한국어의 정보이론적 연구 방향

이 재홍, 이 재학
서울대학교 전자공학과

On Information Theoretical Research of the Korean Language

Jae Hong Lee and Chaehag Yi
Department of Electronics Engineering, Seoul National University

요 약

한국어는 다른 언어와는 달리 초성, 중성, 종성의 자소가 모여서 한 음절을 이룬다. 음절을 이루는 자소는 그 발생의 확률적 성질에 따라 확률변수로 간주된다. 음절 안에서 자소간의 발생의 상관관계는 자소간 조건부 확률 및 엔트로피로 표시된다. 음절이 모여서 단어를 이루고 단어를 이루는 음절은 그 발생의 확률적 성질에 따라 확률변수로 간주된다. 한국어 단어안에서 음절간의 발생의 상관관계는 음절간 조건부 확률 및 엔트로피로 표시된다. 수 있다. 그런데 가능한 음절의 종류가 매우 많기 때문에 음절 발생의 상관관계를 표시하는 지표로서 음절간 조건부 확률 대신 초성, 중성, 종성 단위의 조건부 확률을 사용하는 것이 음절간의 발생의 상관관계를 표시하는데 효과적이다.

이러한 한국어의 정보이론적 연구를 위하여서는 기초자료로서 한국어 단어의 빈도분포가 필요하다. 한국어 단어의 빈도분포의 포괄적인 조사는 1956년의 "우리말 말수 사용의 잦기 조사"가 유일한 실정이다. 시간 경과에 따른 한국어의 정보이론적 특성 변화의 분석을 위하여서는 한국어 단어 빈도의 주기적인 조사가 필요하다. 한국어에서 초성, 중성, 종성 단위의 정보이론적 연구결과는 한국어 음성인식 및 합성, 자연언어처리, 암호법, 언어학, 음성학, 한국어부호 표준화 연구등에 이용될 것으로 기대된다.

남북한의 언어는 분단이 지속됨에 따라 상호 이질화가 진행되고 있다. 이러한 이질화를 극복하려는 부분적인 노력으로 남북한 언어의 한국어 영문표기의 단일화 등이 있었다. 이러한 노력에 병행하여 남한과 북한의 언어에 대한 정보이론적 비교 연구도 있어야 할 것이다.

I. 서 론

언어는 인간의 의사소통의 가장 중요한 수단으로서 언어가 문자화된 것이 글이다. 글은 기호들이 연속적으로 나열되어 형성된 기호열로 볼 수 있다. 이렇게 형성된 기호열은 정보를 내포하는데, 기호열이 내포하는 정보량을 계량하려는 정보이론적 시도가 있어 왔다. 즉 C. Shannon이 정보이론에 기초하여 영어 기호열의 정보량 척도로서 엔트로피(entropy)를 제안하고 그 측정 방법을 제안한 이래 몇몇 언어에 대하여 그 엔트로피가 계산되었다[1][2]. 한글에 있어서는 Shannon의 측정 방법을 사용한 자소단위의 엔트로피에 관한 연구가 몇 차례 있었다[3-5]. 이들 연구에서는 한국어의 24자모를 영어의 26자의 알파벳에 대응되는 것으로 보고 영어에서와 동일한 측정 방법을 사용하여 자소단위의 엔트로피를 구하였다.

한국어는 다른 언어와는 달리 초성, 중성, 종성의 자소가 모여서 한 음절을 이룬다. 음절을 이루는 자소는 그 발생의 확률적 성질에 따라 확률변수로 간주된다. 음절 안에서 자소간의 발생의 상관관계는 자소간 조건부 확률 및 엔트로피로 표시한다. 이러한 특성을 고려한 한국어 음절의 초성, 중성, 종성 단위의 발생확률과 단위간의 조건부 확률에 관한 연구가 발표된 바 있다[6]. 음절이 모여서 단어를 이루고 단어를 이루는 음절은 그 발생의 확률적 성질에 따라 확률변수로 간주된다. 한국어 단어안에서 음절간의 발생의 상관관계는 음절간 조건부 확률 및 엔트로피로 표시된다. 그런데 가능한 음절의 종류가 매우 많기 때문에 음절 발생의 상관관계를 표시하는 지표로서 음절간 조건부 확률 대신 초성, 중성, 종성 단위의 조건부 확률을 사용하는 것이 음절간의 발생의 상관관계를 표시하는데 효과적이다. 이러한 특성을 고려한 한국어 다음절 단어의 초성, 중성, 종성 단위의 음절간 조건부 확률에 관한 연구가 발표된 바 있다[7]. 이러한 연구는 우리 민족의 고유언어이자 세계의 유일무이한 창조문자인 한국어의 정보이론적 분석이라는 관점에서 의미있는 일이다. 그리고 연구결과는 한국어 음성인식 및 합성, 자연언어처리, 암호법(cryptography), 언어학, 음성학, 한국어부호 표준화 연구[8-10]등에 이용될 것으로 기대된다.

