단어의 정보모형작성에서 제기되는 몇가지 문제

안 성 득

위대한 수령 **김일성**동지께서는 다음과 같이 교시하시였다.

《새로운 과학분야를 개척하며 최신과학기술의 성과를 인민경제에 널리 받아들이기 위한 연구사업을 전망성있게 하여야 합니다.》(《**김일성**전집》 제72권 292페지)

단어는 언어에서 기본단위이며 단어가 없는 언어를 생각할수 없다.

언어가 교제의 수단이며 혁명과 건설의 무기라고 할 때 그것은 단어의 존재를 전제로 한다.

언어는 어휘와 문법, 의미를 자기의 연구대상으로 한다.

여기에서도 기본은 어휘 즉 단어이다.

문법과 의미도 언어연구에서 중요한 자리를 차지하지만 단어를 떠난 문법이나 의미에 대하여 생각할수 없다.

단어가 언어의 기본단위로 되는것은 첫째로, 그것이 말소리의 측면에서나 뜻의 측면에서 언어의 기본단위로 되기때문이며 둘째로, 단어가 문법구조를 이루게 하며 그 기초로 되는 언어적단위로 되기때문이다.

문법적범주와 문법적뜻은 단어를 통하여 실현되며 문법적관계는 기본적으로 형태단어들사이의 호상관계에 의하여 표현된다.

단어는 사람들의 사상과 감정을 표현하고 전달하기 위한 기본요소이며 언어의 기본단위이다.

그러므로 언어의 어떤 분야를 연구하든지간에 기본적으로는 단어를 대상으로 하게 되며 모든 언어적연구는 단어와 그에 기초한 연구로 된다.

단어는 언어의 다른 단위들인 음운, 형태부, 단어결합, 문장 등과 본질적으로 구별되면서 언어발전에서 중요한 자리를 차지하는 언어의 기본단위로 된다.

응용언어학견지에서 단어를 정식화할수 있다.

일반자모에서 순서있게 씌여진 글자들의 유한렬을 단어라고 부른다.

단어를 간단히 일반자모 V 에서 얻어진 유한렬로 정의할수 있다.

여기서 일반자모는 글자라고 부르는 유한개 원소들의 모임을 말한다.

일반자모의 글자로서는 조선어, 영어, 일본어 등의 자모도 될수 있고 수자나 기호 등도 될수 있다.

일반자모를 간단히 기호(글자)들의 비지않은 유한모임으로 정의할수 있다.

V 를 조선어일반자모라고 하면 $V=\{$ㄱ, ㄴ, ㄷ, ㄹ, ㅁ, ㅂ, ㅅ, ㅇ, ㅈ, ㅊ, ㅋ, ㅌ, ㅍ, ㅎ, ㄲ, ㄸ, ㅃ, ㅆ, ㅉ, ㅏ, ㅑ, ㅓ, ㅕ, ㅗ, ㅛ, ㅜ, ㅠ, ㅡ, ㅣ, ㅐ, ㅒ, ㅔ, ㅖ, ㅚ, ㅟ, ㅢ, ㅘ, ㅝ, ㅙ, ㅞ$\}$로서 40개의 글자들로 이루어진다.

단어에는 단어를 이루는 매 글자에 1부터 n까지 번호를 붙일수 있다.

즉 주어진 단어에 대하여 왼쪽으로부터 오른쪽으로 가면서 자연수번호를 붙인다.

가령 단어를 α라고 하면

$$\alpha=\alpha_1\alpha_2\cdots\alpha_n$$

여기서 α_n 은 단어 α 를 이루는 n 번째 글자이다.

례: 단어《사이다》가 주어졌다면

$\alpha_1=$ ㅅ, $\alpha_2=$ ㅏ, $\alpha_3=$ ㅇ, $\alpha_4=$ ㅣ, $\alpha_5=$ ㄷ, $\alpha_6=$ ㅏ

일반자모 V 에서 얻어진 가능한 모든 유한렬들의 모임을 V^*이라고 하면 V^*에서는 원소들의 결합법칙이 성립한다.

가령 α, β가 V^*의 원소 즉 단어라면

$$\alpha=\alpha_1\alpha_2\cdots\alpha_n$$
$$\beta=\beta_1\beta_2\cdots\beta_m$$

여기서 임의의 $\alpha_i\ (i=\overline{1,n})$ 와 $\beta_j\ (j=\overline{1,m})$ 는 모두 V 의 원소들이다.

즉 $\alpha_i \in V\ (i=\overline{1,n})$

$\beta_i \in V\ (j=\overline{1,m})$

α와 β가 V^*의 원소이면 $\alpha\beta$ 역시 V^*의 원소로 된다.

실례로 α=인민, β=과학자이라면 $\alpha\beta$=인민과학자로서 역시 단어들의 모임 V^*에 속한다.

단어 α를 이루는 자모들의 개수를 주어진 단어 α의 길이라고 부르며 다음과 같이 표시한다.

