엔트로피를 이용한 한국어 연어 추출

박경미[0] 송만석
연세대학교 컴퓨터과학.산업시스템공학과
{kmpark[0], mssong}@december.yonsei.ac.kr

Extracting Collocations Using Entropy in Korean

Kyung-Mi Park[0] Man-Suk Song
Dept. of Computer Science & Industrial Systems Engineering, Yonsei University

요 약

연어는 습관적으로 같이 자주 나타나는 단어열로 각 단어로 분리하기보다 통합해 처리하는 것이 효율적이기 때문에 기계 번역과 음성 인식등에서 유용한 정보로 사용된다. 이러한 연어를 추출하기 위해 본 논문에서는 2가지 경우를 고려했는데, 첫 번째로 연어를 말뭉치에 자주 나타나는 단어열이라고 했을 때 단어열들의 엔트로피가 일정값 이상이면 연어로 추출했다. 두 번째로 통사적 제약이 있는 연어를 추출하기 위해 앞 또는 뒤에 올 단어를 제약하는 단어의 엔트로피를 구해 일정값 미만이면 그 단어를 포함한 단어열을 연어로 추출했다. 실험은 품사 부착된 HANTEC 말뭉치를 가지고 수행했고, 첫 번째 방법으로 실험했을 때 엔트로피가 2이상인 단어열을 가지고 분리된 연어도 유도해냈다.

1. 서론

연어(collocation)는 둘 이상의 단어가 연결되어 있을 경우 이들을 낱낱의 단어로 분리시키지 않고 통합된 단위로 파악해야 할 어군(word group)을 말한다[1]. 또, 자주 나타나면서 한 단어 이상으로 구성돼 있고 그 순서가 고정적이라는 특징을 갖는다. 주로 대용량의 말뭉치로부터 상호 정보(mutual information)와 같은 확률 정보를 알아내 자동으로 추출하는데[2,3], 일정값보다 크거나 혹은 작아야 한다는 조건을 만족하지 않을 경우 연어가 아닌 것으로 간주돼 제거된다. 이러한 연어는 기계 번역과 음성 인식등에서 유용한 정보로 사용되는데, 특히, 분야를 한정했을 경우 그 분야에서만 통하는 전문어(jargon)등을 알아낼 수 있어 효율적이다. 예를 들어, "법률" 분야의 기계 번역 시스템을 만든다고 할 때 "재판부는 판결문에서", "형의로 구속 영장을 신청했다"와 같은 고빈도 단어열들의 대역어를 미리 구축해 놓음으로써 번역의 질을 높일 수 있다.

본 논문에서는 다음과 같은 엔트로피 개념을 이용해 연어를 추출한다. 즉, 확률 변수 X가 x_1, x_2, x_3, ...등 여러 가지 값을 갖고 각각이 균일한(uniform) 분포를 가질수록 불확실성이 커져 엔트로피는 증가하는데 반해, 확률 변수가 적은 경우의 수를 갖고 한두가지만이 집중적으로 나타날수록 불확실성이 작아져 엔트로피는 감소한다는 것이다.

이 개념을 이용해 두가지 경우로 나눠 연어를 추출하는데, 첫 번째로 2.1절에서는 연어를 말뭉치에서 자주 나타나는 단어열이라고 했을 때 고빈도 단어열들 중 연어를 추출하는 방법이고 두 번째로 2.2절에서는 통사적 제약이 있는 연어를 추출하는 방법을 기술했다. 그리고 2.1절의 방법으로 실험했을 때 엔트로피가 2이상인 단어열을 가지고 분리된 연어(interrupted collocation)도 유도해냈다. 품사 부착된 말뭉치를 사용했기 때문에 본 논문에서 단어는 형태소를 의미하고 단어열의 길이는 단어수를 뜻한다.

2. 연어 추출 방법

2.1 고빈도 단어열들 중 연어 추출

말뭉치에서 자주 나타나는 연어를 추출하기 위해 고빈도 단어열들 중 연어가 아닌 것들을 제거하는 방법이 필요하다. 우선, 고빈도 단어열과 그것의 부분 단어열의 특징을 알아보기 위해 다음과 같이 말뭉치에서 49번 나타난 단어열을 고려하자.