이러한 한국어의 정보이론적 연구는 한국어 단어의 빈도분포를 기초자료로 사용한다. 지금까지 한국어 단어 또는 음절 단위의 빈도분포의 체계적 조사는 빈약하였으며 조사의 내용에 있어서도 다른 언어의 체계적인 분포 조사와 통계적 분석에 비하여 빈약하였다. 해방후 40여년간 한국어 단어별 발생빈도분포의 포괄적인 조사는 단 한번 실시되었을 뿐이고 그 이후에는 전무한 실정이다. 즉, 1956년 문교부에 의하여 한국어 단어가 사용되는 빈도를 조사하여 과학적인 국어의 기본형태를 파악하고 우리말의 합리적인 사용을 꾀하는 것을 목적으로 "우리말 말수 사용의 잦기 조사"[11]가 실시되었다.. 조사의 대상이 된 목적물은 초중등교과서가 50%, 일반간행물 중 문학 예술류가 30%, 신문, 잡지, 방송원고, 국회의사록등이 20%의 비율로 구성되었다. 모두 56,077개의 단어가 수록되었는데 조사와 외래어가 포함되었고 동사와 형용사는 기본형의 형태로 수록되었다.

본 고에서는 한국어의 정보이론적 분석 방법을 서술하고 한국어에서의 초성, 중성, 종성

단위의 엔트로피와 평균상호정보량(average mutual information)의 연구 결과를 소개한다. 그리고 앞으로의 한국어의 정보이론적 연구에서 진행되어야 할 내용에 대하여 살펴본다.

Ⅱ. 한국어의 정보이론적 해석

한국어의 도형적 合字(이하 '음절')를 구성하는 초성, 중성, 종성은 각각 가능한 자소집합으로부터 어떤 확률로 발생하는 확률변수(random variable)로 간주할 수 있다. 초성, 중성, 종성을 표시하는 확률변수를 각각 X, Y, Z라 정의하고, 그 가능한 자소의 집합 즉, 표본공간(sample space)을 각각 A_X, A_Y, A_Z라 하고, 초성 X의 발생확률, 중성 Y의 발생확률, 종성 Z의 발생확률을 각각 $p(x)$, $p(y)$, $p(z)$라 하자. 그러면 A_X, A_Y, A_Z는 다음과 같이 주어진다.

A_X = { ㄱ,ㄲ,ㄴ,ㄷ,ㄸ,ㄹ,ㅁ,ㅂ,ㅃ,ㅅ,ㅆ,ㅇ,ㅈ,ㅉ,ㅊ,ㅋ,ㅌ,ㅍ,ㅎ}

A_Y = {ㅏ,ㅐ,ㅑ,ㅒ,ㅓ,ㅔ,ㅕ,ㅖ,ㅗ,ㅘ,ㅙ,ㅚ,ㅛ,ㅜ,ㅝ,ㅞ,ㅟ,ㅠ,ㅡ,ㅢ,ㅣ}

A_Z = { 공백소,ㄱ,ㄲ,ㄱㅅ,ㄴ,ㄴㅈ,ㄴㅎ,ㄷ,ㄹ,ㄹㄱ,ㄹㅁ,ㄹㅂ,ㄹㅅ,ㄹㅌ,

ㄹㅍ,ㄹㅎ,ㅁ,ㅂ,ㅂㅅ,ㅅ,ㅆ,ㅇ,ㅈ,ㅊ,ㅋ,ㅌ,ㅍ,ㅎ}

공백소(blank)는 자소가 결여된 상태를 표시하며 종성에만 발생한다. 확률변수 X, Y, Z의 표본공간의 크기는 각각 $|A_X|$=19, $|A_Y|$=21, $|A_Z|$=28 이다.

하나의 음절은 초성, 중성, 종성으로 구성되는데 확률변수 X, Y, Z의 결합확률 $p(x,y,z)$는 특정 음절의 발생확률을 나타낸다. 음절은 그 발생의 확률적 성질에 따라 확률변수로 간주될수 있다. 음절을 표시하는 확률변수를 S라 정의할 때 확률변수 S의 표본공간 A_S는 다음과 같이 주어진다.

$$A_S = A_X \times A_Y \times A_Z$$

따라서 확률변수 S의 표본공간의 크기는 $|A_S|=|A_X| \times |A_Y| \times |A_Z|$ 이다. 음절들이 모여서 이루어지는 단어도 확률변수로 간주할수 있다.

