한국어 텍스트 내 용어연관성 분석을 위한 기초 연구

Preliminary Study on the Analysis of Term Associations in Korean Text

정영미, 이재윤, 연세대학교 문헌정보학과

Young-Mee Chung, Jae-Yun Lee

Dept. of Library and Information Science, Yonsei University

텍스트 자동분석을 통해 얻어진 통계적인 용어연관성은 정보검색 및 언어처리와 관련된 여러 분야에서 폭넓게 이용되고 있다. 용어연관성을 구하기 위한 연관계수는 여러 가지가 있지만 적용분야에 관계없이 유사계수 공식이나 상호정보량 공식이 주류를 차지하고 있다. 이런 공식들은 그 통계적 특성이 서로 다르기 때문에 알맞은 적용분야를 파악할 필요가 있다. 이 연구에서는 주요 연관계수 공식의 특성을 이론적으로 파악하였고, 실험으로 검증하기 위하여 240만 어절 분량의 실험용 한국어 신문기사 데이터베이스를 구축하였다.

1. 서론

용어간의 관계는 문헌정보학, 언어학, 심리학 등 여러 분야의 관심 주제가 되고 있으며 다양한 분야에 응용되고 있다. 컴퓨터에 의한 자연언어 처리 기술이 발전하면서 단어/용어간의 관계를 언어학적 이론이나 직관적 분석에 의존하지 않고 텍스트에 대한 통계적인 분석을 통해 자동으로 파악하려는 연구가 활발해지고 있다. 이런 연구를 통계적 용어연관성(statistical term associations)에 관한 연구라고 하며, 연구의 대상이 되는 단어/용어의 동시출현을 공기(共起; co-occurrence), 동시출현 횟수를 공기빈도(共起頻度; co-occurrence frequency)라고 부른다.

용어연관성을 측정하기 위해 공기빈도를 이용하는 경우 특정한 공식(formula) 또는 계수(measure)를 사용하게 된다. 이때 사용하는 공식을 연관계수(association measure), 또는 유사계수(similarity measure)라고 흔히 부른다. 여기서는 비교하는 두 항이 반드시 유사한 가를 평가하는 것이 아니라 뭔지는 모르지만 '어떤' 종류의 연관성이 어느 정도 있는지 판단하기 위한 것이므로 연관계수라고 부르기로 한다.

이 연구에서는 한국어 텍스트를 대상으로 용어연관성을 분석할 때 고려해야 할 여러 가지 연관계수에 대해 그 종류, 정의, 성질을 이론적으로 알아보고, 실험분석 및 검증을 위한 한국어 텍스트 데이터베이스를 구축하였다. 각 연관계수를 적용하는 실제 실험은 후속연구에서 이루어질 것이다.

2. 용어연관성 분석시 고려 사항

연관성을 측정하기 위해서는 다음과 같은 사항을 먼저 결정해야 한다.

1. 분석단위 : 빈도 측정 대상을 결정해야 한다. 응용분야에 따라서는 자모나 음절을 연관성 분석 대상으로 해야 할 때도 있다.

② 공기범위 : 두 분석단위가 공기했다고 판정할 범위를 결정해야 한다. 연어 분석과 같은 경우에는 주로 인접 단어를 범위로 하며 다른 경우에는 3 단어 이내, 혹은 5 단어 이내와 같은 일정한 크기의 문맥창(context window)을 범위로 삼거나 동일 문장, 동일 문헌 등을 이용하게 된다.

③ 공기순서 : 공기범위가 한 문장이거나 몇 개 단어 이내인 경우 분석단위의 출현 순서를 구분할 지 여부를 결정해야 한다.

④ 공기거리 : 공기범위 내에서 두 분석단위 사이의 거리를 고려할지 여부를 결정해야 한다.

⑤ 연관계수 : 각 연관계수의 특성을 고려하여 적절한 연관계수를 결정해야 한다.

