김일성종합대학학보
어문학

주체107(2018)년
제64권 제4호

조건부확률에 의한 본문정보측정의 한가지 방법

안 성 득

위대한 령도자 **김정일**동지께서는 다음과 같이 교시하시였다.

《새로운 과학기술분야를 개척하기 위한 사업도 전망성있게 밀고나가야 합니다. 나라의 과학기술을 세계적수준에 올려세우자면 발전된 과학기술을 받아들이는것과 함께 새로운 과학기술분야를 개척하고 그 성과를 인민경제에 적극 받아들여야 합니다.》(《**김정일**선집》 증보판 제11권 138～139페지)

선진과학기술을 적극 받아들이자면 과학기술정보사업을 활발히 벌려야 한다.

과학기술정보사업은 언어와 떼여놓고 생각할수 없다. 그것은 언어가 인간교제의 가장 중요한 수단이며 정보의 적재와 표현, 전달에서 기본요소로 된다는 사정과 관련된다. 결국 언어가 없이는 정보산업의 발전에 대하여 생각조차 할수 없다.

언어연구에 수학적인 원리와 방법들을 적용하는것은 나라의 정보산업을 빨리 발전시키고 인민경제의 모든 부문을 정보화하는데서 절실한 문제로 나선다.

언어학의 기본연구단위들중의 하나인 본문에 대한 연구에서도 이것은 례외로 되지 않는다.

본문에 대한 언어학적리해에서 지금까지의 견해들을 보면 크게 두가지로 가를수 있다.

하나는 본문이 언어적분석과 처리를 위한 일종의 언어자료라는것이며 다른 하나는 정보교환을 위한 언어적단위라는것이다.

첫째 견해의 의미는 본문이 언어학의 연구대상이 아니라 분석과 처리를 위한 한갖 언어자료라는것이며 둘째 견해는 본문이 언어학의 연구대상으로서 그것은 언어적단위들가운데서 가장 큰 확대된 단위라는것이다. 다시말하여 전통언어학에서 언어적단위를 어음, 형태부, 단어, 단어결합, 문장으로 보던것을 확대하여 본문까지 본다는것이다.

본문의 기본표징은 첫째로, 언어규범에 의하여 기록된 완결된 형식과 론리적체계에 따라 서술된 방향성이 있는 완결된 사상을 가져야 한다는것이다.

둘째로, 사상적내용을 보존하면서 형식과 형태를 여러가지 목적과 필요에 따라 변경시킬수 있어야 한다는것이다.

셋째로, 언어적형식과 사상적내용이 서로 엉켜져있어야 한다는것이다.

본문은 문구와 문단으로 이루어진다.

문구란 본문을 구성하는 가장 작은 글토막이다. 많은 경우 문구는 1개의 문장으로 이루어진다.

문단은 문구를 기본줄거리로 하여 본문을 론리적으로 구획지을수 있는 단락이다. 문단은 1개이상의 문장들로 이루어진다.

본문을 구성하는 문구와 문단은 상대적이다. 다시말하여 하나의 문구, 하나의 문단으로 이루어진 본문도 있을수 있다. 그것은 본문의 형식과 형태를 자유로이 변경시킬수 있기때문이다.

여기로부터 본문의 크기에 대한 개념을 설정할수 있다. 본문의 크기가 가장 작은것은 발자크와 편집원사이에 오고간 편지의 내용인 《?》, 《!》을 들수 있고 본문의 크기가 가장 큰것은 수학자 오일레르의 전집을 들수 있다.

본문은 언어정보론적측면에서 기본단위일뿐아니라 최소단위로 된다. 그것은 콤퓨터에 의한 정보의 분석과 처리를 원만히 실현하자면 1~2개의 본문만을 대상으로 하여서는 안되며 이미 출판된 본문들을 모두 고려해야 하는 사정과 관련된다.

본문만이 언어자료처리를 원만하게 할수 있는 정보적내용을 가진다.

어떤 사람이 이야기를 하거나(입말) 글을 썼을 때(글말) 이것이 얼마만한 정보를 가지며 어떤 통로를 거쳐 어떻게 전달되는가를 연구하는것이 언어정보론의 연구대상이다.

일반적으로 정보란 자료를 통하여 얻는 지식의 총체, 다시말하여 자료를 통하여 이전에 알지 못했던 사실을 새롭게 알게 되거나 어떤 사실을 명백하게 알게 될 때 그 자료에 포함된 내용을 말한다.

정보를 전달하는 수단을 통로, 정보를 발생하는 사람이나 물질적대상을 정보원천이라고 한다.

언어가 통신적기능 즉 정보전달적기능을 수행하는것만큼 언어를 정보와의 련관속에서 연구하는것은 언어의 특성으로부터 필연적으로 제기된다고 볼수 있다.

언어학에서의 정보는 이야기하는 사람(화자)에 의하여 이야기듣는 사람(청자)에게 전달되는 소식의 내용을 말한다.

정보론적견지에서 보면 언어란 한 민족성원들이 각이한 사회생활분야에서 정보전달에 공통적으로 리용하는 사회적으로 약속된 기호의 체계라고 말할수 있다. 따라서 언어적형식은 정보의 형식이고 내용도 정보의 내용으로 볼수 있다.