정보량을 나타내는 척도로 엔트로피와 평균상호정보량을 사용한다. 엔트로피의 유형에는 초성 X의 엔트로피 $H(X)$, 초성, 중성 (X,Y)의 결합 엔트로피 $H(X,Y)$, 초성, 중성, 종성 (X,Y,Z)의 결합 엔트로피, 즉 음절의 엔트로피 $H(X,Y,Z)$, 중성 Y이 주어질 때 초성 X의 조건부 엔트로피 $H(X|Y)$, 종성 Z이 주어질 때 초성, 중성 (X,Y)의 조건부 엔트로피 $H(X,Y|Z)$, 중성,종성 (Y,Z)이 주어질 때 초성 X의 조건부 엔트로피 $H(X|Y,Z)$가 있으며 각각 다음과 같이 주어진다.

$$H(X) = -\sum_x p(x)\log p(x) \qquad (1)$$

여기서 $p(x)$는 초성 X의 발생확률이다.

$$H(X,Y) = -\sum_{x,y} p(x,y)\log p(x,y) \tag{2}$$

여기서 $p(x,y)$는 초성, 중성 (X,Y)의 결합확률이다.

$$H(X,Y,Z) = -\sum_{x,y,z} p(x,y,z)\log p(x,y,z) \tag{3}$$

여기서 $p(x,y,z)$는 초성, 중성, 종성 (X,Y,Z)의 결합확률 즉 음절의 발생확률이다.

$$H(X|Y) = -\sum_{x,y} p(x,y)\log p(x|y) \tag{4}$$

여기서 $p(x|y)$는 중성이 Y때 초성이 X일 조건부 확률이다.

$$H(X,Y|Z) = -\sum_{x,y,z} p(x,y,z)\log p(x,y|z) \tag{5}$$

여기서 $p(x,y|z)$는 종성이 Z일때 초성이 X이고 중성이 Y일 조건부 확률이다.

$$H(X|Y,Z) = -\sum_{x,y,z} p(x,y,z)\log p(x|y,z) \tag{6}$$

여기서 $p(x|y,z)$는 중성이 Y이고 종성이 Z일때 초성이 X일 조건부 확률이다.
이러한 유형의 엔트로피들 간에 다음의 관계가 성립한다.

$$H(X,Y) = H(X)+H(Y|X) \tag{7}$$

$$H(X,Y,Z) = H(X,Y)+H(Z|X,Y) \tag{8}$$

$$H(X,Y,Z) = H(X)+H(Y,Z|X) \tag{9}$$

이 엔트로피로부터 초성 X와 중성 Y간의 평균상호정보량 $I(X;Y)$와 초성, 중성 (X,Y)와 종성 Z간의 평균상호정보량 $I(X,Y;Z)$와 초성 X와 중성, 종성 (Y,Z)간의 평균상호정보량 $I(X;Y,Z)$이 다음과 같이 구해진다.

$$I(X;Y) = H(X)-H(X|Y) \tag{10}$$

$$I(X,Y;Z) = H(X,Y)-H(X,Y|Z) \tag{11}$$

$$I(X;Y,Z) = H(X)-H(X|Y,Z) \tag{12}$$

평균상호정보량간에 다음의 관계가 성립한다.

$$I(X;Y) = I(Y;X) \tag{13}$$

$$I(X;Y,Z) = I(X,Y;Z) \tag{14}$$

Ⅲ 한국어의 엔트로피와 정보량

"우리말 말수 사용의 잦기 조사"[11]에 수록된 56,077개의 단어에 대한 빈도순서와 잦기로부터 다음절 단어의 길이 분포와 음절의 발생확률을 계산하고 전체 음절의 초성, 중성, 종성 단위간의 엔트로피와 평균상호정보량을 계산하며 다음절 단어의 초성, 중성, 종성 단위의 음절간 조건부 엔트로피를 계산한다.

한국어 단어의 길이 분포를 표 1에 보인다. 단어의 길이를 나타내는 확률변수를 L이라 하자. 여기서 L=10은 길이가 10인 단어와 그 이상인 단어를 나타낸다. 한국어 단어는 1음절 단어의 빈도가 가장 높음을 표 1에서 볼수 있다. 한국어 단어에서 1음절 단어의 발생빈도가 가장 높은 것은 조사가 있기 때문이다. 단어의 평균 길이는 1.82 음절/단어 임을 표 1에서 볼수 있다. 대부분의 한국어 단어의 길이는 1 또는 2임을 의미한다.