3. 용어 연관계수

두 용어의 연관성 측정에 연관계수를 적용하기 위해서는 각 용어의 출현빈도와 공기빈도가 있어야 한다. 이때 두 용어의 단순 공기빈도를 그대로 두 용어 사이의 연관계수로 이용하면 빈도가 높은 용어일수록 다른 용어와의 공기빈도가 높게 나타나기 마련이므로 여러 쌍의 연관성 값을 상대적으로 비교하기가 어렵다.

일반적으로는 단순 공기빈도를 이용하는 방식 대신 개별 용어의 빈도와 전체 용어 빈도를 함께 이용하여 용어 사이의 통계적인 연관성을 객관적으로 평가하는 상대 공기빈도 방식의 연관계수가 주로 이용된다. 상대 공기빈도 방식의 연관계수는 다시 크게 유사계수 방식, 정보이론 기반 방식, 통계적 검증방식의 세 종류로 나눌 수 있다.

3.1.1 유사계수

유사계수 공식은 정보검색에서 문헌과 문헌, 문헌과 용어 사이의 유사도를 측정하거나 용어 클러스터링에서 클러스터 사이의 유사도를 측정하는 데 널리 쓰여왔다. 여러 가지의 유사계수 공식이 있으나 벡터공간 검색모형에서 출발한 코사인 계수와 클러스터링에 주로 쓰이는 타니모토 계수가 주로 이용된다.

$$Cosine(x,y) = \frac{f_{xy}}{\sqrt{f_x}\times\sqrt{f_y}}$$

$$Tanimoto(x,y) = \frac{f_{xy}}{f_x+f_y-f_{xy}}$$

3.1.2 정보이론 기반 연관계수

Shannon으로부터 시작된 정보 이론에 기반하여 연관성 측정에 쓰이는 것으로 상호정보량과 상대 엔트로피 공식이 있다.

(1) 상호정보량

상호정보량이란 두 독립사건의 확률변수 X와 Y 사이의 의존관계를 정량적으로 나타낸 것으로서 공식은 다음과 같다.

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

상호정보량은 두 확률이 완전히 독립적일 경우 0이 되고 의존 관계가 깊을수록 높은 값을 가진다. 상호정보량은 대칭성을 만족하는 함수이므로 $MI(x,y) = MI(y,x)$가 성립한다.

연관성 분석에 상호정보량을 이용할 때 문제점으로는 저빈도 단어 사이의 상호정보량이 고빈도 단어 사이의 상호정보량보다 상대적으로 과대평가되는 경향이 주로 지적된다.

(2) 상대 엔트로피

상대 엔트로피는 KL 거리 (Kullback-Leibler distance; D_{KL}), 교차 엔트로피(cross entropy)라고도 불리며 두 확률 분포 p(x)와 q(x) 사이의 평균적인 차이를 측정하는 것으로서 연관성 측정을 위해서는 주로 다음 형태의 공식이 쓰인다(정석경, 1997).

$$D_{KL}(p(x)||q(x)) = \sum_x p(x)[\log\frac{1}{q(x)} - \log\frac{1}{p(x)}]$$

상대 엔트로피 값은 항상 0보다 크거나 같으며 두 확률이 일치할 경우에만 0이 되고, 대칭성을 만족하지 않는다.

3.1.3 통계학 이론에 근거한 연관계수

분석 대상 텍스트의 전체적인 규모가 작거나 대상 단어의 빈도가 낮은 경우에는 공기빈도도 자연히 작은 값을 가지므로 통계학적인 검증기법을 적용할 필요가 있다.

통계학적인 검증기법으로 공기빈도를 분석할 경우에는 주로 <표 1>과 같은 2×2 분할표(contingency table)를 작성하여 분석한다.

