세종 말뭉치로부터 용언연어 추출

이정태○, 천민아, 김재훈
한국해양대학교
make8286@naver.com, minah014@outlook.com, jhoon@kmou.ac.kr

Verbal Collocation Extraction from Sejong Tagged Corpus

Jeong-Tae Lee○, Min-Ah Cheon, Jae-Hoon Kim
Korea Maritime and Ocean University

요 약

연어는 둘 이상의 단어로 구성된 표현으로 연어에 속하는 개개의 단어의 의미로써 연어의 의미를 유추할 수 없다. 따라서 연어의 의미를 분석하거나 번역할 경우 개개의 단어보다는 연어 그 자체를 하나의 분석 단위로 간주하는 것이 훨씬 더 효과적이다. 이를 위해 본 논문에서는 통계기법을 활용하여 세종 말뭉치로부터 용언연어의 추출 방법을 제시하고 그 성능을 평가한다. 연어 패턴과 통계 정보를 이용해서 연어를 추출한다. 평가를 위해서 연어 사전과 전문가의 주관적 평가를 동시에 수행했다.

주제어: 용언연어(verbal collocation), 연어추출, 연어사전, 주관적 평가

1. 서론

연어는 둘 이상의 단어로 구성된 표현으로 연어에 속하는 개개의 단어의 의미로써 연어의 의미를 유추할 수 없다[1]. 따라서 연어의 의미를 분석하거나 번역할 경우 개개의 단어보다는 연어 그 자체를 하나의 분석단위로 간주하는 것이 훨씬 더 효과적이다. 예를 들어, “모자를 쓰다”에서 “모자”는 “hat / cap / mother and son”으로 번역될 수 있고 “쓰다”는 “write / compose / use / bitter / wear”로 번역될 수 있다. 하지만 “모자를 쓰다”를 “wear a hat”으로 번역할 수 있도록 연어사전이 구축되면 의미의 중의성뿐 아니라 시스템의 성능 개선에 크게 도움이 될 것이다. 또 다른 예로는 “실패로 돌아가다”에서 “실패”는 “failure”이고 “돌아가다”는 “return / go back to / get back to”이며 “실패로 돌아가다”는 “turn out a failure”이므로 정확한 번역을 찾을 수 없다. 이런 문제를 완화시키기 위해서 연어의 대역사전이 있다면 커다란 도움이 될 것이다. 특히 동사는 문장에서 가장 핵심적인 구성요소이기 때문에 대역어의 선택이 매우 중요하다. 본 논문은 세종 말뭉치로부터 동사 연어(verbal collocation)를 추출하는 방법을 제안한다. 연어 후보를 추출하기 위하여 빈도, 품사정보, 용언 위치에 따른 결합 등 다양한 정보를 이용한다. 이러한 정보를 바탕으로 본 논문에서는 더 제한적인 결합 조건과 진보된 확률 정보를 통하여 연어를 추출한다. 용언을 기준으로 추출한 연어 후보들은 매우 높은 확률로 연어라고 판단할 수 있어 앞으로 연어 사전 구축 연구에 도움이 될 것이라 생각된다.

논문의 구성을 다음과 같다. 2장에서 용언연어 추출 방법을 기술하고 3장에서 실험 및 평가를 기술하고 끝으로 4장에서 결론 및 향후 연구에 대해 기술한다.

2. 용언연어 추출 방법

본 논문에서 제안하는 방법은 용언연어 추출 방법을 단순화하기 위하여 여러 개의 단계로 나뉘어 진행되며 각 단계는 이하의 절에서 자세히 설명할 것이다.

2.1 전처리 단계

전처리 단계는 세종 말뭉치의 품사 부착을 부분적으로 수정한다. 예를 들면 세종 말뭉치의 품사 부착이 “공부/xr + 하/xsv”라며 “공부하/vv”로 수정하여 의미적으로 완전한 동사를 쉽게 찾을 수 있도록 변환한다. 아래는 전처리 단계에서 변환되는 규칙들이다.