확률변수 X, Y, Z의 표본공간의 크기는 각각 $|A_X|=19$, $|A_Y|=21$, $|A_Z|=28$ 이므로 가능한 초성, 중성, 종성의 결합 가지수는 $19 \times 21 \times 28 = 11{,}172$개이다. 그런데 "우리말 말수 사용의 잦기 조사"[11]에서 조사된 단어에는 모두 1,535개의 음절만이 사용되었다. 이는 모든 가능한 초성, 중성, 종성의 결합의 약 13.7%만이 발생됨을 뜻한다. 한국어 단어에 있어서 발생되는 음절의 종류는 모든 가능한 초성, 중성, 종성의 결합에 비하여 매우 적음을 알수 있다. 발생되는 음절을 발생확률이 높은 순서로 나열했을 때 가장 많이 발생되는 음절 열 개와 그것의 발생확률을 표 2에 보인다. 위의 자료에서 동사와 형용사는 기본형으로 수록되어 있으므로 음절 '다'와 '하'가 다른 음절에 비하여 발생빈도가 높다. 실제로 발생되는 1,535개의 음절중 가장 많이 발생되는 열 개의 음절이 약 36.0%를 차지한다.

전체 단어에서 발생 빈도를 고려하여 음절별 누적 빈도수를 구하였다. 이로부터 음절의 초성, 중성, 종성 단위의 엔트로피와 평균상호정보량을 계산하여 표 3에 보인다. 초성, 중성, 종성에서 초성은 그 종류가 가장 적지만 엔트로피는 제일 크고 종성은 그 종류는 가장 많지만 엔트로피는 제일 작다. 초성, 중성, 종성의 결합 엔트로피 $H(X,Y,Z)$는 7.24 bits 로, 매우 큰 음절의 다양성 비하면 작은 값이어서 음절의 리던던시(redundancy)가 큼을 알수 있다.

한국어 다음절 단어안에서 음절간의 발생의 상관관계를 표시하는 지표로서 초성, 중성, 종성 단위의 음절간 조건부 엔트로피를 사용한다. 앞음절 초성 X_p 와 다음 음절 초성 X_n의 조건부 엔트로피, 앞음절 중성 Y_p 와 다음 음절 초성 Y_n의 조건부 엔트로피, 앞음절 종성 Z_p 와 다음 음절 종성 Z_n의 조건부 엔트로피와 앞음절 종성 Z_p 와 다음 음절 초성 X_n의

조건부 엔트로피를 계산하여 표 4에 보인다.

Ⅳ. 지금까지의 연구결과 및 향후의 연구방향

한국어는 다른 언어와는 달리 초성, 중성, 종성의 자소가 모여서 한 음절을 이룬다. 음절을 이루는 자소는 그 발생의 확률적 성질에 따라 확률변수로 간주된다. 음절 안에서 자소간의 발생의 상관관계는 자소간 조건부 확률 및 엔트로피로 표시한다. 이러한 특성을 고려한 한국어 음절의 초성, 중성, 종성 단위의 발생확률과 단위간의 조건부 확률에 관한 연구가 발표된 바 있다. 음절이 모여서 단어를 이루고 단어를 이루는 음절은 그 발생의 확률적 성질에 따라 확률변수로 간주된다. 한국어 단어안에서 음절간의 발생의 상관관계는 음절간 조건부 확률 및 엔트로피로 표시된다. 그런데 가능한 음절의 종류가 매우 많기 때문에 음절 발생의 상관관계를 표시하는 지표로서 음절간 조건부 확률 대신 초성, 중성, 종성 단위의 조건부 확률을 사용하는 것이 음절간의 발생의 상관관계를 표시하는데 효과적이다. 이러한 특성을 고려한 한국어 다음절 단어의 초성, 중성, 종성 단위의 음절간 조건부 확률에 관한 연구가 발표된 바 있다. 이러한 연구결과는 한국어 음성인식 및 합성, 자연언어처리, 암호법(cryptography), 언어학, 음성학, 한국어부호 표준화 연구등에 이용될 것으로 기대된다.

이러한 연구를 위한 기초자료로서 한국어 단어의 빈도분포가 필요하다. 해방 이래 한국어의 단어 또는 음절 단위의 빈도 분포의 체계적 조사는 1956년 단 한번 뿐이었다. 조사 시기 이후의 사회적 변화, 인구의 이동 등에 의한 한국어 표현 방법의 변화에 따라 한국어 단어의 빈도 분포도 변화하였을 것으로 추정되는 바 지금까지의 저자에 의한 연구는 사용된 자료의 조사 시점에 한한 연구라는 한계점을 가진다. 시간 경과에 따른 한국어의 정보이론적 특성 변화의 분석을 위하여서는 한국어 단어의 빈도의 주기적인 조사가 필요하다.