"일/NNDE2 이/NNIN2 같은/PPCA1 내용/NNIN2 을/PPCA2 골자/NNIN2 로/PPAD 하/VBMA 는/ENTR1"

표1 : 고빈도 단어열의 부분 단어열

str	F(str)	pos	adj	max	word
일/ 이/ 같은/	237	뒤	22	211	내용
이/ 같은/ 내용/	299	뒤	6	165	을
같은/ 내용/ 을/	377	앞	3	209	이
내용/ 을/ 골자/	257	뒤	2	253	로
을/ 골자/ 로/	432	뒤	20	410	하
골자/ 로/ 하/	465	앞	13	410	을
로/ 하/ 는/	712	앞	100	261	골자

여기서 각각, str은 단어열을, F(str)은 단어열의 빈도수를, pos는 앞뒤 중 인접 단어(adjacent word)의 가지수가 적은 쪽을, adj는 pos에서 인접 단어의 가지수를, max는 pos에서 가장 많이 나타난 인접 단어의 빈도수를, word는 pos에서 가장 많이 나타난 인접 단어를 의미한다. 표1에서 예를 들어 "일/ 이/ 같은"이란 단어열이 말뭉치에서 237번 나타났는데, 이 단어열 다음에 나타난 22가지 단어 중 "내용"이란 단어가 211번이나 집중적으로 나타난다는 것을 알 수 있다.

즉, 고빈도 단어열의 앞뒤 인접 단어로 여러 가지 단어가 올 때 그 단어열은 연어일 가능성이 크지만, 앞뒤 한 쪽이라도 인접 단어로 한두가지만이 집중적으로 나타나면 연어의 부분 단어열일 가능성이 크다는 것이다

[4,5]. 이 특징을 이용해 인접 단어의 확률값을 가지고 단어열의 엔트로피를 구하면, 연어인 단어열들은 앞뒤 모두에서 엔트로피가 클 것이고, 연어의 부분 단어열들은 집중적인 분포로 인해 둘 중 어느 한쪽에서는 엔트로피값이 작아 제거될 수 있다.

이렇게 연어의 부분 단어열을 제거하기 위해 인접 단어의 확률 $p(w_i)$를 다음과 같이 정의한다.

$$p(w_i) = \frac{F(w_i)}{F(str)} \quad (1)$$

여기서 str은 단어열을, w_i는 인접 단어를 의미하고, F(str)은 단어열의 빈도수이며, $F(w_i)$는 str의 인접 단어로 w_i가 나타난 빈도수이다. 위의 확률값을 이용해 단어열 str의 엔트로피를 구하게 되는데, 그 식은 다음과 같고 앞뒤에서 구한 2가지 값 중 작은값을 H(str)로 하게 된다. 예를 들어, str의 인접 단어로 20가지가 오고 모두 균일한 분포를 갖는다고 하면 H(str)=-20*(1/20)*log (1/20)= 2.9957이다.

$$H(str) = \sum_{i=1}^{n} -p(w_i)\log p(w_i) \quad (2)$$

이제, 엔트로피에 관한 다음과 같은 조건을 제시해 이 조건을 만족하는 단어열만을 연어로 추출하게 된다. 즉, 엔트로피가 일정값 이상이어야만 연어로 추출된다.

$$H(str) \geq k_1 \quad (3)$$

부록1에 추출된 연어의 예를 보였다.

2.2 통사적 제약이 있는 연어 추출

통사적 제약이 있는 연어는 연어 내에 특정 단어가 앞 또는 뒤에 올 단어를 제약하는데, 일반적으로 말뭉치에 자주 나타나지 않는 단어열인 경우가 많다. 그래서 2.1절의 방법에 의하면 추출되지 않을 수 있다. 따라서 이와 같은 연어를 추출하기 위해 앞 또는 뒤에 올 단어를 제약하는 단어의 엔트로피를 구한다. 예를 들어, "간담"이란 단어 다음에 올 수 있는 단어들을 확률 변수로 놓고 엔트로피를 구했을 때, 특정 단어의 집중적인 분포로 불확실성이 작아져 엔트로피가 작은값을 가지면 "간담"과 특정 단어를 포함한 단어열을 연어로 추출한다.

이러한 제약 관계에 있는 단어열들을 살펴보면 "침을 뱉다"처럼 용언이 앞에 올 명사를 제약하는 용언으로 인한 선택 제약이 있고 "간담이 서늘하다"처럼 명사가 뒤에 올 용언을 제약하는 명사에 의한 선택 제약, 또 "찢어지게 가난하다"처럼 부사에 의한 것이 있다[1]. 따라서 말뭉치에서 길이 3 또는 2인 단어열들 중 '명사+조사+용언' 또는 '부사+용언'으로 이뤄진 것들을 추출해 그 중 연어를 찾는다.