단어 α의 길이$=\|\alpha\|$

우의 실례에서 $\|\alpha\|=6$이고 $\|\beta\|=7$이다.

결국 $\|\alpha\|=n$, $\|\beta\|=m$ 이라면 $\|\alpha\beta\|=n+m$이 성립한다.

어떤 유한렬을 이루는 자모들의 개수가 0이면 그런 유한렬을 ε이라고 표시하고 V^*의 단위원소라고 부른다.

즉 임의의 유한렬 α에 대하여 $\varepsilon\alpha=\alpha\varepsilon=\alpha$가 성립된다.

이때 ε을 빈단어라고 부른다.

실례로 A를 로어일반자모라고 하면 студент, математика, множество,...등은 A^*의 원소 즉 로어단어들로 되며 그 길이는 각각 7, 10, 9,…등이다.

B를 영어일반자모라고 하면 student, set, work,…등은 B^*의 원소 즉 영어단어들로 되며 그 길이는 각각 7, 3, 4,…등이다.

실례를 통하여 우리는 어떤 언어의 단어들의 길이는 다른 언어의 단어들의 길이보다 길거나 짧다는것을 알수 있다. 이 사실을 고려하면 주어진 언어의 단어들의 평균길이가 얼마인가 하는것을 계산할수 있다.

단어를 이루는 자모들을 왼쪽으로부터 오른쪽으로 점차 이동해가면서 정보량의 변화가 어떻게 일어나는가에 대하여 보기로 하자.

단어를 이루는 매 자모들이 가지고있는 불확정성은 서로 다르다.

실례로 《ㅎㅕㄱㅁㅕㅇ》이라는 단어에서 첫 글자 《ㅎ》가 가지는 불확정성이 가장 크고 다음에는 《ㅕ》, 그 다음에는 《ㄱ》 등으로 된다.

마지막글자 《ㅇ》은 유한렬 《ㅎㅕㄱㅁㅕ》가 주어진 조건에서 매우 작은 불확정성을 가진다. 즉 조선어를 잘 아는 사람이라면 사실 《ㅎㅕㄱㅁㅕ》가 주어진 다음에 어떤 글자가 출현하겠는가고 하면 확신성있게 《ㅇ》이 출현한다고 대답할것이다.

즉 《ㅇ》의 출현사건은 확실한 사건으로 된다.

일반적으로 단어를 X, 자모를 $X=X_1X_2\cdots X_n$라고 하면 단어 X의 불확정성을 $H(X)$로 표시하면 단어에서는 일반적으로 다음과 같은 식이 성립한다.

$$H(X_1)\geq H(X_2)\geq\cdots\geq H(X_n)$$

불확정성에 대해서는 다음의 성질들이 성립한다.

① 시행의 결과개수 $n=1$이면 그의 불확정성 $H(n)=H(1)=0$

② 2개의 시행에서 $n_1>n_2$이면 불확정성은 $H(n_1)>H(n_2)$

③ $H(n_1\cdot n_2)=H(n_1)+H(n_2)$

여기서 n_1은 첫 시행의 결과개수이고 n_2는 두번째 시행의 결과개수이다.

단어의 정보량을 측정하는것은 여러가지 문제와 관련되며 특히 입말, 글말 등에서 군더더기를 줄이고 정보를 정확하고 명료하게 전달하는데서 중요한 의의를 가진다.

만일 단어가 문장이나 본문에서 아무런 정보도 가지고있지 못하거나 매우 적은 량을 가지고있다면 그 단어를 군더더기라고 한다.

군더더기는 입말에서 많이 나타나며 글말에서도 쓰인다.

그런데 과학기술문제에서, 특히는 그 콤퓨터처리에서는 아무런 정보도 없는 단어 즉 군더더기가 전혀 필요없다.

한마디로 문장구성에서 군더더기인 단어들은 제거하여야 한다.

그러자면 문장에서 핵심단어를 확정하고 그의 정보무게를 결정하여야 하며 정보무게가 허용한계보다 작다면 즉 정보의 크기가 매우 작으면 그 단어를 문장구성에서 제거하여야 한다.

단어의 정보모형을 작성하기 위하여 수렬과 합렬의 개념을 도입하자.

어떤 수값들의 렬 $u_1, u_2, \cdots, u_n$이 주어졌을 때 이 렬을 수렬이라고 하며 $\{u_n\}$으로 표시한다.

그리고 u_n을 수렬 $\{u_n\}$의 일반항이라고 부른다.

만일 $S_1 = u_1$

$S_2 = u_1 + u_2$

$S_3 = u_1 + u_2 + u_3$

$\cdots \quad \cdots$

$S_n = u_1 + u_2 + \cdots + u_n$

이라고 하면 $\{S_n\}$을 합렬이라고 하며 S_n을 합렬 $\{S_n\}$의 부분합이라고 한다.