<표 1> 공기빈도의 2×2 분할표

		단어 y		
		출현	미출현	합계
단어 x	출현	f_{11}	f_{12}	$f_{\cdot 1}$
	미출현	f_{21}	f_{22}	$f_{\cdot 2}$
	합계	$f_{1\cdot}$	$f_{2\cdot}$	$f_{\cdot\cdot}$

(1) z점수

z점수는 표준점수라고도 하며, 평균과 각 관찰값과의 차를 표준편차로 나누어 구한 값으로 평균이 0이고 표준편차가 1인 분포로 전환된다.

$$z = \frac{\mu - f_n}{\sigma}$$

텍스트 안의 주어진 단어에 공기한 모든 다른 단어의 실제 출현 빈도를 기대 빈도와 비교하는 것이 z검증이다. z점수가 높을 수록 평균을 벗어나 연관성이 높다고 볼 수 있다. z점수도 분석대상 빈도가 적은 경우 과대평가되는 경향이 있다고 지적된다.

(2) χ^2 통계량

χ^2는 관찰값과 기대값 사이의 차이에 대한 통계적인 검증에 이용되는 값으로서 2×2 분할표에 적용하면 다음과 같은 공식이 된다.(Kageura, 1997, 3)

$$X^2 = \frac{f_{\cdot\cdot}(f_{11}f_{22} - f_{12}f_{21})^2}{f_{1\cdot}f_{2\cdot}f_{\cdot 1}f_{\cdot 2}}$$

2×2 분할표에서는 $\chi^2 = z^2$ 인 관계가 성립한다. 따라서 χ^2도 z값처럼 연관성 분석에서는 분석대상 빈도가 적은 경우 과대평가할 여지가 있다. 또한 위 공식에서도 알 수 있듯이 표본의 크기가 커지면 분자의 $f_{\cdot\cdot}$(= 총 단어 수 N)도 커지므로 비례해서 χ^2도 커지는 문제가 있다.

(3) 율(Yule)의 결합계수 Y

율의 결합계수 Y는 2×2 분할표에서 다음 공식으로 구한다.

$$Y = \frac{\sqrt{\frac{f_{11}f_{22}}{f_{12}f_{21}}} - 1}{\sqrt{\frac{f_{11}f_{22}}{f_{12}f_{21}}} + 1}$$

이 값은 -1 이상 1이하의 값을 가지며 1에 가까울수록 긍정적 연관성을, -1에 가까울수록 부정적 연관성을 의미한다.

공식과 <표 1>의 분할표에서 알 수 있듯이 이 값은 합계를 이용하지 않고 교차곱의 비율을 이용하므로 대상 자료의 규모에 영향을 받지 않는다. 또한 비교대상인 두 단어의 출현빈도가 많거나 적은 것에 영향받지 않는다. 따라서 여러 표본에서 얻어낸 공기빈도를 비교분석하기에 적절하다.

(4) 우도비 검증법

우도비(尤度比; likehood ratio)는 관찰된 현상의 확률과 이 현상이 이론적으로 발생할 확률간의 비율을 뜻한다. 확률은 더하거나 빼기가 곤란하므로 차이를 구하기 위해서는 각각의 우도를 로그화하여 비율을 구해야 한다. <표 1>과 같은 2×2 분할표에 적용하기 위한 우도비 공식은 여러 형태가 있지만 Kageura(1997)가 제시한 다음 공식이 비교적 간단하다.

$$-2\log\lambda = 2[\sum_c \log L(f_{1c}/f_{\cdot c}, f_{1c}, f_{\cdot c}) - \sum_c \log L(f_{1\cdot}/f_{\cdot\cdot}, f_{1c}, f_{\cdot c})]$$

우도비와 χ^2은 둘 다 모두 표본의 크기가 커짐에 비례해서 값이 증가한다. 또한 분석 대상인 두 단어의 빈도에도 비례한다.

여타 연관계수와 다른 우도비의 특징은 표본의 크기가 클수록 개별 빈도와 공기빈도가 높고 둘 다 나타나지 않은 빈도가 상대적으로 적은 경우에 높은 값을 가진다는 점이다.

4. 한국어 텍스트에 대한 실험

4.1 실험용 신문기사 데이터베이스 구축

텍스트 데이터베이스를 구축할 때 고려한 사

항은 다음과 같다.

① 언어 : 한국어 자료를 대상으로 한다.

② 규모 : 명확한 기준은 없으나 일반적으로 최소 100만 어절 이상을 필요로 한다.

③ 분야 : 주제분야별 구성이 균등해야 한다.