예를 들어 명사가 뒤에 올 용언을 제약하는 명사로 인한 선택 제약이 있는 연어들을 추출한다고 하면, 특정 명사 다음에 올 수 있는 용언의 확률을 다음과 같이 정의한다.

$$p(verb_i) = \frac{F(verb_i)}{F(noun)} \quad (4)$$

여기서 F(noun)은 특정 명사의 빈도수이고 $F(verb_i)$는 특정 명사 다음에 온 용언의 빈도수이다. 이 때, F(noun)=1 또는 2이면 그 명사는 고려하지 않는다. 이제 이 확률값을 엔트로피에 관한 식 (2)에 대입하면 특정 명사의 엔트로피 H(noun)을 구할 수 있는데, 이 때 다음과 같은 조건을 만족하면 길이 3인 단어열 "noun+조사+$verb_j$"를 연어로 추출한다.

$$H(noun) < k_2 \quad (5)$$

여기서 $verb_j$는 특정 명사 다음에 가장 많아 나타난 용언을 의미한다. 부록2에 k_2=0.56일 때 추출된 연어의 예를 보였다.

2.3 분리된 연어 추출

분리된 연어는 연속된 연어(uninterrupted collocation)와 달리 중간에 다른 단어열의 삽입이 가능한 단어열이다. 이 절에서는 이러한 분리된 연어를 유도해내기 위해 먼저 2.1절의 방법으로 실험했을 때 연어로 추출되지 않은 엔트로피가 k_1보다 작고 2이상인 단어열들을 str_k로 놓고 말뭉치에서 str_k를 포함하는 문장들을 추출한다. 그리고, 문장내에서 공기하는 단어열 str_i가 2.1절의 방법으로 실험했을 때 엔트로피가 1이상이고 다음의 조건을 만족하면 str_k와 str_i로 이뤄진 분리된 연어를 추출한다.

$$F(str_k, str_i) > \frac{F(str_i)}{10} \quad (6)$$

이 조건은 단어열 str_k와 str_i가 같이 나타난 빈도수 $F(str_k, str_i)$가 str_i의 빈도수 $F(str_i)$의 10분의 1보다는 커야 한다는 것이다. 마찬가지로 $F(str_k, str_i)$가 str_k의 빈도수 $F(str_k)$의 10분의 1보다는 커야 한다는 조건도 만족해야 한다. 또 $F(str_i)$>=10이어야 한다. 부록3에 추출된 연어의 예를 보였다.

3. 실험

3.1 실험 말뭉치

실험은 품사 부착된 HANTEC 말뭉치를 가지고 수행했는데, 특정 분야에 속하는 말뭉치는 아니다. 실험 말뭉치의 크기는 문장수 541,668개, 어절수 7,281,736개, 형태소수 14,335,421개이다. 부착된 품사는 [6]에 소개돼 있다.

3.2 실험 방법 및 결과

2.1절에서 단어열 str이 문장의 시작이나 끝에 위치하는 횟수가 F(str)의 과반수 이상일 때, 그 단어열은 시작이나 끝에 위치하는 것으로 간주했고 인접 단어가 없기 때문에 나머지 한쪽에서 구한 엔트로피값을 H(str)로 했다. 그런데 단어열이 제목이나 완전한 문장처럼 앞뒤 모두에서 그러하면 그 단어열은 제거했다.

2.1절의 방법에 의해 추출된 연어의 개수와 정확률이 표2에 제시돼 있다. 여기서 n은 단어열의 길이이고 각 길이마다 k_1값을 다르게 줬다.

표2 : 추출된 연어의 개수와 정확률(2.1절)

n	k_1	추출된 개수	정확률
3	3.68	1056	80.1%
4	3.40	574	90.6%
5	3.21	387	88.6%

2.2절의 방법에 의한 결과가 표3에 나타나 있는데, 3가지 추출 방법 모두 k_2=0.56이다.

표3 : 추출된 연어의 개수와 정확률(2.2절)

추출방법	n	추출된 개수	정확률
용언으로 인한 선택 제약	3	118	90.7%
명사로 인한 선택 제약	3	706	86.7%
부사로 인한 선택 제약	2	49	93.9%

3.3 평가 및 분석

2.1절에 대한 표2에서 단어열의 길이가 증가할수록 고빈도 단어열의 수가 현격하게 줄어 k_1값을 각 길이마다 다르게 줬다. 또, 표에서 길이 5까지의 결과만 나타냈는데 긴 단어열도 엔트로피를 이용해 추출할 수 있다. 2.2절에 대한 표3에서 명사로 인한 선택 제약의 경우 "(여시니아균, 이, 검출되)"처럼 우연히 말뭉치에 나타난 단어열들에 의해서 추출된 개수는 많았지만 정확률이 낮아졌다. 2.3절의 방법으로 실험했을 때 엔트로피가 2이상인 80개의 str_k중 25%만이 분리된 연어를 구성했다.