1. 접두사(xpn) + 명사 파생 접미(xsn) -> 일반 명사
2. 접두사(xpn) + 어근(xr) -> 일반 명사
3. 어근(xr) + 명사 파생 접미 -> 일반 명사
4. 어근 + 동사 파생 접미(xsv) -> 동사
5. 어근 + 형용사 파생 접미(xsa) -> 형용사
6. 일반 명사(nng) + 명사 파생 접미 -> 일반 명사
7. 일반 명사 + 동사 파생 접미 -> 일반 동사
8. 일반 명사 + 형용사 파생 접미 -> 형용사
9. 부사(mag) + 동사 파생 접미 -> 일반 동사
10. 부사 + 형용사 파생 접미 -> 형용사

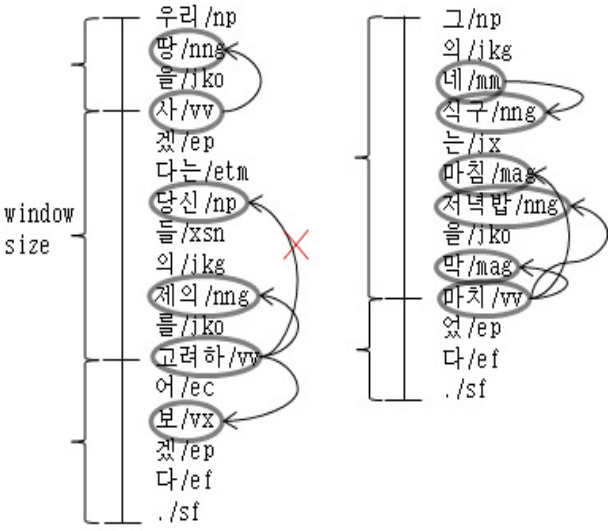
부가적으로 이러한 변환은 같은 의미를 지님에도 불구하고 다르게 해석되어 통계 정보가 손상되는 것을 막기 위함이다. 예를 들어, “불/xpn+가능/xr”같은 경우 "도전하다"와 같은 동사와 잘 어울린다. 하지만 따로 “불/xpn+가능/xr”을 그대로 두면 "가능" + "도전하다"와 같은 연어 후보가 수집될 수 있다. 이러한 경우를 막기 위해 전처리 단계를 수행한다.

2.2 용언연어 후보 추출

품사정보를 활용하여 연어를 추출하는 방법은 매우 다양하며 품사 간의 특정한 결합구조를 보인다[5,6]. 본 논문에서 사용된 용언연어 연어 유형은 다음과 같다.

1. 부사(mag) + 용언(va/vv) : 발칵/mag + 뒤집히/vv
2. 관형사(mm) + 체언(n+) : 오랜/mm + 세월/nng
3. 체언(n+) + 용언(va/vv) : 다정/nng + 다감하/va
4. 체언(n+) + 체언(n+) : 시장/nng + 경제/nng
5. 용언(va/vv) +보조용언(vx) : 도리/vv + 내/vx

위 다섯 가지 유형의 용언연어 후보의 통계 정보를 통해 연어 후보를 추출할 것이다.



[그림 1] 용언연어 후보 추출을 위한 탐색범위

[그림 1]은 유형 1~5가 어떻게 결합되어 추출되는지를 보여준다. 후보 추출 범위(window)는 주용언(vv/va)에 의해 결정되며 유형 1, 3, 5번의 경우에 그 범위에 있는 주용언으로부터 가장 가까운 단어 하나만 선택한다. [그림 1]의 왼쪽 그림을 보면 탐색범위 내에 용언과 결합 가능한 체언이 "당신/np"과 "제의/nng" 두 종류가 있지만 "제의/nng"와만 결합하도록 하였다. 이는 문장의 목적어 술어 관계는 하나 이상 나타나지 않는 문법적 특성을 이용한 것이다. 5번 또한 "내버려두다"와 같이 "내버리다" + "두다"와 같은 표현은 한 품사가 여러 품사에 걸쳐 공기관계를 갖지 않는다. 그에 반해서 부사와 주용언의 결합은 오른쪽 그림과 같이 여러 개와 결합하여도 그 의존성이 크게 변하지 않는다. 이러한 언어의 특징을 고려해 부사 용언 결합은 탐색범위 내에서는 여러 차례 나타날 수 있다고 생각하고 연어후보 추출을 시행한다. 2, 4번 경우에는 거리가 1이내인 결합만을 허용한다. 여기서 거리는 두 단어 사이에 존재하는 형태소의 개수를 말한다. 즉, 바로 연결되는 두 단어만을 추출하였다.