남북한의 언어는 분단이 지속됨에 따라 상호 이질화가 진행되고 있다. 이러한 이질화를 극복하려는 부분적인 노력으로 남북한 언어의 한국어 영문표기의 단일화 등이 있었다. 이러한 노력에 병행하여 남한과 북한의 언어에 대한 정보이론적 비교 연구도 있어야 할 것이다.

參 考 文 獻

[1] C. Shannon, "Prediction and entropy of printed English," *Bell Syst. Tech. J.*, vol. 30, pp. 50-64, Jan. 1951.

[2] R. Manprino, "Printed Portuguese entropy-statistical calculation," *IEEE Trans. on Inform. Theory*, vol. IT-16, p. 122, Jan. 1970.

[3] 이주근, 최홍문, "한국어 음절의 entropy에 관한 연구," 전자공학회지, 제 11권,제 3호, pp. 15-21, 1974년 6월.

[4] 안수길, 안지환, "공백소를 포함한 한글 자소발생 확률과 엔트로피," 전자공학회지, 제

17권, 제 2호, pp. 23-28, 1980년 4월.

[5] 남궁건, 한글 낱말의 발생빈도 분포의 엔트로피에 관한 연구. 석사학위논문, 서울대학교, 1979.

[6] 이재홍, 오상현, "한글 음절의 초성, 중성, 종성 단위의 발생확률, 엔트로피, 평균상호정보량," 전자공학회논문지, 제 26권, 제 9호, pp. 1-9, 1989년 9월.

[7] 이재홍, 이재학, "한국어 다음절 단어의 초성, 중성, 종성 단위의 음절간 조건부 확률," 전자공학회논문지, 제 28권 B편, 제 9호, pp. 1-11, 1991년 9월.

[8] 과학기술처, 컴퓨터 관련 표준화 규격. 과학기술처, 1982.

[9] 한국표준연구소, 한글 정보처리 표준화연구. 과학기술처, 1986.

[10] 한국표준연구소, 한자부호 표준시안 작성을 위한 연구. 과학기술처, 1987.

[11] 문교부, 우리말 말수 사용의 잦기 조사. 문교부 1956.

표 1. 한국어 다음절 단어의 길이 분포
Table 1. Word length distribution of Korean words.

L	1	2	3	4	5	6	7	8	9	10
p(l)	.421480	.413039	.099342	.061079	.004240	.000730	.000071	.000011	.000004	.000003

표 2. 음절의 발생확률
Table 2. Syllable probability.

S	다	이	하	에	가	의	을	는	리	그
p(s)	.152224	.045519	.044541	.023764	.018903	.018675	.018591	.013200	.012473	.012175

표 3. 초성, 중성, 종성의 엔트로피와 평균 상호 정보량
Table 3. Entropies and average mutual information for a 'choseong', a 'jungseong', and a 'jongseong'.

(a) Entropies

	bits
H(X)	3.39
H(Y)	3.11
H(Z)	2.09
H(X,Y)	5.94
H(Y,Z)	4.87
H(X,Z)	5.24
H(X¦Y)	2.83
H(Y¦X)	2.54
H(Y¦Z)	2.78
H(Z¦Y)	1.76
H(Z¦X)	1.85
H(X¦Z)	3.15
H(X,Y,Z)	7.24
H(X,Y¦Z)	5.15
H(Y,Z¦X)	3.85
H(X,Z¦Y)	4.14
H(X¦Y,Z)	2.37
H(Y¦X,Z)	2.00
H(Z¦X,Y)	1.31

(b) Average mutual information

	bits
I(X;Y)	0.56
I(Y;Z)	0.32
I(Z;X)	0.24
I(X;Y,Z)	1.02
I(Y;X,Z)	1.10
I(Z;X,Y)	0.78

표 4. 단어내에서의 음절간 초성, 중성, 종성의 조건부 엔트로피
Table 4. Conditional entropies for a 'choseong', a 'jungseong' and a 'jongseong' between syllables in a word.

	bits
H(Xn¦Xp)	2.95
H(Xp¦Xn)	3.35
H(Yn¦Yp)	2.46
H(Yp¦Yn)	3.00
H(Zn¦Zp)	1.51
H(Zp¦Zn)	2.13
H(Xn¦Zp)	2.98
H(Zp¦Xn)	1.99