4. 결론

본 논문에서는 엔트로피 개념을 이용해 2가지 경우로 나눠 연어를 추출했다. 2.1절의 방법에 의해서는 말뭉치에서 자주 나타나는 연어들을 추출했고, 2.2절의 방법에 의해서는 통사적 제약이 있는 연어들을 추출했다. 또, 2.3절에서 분리된 연어도 유도해냈는데 날씨, 호텔 예약, 컴퓨터 매뉴얼등 특정 분야에 속하는 말뭉치를 이용했다면 좀 더 좋은 결과를 얻었을 것이다. 이렇게 추출된 연어들은 기계 번역등 다양한 분야에 응용되어 그 질을 높일 수 있다.

참고 문헌

[1] 이희자, "현대 국어 관용구의 결합 관계 고찰", 한글 및 한국어 정보처리 학술발표 논문집, pp.333-352, 1994.

[2] 이공주, 김재훈, 김길창, "품사 태깅된 말뭉치로부터 한국어 연어 추출", 한국 정보과학회 추계 학술발표 논문집, pp.623-636, 1995.

[3] Christopher D. Manning and Hinrich Schutze, "Foundations of Statistical Natural Language Processing", MIT press, pp.151-227, 1999.

[4] Sayori Shimohata, Toshiyuki Sugio and Junji Nagata, "Retrieving Collocations by Co-occurrences and Word Order Constraint", In the 35th Annual Meeting of ACL, pp.476-481, 1997

[5] Makoto Nagao and Shinsuke Mori, "A New Method of N-gram Statistics for Large Number of n and Automatic Extraction of Words and Phrases from Large Text Data of Japanese", In the 15th COLING, pp.611-615, 1994.

[6] 윤준태, "공기 관계 기반 어휘 연관도를 이용한 한국어 구문 분석", 연세대학교 박사학위 논문, pp.97-99, 1998.

부록1 : 추출된 연어의 예(2.1절)

여기서 n=3일 때 k_1=3.68, n=4일 때 k_1=3.40, n=5일 때 k_1=3.21이다.

n	str	F(str)	H(str)
3	에/PPAD 대하/VBMA ㄴ/ENTR1	20,596	6.93
	전국/NNIN2 경제인/NNIN2 연합회/NNIN2	399	4.03
4	르/ENTR1 수/NNDE1 있/AJMA 다/ENTE	2,806	4.81
	재판부/NNIN2 는/PPAU 판결문/NNIN2 에서/PPAD	213	4.51
5	가/PPCA1 자지하/VBMA 는/ENTR1 비중/NNIN2 은/PPAU	38	3.28
	을/PPCA2 상대/NNIN2 로/PPAD 내/VBMA ㄴ/ENTR1	94	3.58

부록2 : 추출된 연어의 예(2.2절)

여기서 F(str)의 오른쪽에 있는 숫자는 제약을 가하는 단어의 빈도수이다. 예를 들어, 명사로 인한 제약에서는 "대미", "생색", "골탕"등의 빈도수를 의미한다.

추출방법	str	F(str)	H(str)
용언으로 인한 제약	(죽, 을, 쑤)	5/5	0.00
	(목, 을, 조르)	7/8	0.38
	(눈, 을, 흘기)	5/5	0.00
명사로 인한 제약	(대미, 를, 장식하)	6/6	0.00
	(생색, 을, 내)	6/7	0.41
	(골탕, 을 먹이)	7/8	0.38
부사로 인한 제약	(골똘히, 생각하)	3/3	0.00
	(텅, 비)	7/7	0.00
	(물씬, 풍기)	12/14	0.51

부록3 : 추출된 연어의 예(2.3절)

여기서 각 단어열 str_i의 앞쪽에 있는 빈도수는 F(str_k, str_i)를 의미하고, str_i중 다른 str_i에 포함되는 경우는 제외시켰다. 즉, 단어열의 길이가 길수록 의미있다고 여겼다. 여기서 예를 들어 "에/ 대하/ ㄴ/ 일제/ ..... 단속/ 을/ 벌이/ 어/"와 같은 분리된 연어를 유도해낼 수 있고 두 단어열 사이에는 "강제", "집중", "심야"등이 들어갈 수 있다.

		str	F(str)	H(str)
str_k		단속/ 을/ 벌이/ 어/	23	2.13
str_i	3	개/ 업소/ 를/	12	1.42
	4	명/ 을/ 구속하/ 고/	15	2.49
	3	명/ 을/ 적발/ ,/	23	1.93
	3	범칙금/ 을/	13	2.03
	6	에/ 대하/ ㄴ/ 일제/	48	1.75
	3	위반/ 사범/	24	2.10
	3	적발하/ 었/ 다고/	22	1.43
	3	측심/ 에/	15	1.23