④ 시기 : 가급적 최근 자료이되 시기별 비교가 가능해야 한다.

⑤ 품질 : 오·탈자에 대한 일차 검증을 거친 기계가독형 텍스트를 입수하는 것이 비용/효과 면에서 유리하다.

⑥ 가공 : 원하는 통계 데이터를 분석하기 쉽도록 헤더를 비롯한 태그가 적절히 부착되어야 한다.

이상과 같은 원칙을 고려하여 PC통신 하이텔을 통해 제공되는 KINDS 서비스를 이용하여 직접 구축하였다. 구축 대상으로 C일보 신문기사 중 주요 면에 해당하는 종합/정치, 경제, 사회 3면을 선택하여 1996년 기사와, 1997년 기사가 각각 100만 어절을 넘기도록 5개월분량을 다운로드받아 가공하여 헤더와 태그를 부착하였다. 구축 결과는 아래 표와 같다

<표 2> 전체 기사 건수와 월 평균 건수

구분 / 시기	정치		경제		사회		합계	
	계	평균	계	평균	계	평균	계	평균
1996년	2321	464.2	2968	593.6	2929	585.8	8218	1643.6
1997년	2848	569.6	2876	575.2	2613	522.6	8337	1667.4
합 계	5169	516.9	5844	584.4	5542	554.2	16555	1655.5

<표 3> 전체 어절 수와 월 평균 어절 수

구분 / 시기	정치		경제		사회		합계	
	계	평균	계	평균	계	평균	계	평균
1996년	312778	135.1	438739	147.9	383650	131.3	1135167	138.1
1997년	501333	164.4	422405	146.3	360111	137.8	1283849	149.5
합 계	814111	157.5	861144	147.4	743761	134.2	2419016	146.1

<표 4> 전체 색인 단어 수와 기사 당 평균

구분 / 시기	정치		경제		사회		합계	
	계	평균	계	평균	계	평균	계	평균
1996년	197616	85.1	287349	96.8	239058	81.6	724023	88.1
1997년	304661	107.0	273269	95.0	220046	84.2	797976	95.7
합 계	502277	97.2	560618	95.9	459104	82.8	1521999	91.9

4.2 연관계수 적용 실험

앞에서 파악한 각 연관계수를 이용하고 공기범위는 동일 기사/문단/문장을 각각 적용하여 연관성을 측정한다. 특히 빈도수준과 연관성의 상관관계를 파악하기 위해 고빈도어-저빈도어, 저빈도어-저빈도어, 고빈도어-고빈도어의 세 가지 유형의 용어쌍의 공기 특성을 분석한다.

5. 결론

Church와 Hanks(1990)가 연어(collocation) 분석에 사용할 것을 제안한 이후 상호정보량은 다양한 분야에서 가장 폭넓게 이용되고 있다.

그러나 여러 연관계수는 각각 그 유래에 따라서 다른 특성을 가지며, 둘 이상의 연관계수를 사용한 실험적 연구에서는 연관계수에 따라 결과가 달라지는 사례가 나타나고 있다.

따라서 한국어 텍스트에서 용어연관성을 분석하고 이를 적절히 이용하기 위해서는 본 연구에서 파악되는 여러 연관계수의 특성을 각 응용분야에 적용해보는 실험이 필요하다.

참고문헌

정석경. 1997. 분포 정보를 이용한 명사 소프트 클러스터링 연구. 연세대학교 석사학위논문.

Church, K. W., and P. Hanks. 1990. "Word Association Norms, Mutual Information, and Lexicography," *Computational Linguistics* 16(1): 22-29.

Kageura, Kyo. 1997. *Bigram Statistics Revisited: A Comparative Examination of Some Statistical Measures in Morphological Analysis of Japanese Kanji Sequences*. <http://www.dcs.shef.ac.uk/~kyo/llc.ps>.

Nobesawa, Shiho et al. 1996. "Segmenting Sentences into Linky Strings Using D-bigram Statistics," *Proceedings of the 16th International Conference on Computational Linguistics*, 586-591.