2.3 통계 정보 측정 방법

2.1과 2.2를 거쳐 뽑은 연어후보를 이용하여 연어 후보들의 순서를 정한다. 말뭉치의 통계 분석에서 다루는 자료는 대부분 질적인 자료(qualitative data)이며, 이러한 데이터는 주로 분할표(contingency table)를 활용하여 분석된다. 여기서 살펴볼 자료의 변수는 왼쪽 단어와 오른쪽 단어 두 개의 변수를 가지고 있는 2차원 분할표를 통해 분석한다. 본 논문에서는 χ^2검정(chi-square, 식 (1))을 통한 분석과 PMI(Pointwise Mutual Information, 식 (2))를 통한 분석을 이용할 것이다.

$$\chi^2 = \sum_{j=1}^{M}\sum_{j=1}^{N}\frac{(O_{ij}-E_{ij})}{E_{ij}} \quad (1)$$

$$PMI(x,y) = \log\frac{p(x,y)}{p(x)p(y)} \quad (2)$$

식 (1)은 일반적으로 편포(skewed distribution)을 이용한 검정 방법으로 말뭉치 자료가 정규 분포를 따르지 않는다고 생각할 때 분석에 유리한 장점이 있으며 실험에서도 더 좋은 성능을 보인다[2]. 식 (2)는 두 단어가 어떤 개연성을 가지고 출현할 확률과 두 단어가 아무런 관계도 없이 우연히 함께 출현할 확률의 비를 보여 이 비율이 높으면 연어일 가능성이 높다는 것이다.

[표 1] 연어추출을 위한 2차원 분할표

	w1=우리	w1!=우리
w2=나라	o_{11}=675 freq(w1,w2)	o_{12}=1816-675 freq(w2) -freq(w1,w2)
w2!=나라	o_{21}=8858-675 freq(w1) -freq(w1,w2)	o_{22}=N- (o_{11}+o_{12}+o_{21})

[표 1]은 세종 말뭉치의 일부를 가지고 '우리/np', '나라/nng'를 2x2 분할표로 나타낸 것이다. 이를 통해 식 (1)과 (2)를 나타내면 아래와 같다.

$$\chi^2 = \frac{N(O_{11}O_{22}-O_{12}O_{21})^2}{(O_{11}+O_{12})(O_{11}+O_{21})(O_{12}+O_{22}(O_{21}+O_{22})} \quad (3)$$

2x2 분할표에서는 자유도가 1이며 이런 경우는 이산 통계량에 가깝기 때문에 Yates 보정을 수행한다. 이 방법을 사용하면 χ^2분포에 가까운 통계량을 얻을 수 있다[7]. 이 방법을 통해 수식을 수정하면 (4)와 같다. [표 1]을 활용하여 PMI를 계산하는 수식은 (5)와 같다.

$$Yates = \frac{N(|O_{11}O_{22}-O_{12}O_{21}|-N/2)^2}{(O_{11}+O_{12})(O_{11}+O_{21})(O_{12}+O_{22}(O_{21}+O_{22})} \quad (4)$$

$$PMI(x,y) = \log(\frac{O_{11}N}{(O_{11}+O_{12})(O_{11}+O_{22})}) \quad (5)$$

통계 정보를 추출할 때에 일반적으로 o_{11}의 빈도가 5이하이면 실험에서 제외시킬 것을 권장하고 있다. 연어의 특성상 습관적으로 같이 자주 나타나야 하는데 그렇지 않고 서로의 의존관계가 높아 상위에 랭크되는 단어들을

제외하기 위함이다. 본 실험에서는 권장되는 빈도 이하의 연어후보에 대해서는 전혀 고려하지 않고 실험을 진행한다. 충분히 크지 않은 말뭉치는 도메인에 따라 특정 단어들을 뽑아내곤 하는데 효과적인 방법의 도입을 통해 특정 단어를 추출하는데 도움을 줄 수 있다[8]. 그 수식은 (6)과 같다. 이 방법은 χ^2검증에 적용 가능하며 이를 적용하면 세종 말뭉치에서 좀 더 좋은 결과를 기대할 수 있다.

$$Yates = sign(O_{11}O_{22} - O_{12}O_{21})^* Yates_0$$
$$sign(z) = \begin{cases} +1 & \text{if } z > 0 \\ -1 & otherwise \end{cases} \quad (6)$$

3. 실험 및 결과

실험은 세종 말뭉치의 45만여 개의 태그된 문장을 대상으로 수행하였다. 평가는 사람이 수동으로 연어임을 판단하는 방법과 기존에 존재하는 한국어 연어 목록[9]을 이용하여 평가를 실시하였다. 한국어 연어 목록은 어절 단위로 사전이 구축되어 있기 때문에 어절을 분석하여 조사나 어미를 제외한 후 연어 후보와 일치하면 정답으로 표기하였다. 제안한 연어 추출 방법으로부터 평가한 결과는 [표 2]와 같다.

