서 주술관계에서 얻어낸 연어가 30.11%에 비

교하면 7.52%의 성능향상을 가져온 것이라 할 수 있다.

5. 결론

본 연구는 태깅이 된 코퍼스에 여러 가지 통계 기법을 적용하여 자동적으로 연어를 추출하는 방법을 연구하였다. 기준에 제시된 여러 통계 기법을 사용하여 어떤 통계 기법을 적용시킬 때, 연어 추출을 극대화 할 수 있는지 살펴보았다. MI와 T-score, $I^* CP(w^1|w^2)$ 값을 적용한 수치가 연어의 손실을 줄이면서 코퍼스에서 연어가 포함된 Bigram을 확보할 수 있는 가장 좋은 방법임을 증명하였다. 그러나 주술관계에만 극한된 연구였기에 이를 다른 구성에 적용시켜 일반성을 증명하는 것이 향후 연구 과제로 남는다 하겠다.

참고문헌

- [1] 신효필, “연어의 통계적 접근을 통한 로그 우도비 중심의 연어 검증,” 한국언어학회 언어학 제 47호, pp. 107-138, 2007.
- [2] Firth, J. R., Modes of meaning, In J. R. Firth, London, Oxford University Press, Paper in Linguistics 1934-1951, 1957.
- [3] 이공주, 김재훈, 김길창, “품사 태깅된 말뭉치로부터 한국어 연어 추출,” 한국정보과학회 추계 학술발표 논문집, pp. 623-636, 1995.
- [4] 홍종선, 강범모, 최호철, “한국어 연어 정보의 분석 응용에 관한 연구,” 한국언어학회 한국어학 제11호, 2000.
- [5] 박경미, 송만석, “엔트로피를 이용한 한국어 연어 추출,” 한국정보과학회 봄 학술발표논문집, pp. 451-453, 2002.
- [6] 임근석, “현대 국어의 어휘적 연어 연구,” 서울대 석사학위 논문, 2002.
- [7] “21세기 세종계획_색인_문어_news,” 문화관광부, 국립국어원, 2007.
- [8] 박성숙, “한불 사전에서의 연어 처리.” 불어불문학 연구 34, 한국불어불문학회, pp. 571-587, 1997.
- [9] Fano, R. M., Transmission of Information: A Statistical Theory of Communications, MIT Press, 1961.
- [10] Manning, C. D. and H. Schütze, Foundations of Statistical Natural Language Processing, The MIT Press, 1999.
- [11] Fontenelle, Thierry, Walter Brüls, Luc Thomas, Tom Vanallemeersch, and Jacques Jansen, deliverable D-la:survey of collocation extraction tools, University of Liege, Liege, Belgium, 1994.
- [12] 김진해, “기능동사는 어휘적 의미가 없는가?,” 경희대 민속학연구소 한국문화연구 3, pp. 221-241, 2000.
- [13] “21세기 세종계획_전자사전,” 문화관광부, 국립국어원, 2007.