주어진 합렬 $\{S_n\}$에 대하여

$$\lim_{n\to\infty} S_n = S$$

가 성립하면 $\{S_n\}$을 수렴하는 합렬이라고 하며 S를 그의 합이라고 부른다.

만일 극한이 존재하지 않거나 ∞이면 합렬 $\{S_n\}$은 발산하는 합렬이라고 한다.

단어 $X = X_1 X_2 \cdots X_n$이 주어졌다면 앞에서 본바와 같이 단어를 이루는 자모들이 가지고있는 정보량은 단어의 왼쪽으로부터 오른쪽으로 옮겨가면서 점차 감소된다. 즉 $I(X_1) \geq I(X_2) \geq \cdots \geq I(X_n)$

단어를 이루는 자모들이 가지는 정보량을 알려면 초기엔트로피 H_0과 정보의 감소결수 μ가 주어진 경우 n개의 자모들로 이루어진 단어의 마지막자모 X_n가 가지는 불확정성을 모형화하여야 한다.

단어는 유한개의 원소(자모, 음절)로 이루어져있다.

현실에서는 단어의 길이(자모 또는 음절의 개수)가 매우 긴 경우와 자주 부닥치게 된다.

일반적으로 단어길이에 대하여 말할 때에는 단어의 평균길이의 개념을 도입하고 평균길이와 크게 차이없는 값을 념두에 둔다.

언어의 어휘구성안에 새로 생기는 단어들은 아무것도 없는 빈터우에서 창조되거나 마음대로 만들어지는것이 아니라 주어진 언어의 단어조성수법 혹은 외국어단어들을 차용하는 방법에 의하여 만들어진다.

특히 어근합성에 의하여 만들어진 합성어는 그 길이가 매우 길어질수 있다.

실례로 《조선민주주의인민공화국 사회주의헌법》을 들수 있다.

이 단어의 길이는 42이다. 지어 이보다 더 긴 단어들도 얼마든지 만들수 있다.

따라서 단어길이를 나타내는 량 ξ를 련속량으로 고찰하여도 언어연구에서 별로 큰 차이를 가져오지 않는다.

단어의 길이를 특징짓는 련속량 ξ를 띠염옹근수 n으로 바꾸고 불확정성(엔트로피)과 정보가 값에 있어서 같다는것을 리용하면 다음의 정보모형이 얻어진다.

즉 $\qquad I_n = I_0 e^{-\mu n}$

여기서 I_n은 단어 $X = X_1 X_2 \cdots X_n$의 n번째 자리에 있는 자모가 가지고있는 정보량이다.

얻어진 단어의 정보모형을 리용하면 단어의 첫 자모가 가지는 정보량은 $I_1 = I_0 e^{-\mu}$이고 둘째 자모가 가지는 정보량은 $I_2 = I_0 e^{-2\mu}$임을 알수 있다.

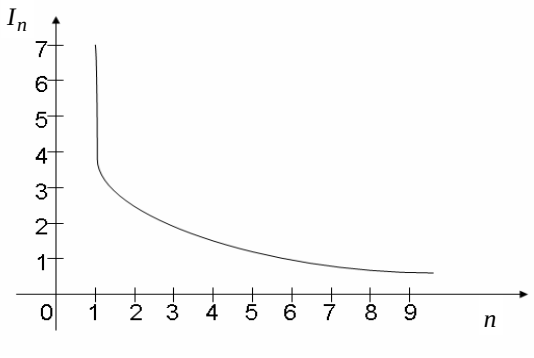
단어의 정보모형을 리용하여 단어의 최대 글자정보량이 얼마인가를 계산할수 있다.

즉 단어의 글자정보는 단어를 이루는 매개 자모들이 가지는 정보들의 총합과 같다.

단어의 최대정보량을 $I_{최}$라고 하면

$I_{최} = \dfrac{I_0}{1 - \dfrac{1}{e^{\mu}}}$으로 된다.

단어의 정보모형을 그라프로 그리면 다음과 같다.



단어의 정보모형그라프는 단어의 정보분포상태를 보여주며 이 그라프에 의하여 단어를 이루는 임의의 자모의 정보를 량적으로 계산할수 있다. 물론 언어마다 정보의 감소곁수 μ가 서로 다른 조건에서 그라프의 모양은 조금씩 달라진다. 단어의 정보모형에 있는 보조변수 μ의 특성을 보면 보조변수 μ는 단어의 정보량 I_ξ의 변화속도를 반영하는 곁수로서 주어진 단어의 특성에 따라 각이하게 계산되는 량이다. 보조변수 μ의 값이 클수록 정보량 I_ξ의 값은 작아지고 반대로 μ의 값이 작을수록 I_ξ의 값은 커진다. μ의 값이 단어에 따라 변하는 조건에서 실천에서는 통계적방법으로 그의 근사값을 택한다.

앞으로 응용언어학에 대한 연구를 심화시켜 나라의 과학기술을 하루빨리 세계적수준으로 발전시켜야 할것이다.