표 2

[표 2] 용언연어 추출에 대한 성능 평가

언어 유형	Chi-Square		PMI	
	사전	수동	사전	수동
부사+용언	33%	97%	14%	97%
관형사+체언	22%	70%	13%	69%
체언+용언	20%	82%	3%	70%
체언+체언	1%	10%	0%	12%
용언+보조용언	33%	97%	16%	95%

[표 2]는 각각의 유형에 따라 PMI와 Chi2값의 상위 100개의 후보들에 대해 정확도를 나타낸 결과표이다. 각 언어 유형별로 적게는 수천, 많게는 수만 개의 후보들이 추출되기 때문에 상위의 후보들만 가지고 수동으로 평가하였으며 Chi-Square와 PMI, 두 가지 평가 방법에서 특정 수치 이내에서는 상당히 신뢰도가 있는 추출률을 보여준다. 두 평가 방법을 비교해 보았을 때엔 Chi-Square 방식이 PMI보다 더 뛰어난 결과를 보여준다. 사전으로 엄격하게 평가하였을 때나 사람이 수동으로 채점하였을 때나 두 가지 모든 경우에서 평균적으로 더 좋은 결과를 보였다. “부사+용언”과 “용언+보조용언”은 사전으로 평가하던 수동으로 하던 좋은 결과를 보이고 있어 상당히 유용할 것으로 생각된다. “체언+체언”의 결과는 상위 100개에 랭크된 단어들이 대부분 고유 명사이기 때문에 사전에 없는 것과 더불어 연어라고 판단하기 힘들어 좋은 결과가 나오지 않았다. 하지만 상위에 랭크된 고유명사들을 제외하면 많은 합성어들이 랭크되는 것을 볼 수 있다.

4. 결론

본 논문에서는 통계적 기법을 통하여 자동으로 연어를 추출하는 방법에 대해서 실험하였다. 현재 연어는 명확한 정의가 없기 때문에 사람이 평가하여 Kappa 계수를 이용하거나 평균 등의 값으로 평가하는 경우가 많았다. 본 실험에서는 객관성을 위하여 사전으로 평가한 후 수동으로 평가하여 그 차이가 어느 정도인지를 보였다. 본 실험의 결과로는 현존하는 연어사전의 어휘가 부족한 것으로 보이며 이 후 완성도 있는 사전이 만들어져야 객관성 있는 평가가 가능 할 것으로 보인다. 해당 실험에서 수동으로 평가한 결과가 좋기 때문에 앞으로 연어 사전을 만들 때 큰 도움이 될 것이라고 생각한다. 차후 연어의 길이를 늘여 찾고자 한다면 n개의 품사 조합이나 연어의 이행적(transitive) 관계를 이용하여 확장할 수 있을 것이다.

감사의 글

본 연구는 미래창조과학부 및 정보통신기술진흥센터의 정보통신·방송 연구개발사업의 일환으로 수행하였음. [10041807, 지식학습 기반의 다국어 확장이 용이한 관광/국제행사 통역률 90%급 자동 통번역 소프트웨어 원천 기술 개발]

참고문헌

[1] H. E. Palmer, "영어학사전", 1990.
[2] 주은석, "대규모 연어추출을 위한 통계적 기법 비교", 언어사실과 관점, 제25권, pp. 189-210, 2010.
[3] 이공주, 김재훈, 김길창, "품사 태깅된 말뭉치로부터 한국어 연어 추출", 한국 정보과학회 추계 학술발표 논문집, pp.623-636, 1995
[4] F. Samadja, "Retrieving collocations from text: Xtract", In Computatuinal Linguistics, 19(1), pp.143-177, 1993.
[5] 서상규, 홍종선, "한국어 정보 처리와 연어 정보", 국어학회, 제 39집, pp. 321-360, 2006.
[6] 임근석, "통계적 방법을 이용한 문법적 연어 후보 추출", 한국어학회, 제45권, pp. 305-333, 2009
[7] F. Yates, "Contingency Tables Involving Small Numbers and the χ^2 Test", Journal of the Royal Statistical Society, Vol. 1, No. 2, pp. 217-235, 1934
[8] Kiyomi chujo, Masao Utiyama, Takahiro Nakamura, Kathryn Oghigian. "Evaluating Satistically-extr-
cted Domain-Specific Word Lists", Nihon University, 2010
[9] 김하수 외 8명, "한국어 교육을 위한 한국어 연어목록", 2007