언어정보론에 대한 인식을 바로가지자면 우선 정보의 개념부터 잘 알아야 한다.

원래 정보라는 말은 군사, 정탐분야에서 정황에 대한 소식이나 통보자료의 뜻으로 쓰이기 시작하였지만 오늘에 와서는 과학기술, 경제를 비롯한 사회생활의 거의 모든 분야에서 쓰이고있다.

그러나 정보가 모든 분야에서 다 동일한 의미로 쓰이는것은 아니다.

조종학에서는 정보를 자동조종체계가 외부세계와 수감장치를 통하여 진행하는 일체 교환내용이라고 본다.

통보론에서는 정보란 불확정적인것이 새롭게 확정되는 내용이라고 주장한다.

이외에도 정보는 새로 알게 되는 지식, 결심채택에 필요한 지식, 일정한 시간에 일정한 목적을 가지고 전달되는 사용가치가 있는 지식, 사람과 사람사이에서 전달되는 부호화된 지식 등 여러가지 의미로 쓰이고있다.

정보개념의 이러한 정의들에는 정보의 속성이 단편적으로 반영되여있다.

이 모든 정의들을 론리적으로 종합분석해놓고보면 정보가 지식성, 유용성, 전달성을 공통적인 속성으로 하고있다는것을 알수 있다.

이러한 판단에 기초하여 정보를 다음과 같이 정식화할수 있다.

정보란 다른 대상으로부터 새로 알게 되는 쓸모있는 지식을 말한다.

정보는 우선 지식이다. 그렇다고 하여 모든 지식이 다 정보로 되는것은 아니다. 정보는 다른 대상으로부터 새로 알게 되는 쓸모있는 지식이다.

정보는 또한 새로 알게 되는 지식이여야 한다.

아무리 다른 대상으로부터 알게 되는 쓸모있는 지식이라고 하여도 이미 알고있는 지식은 정보라고 말할수 없다.

정보는 다음과 같은 속성을 가진다.

정보의 속성으로서는 우선 지식성을 들수 있다.

정보가 지식성을 가진다는것은 정보가 다름아닌 지식이라는것을 의미한다.

지식은 인식의 결과에 이루어진다. 인식을 감성적인식과 론리적인식으로 가를수 있는것만큼 지식도 감성적인식결과로 이루어진 지식과 론리적인식결과로 이루어진 지식으로 나누어볼수 있다.

전자는 직접 실물을 통하여 그의 외적인 상태 혹은 정황을 반영하는 자료적인 지식이라면 후자는 개념을 통하여 실물의 내적인 본성과 합법칙성을 반영하는 리론적인 지식이다.

원래 정보의 개념은 신호, 소식, 자료 등을 내용으로 하는 전자만을 외연으로 하고있었지만 그것이 널리 쓰임에 따라 후자까지도 외연에 포함시키면서 더욱 확장되였다.

정보의 속성으로서는 또한 유용성을 들수 있다.

정보가 유용성을 가진다는것은 정보가 쓸모있는 지식이라는것이다.

지식의 쓸모는 그것이 자주성을 위한 인간의 요구실현에 얼마나 리로운가 하는데 따라 결정된다. 지식의 쓸모는 상대적이며 가변적이다.

지식은 원래 쓸모를 가지며 또 그에 의하여 이루어지지만 모든 사람들에게 그 쓸모가 절대적으로 동일한것도 아니며 영원히 고정불변한것도 아니다. 개별적인 사람들은 자기가 처한 구체적인 환경에 따라 각이한 요구를 제기하며 그것은 시간의 흐름과 함께 부단히 변화되는것만큼 지식의 쓸모는 개인별로, 시기별로 각이하게 결정된다.

정보의 속성으로서는 또한 전달성을 들수 있다.

정보가 전달성을 가진다는것은 정보가 다른 대상으로부터 알게 되는 지식이라는것을 의미한다.

아무리 새로 알게 되는 쓸모있는 지식이라고 하여도 순수 자체의 론리적사유를 통하여 자기스스로 알게 되는 지식은 정보라고 말할수 없다.

다른 대상으로부터 전달되여 알게 되는 지식만이 정보로 될수 있다.

그러한 대상으로서는 지식을 이미 알고있는 사람, 지식을 기록한 교과서나 참고서, 론문, 특허문헌 등 물질적인 대상을 들수 있다.

일반지식과는 달리 정보에 대해서는 그에 대한 수요와 그의 원천문제가 제기된다.

정보수요란 정보를 알려는 요구의 크기를 말하며 정보원천이란 정보를 체현하고있는 대상을 말한다.

정보를 각이한 기준에 따라 여러가지로 나누어볼수 있다.

정보는 그것이 반영하고있는 대상에 따라 정치정보, 군사정보, 경제정보, 대외정보, 과학기술정보, 체육정보 등 사회분야별로 가를수 있으며 그것을 더 세분화할수 있다.

실례로 과학기술정보는 정보의 과학분류에 따라 사회과학정보, 자연과학정보, 기술공학정보 등으로 가를수 있으며 이것을 다시 세분화할수 있다.

정보는 그 원천에 따라 실물정보, 구두정보, 문헌정보로 가를수 있다.

실물정보는 실물에 체현되여있는 정보이고 구두정보는 정보소유자의 정보이며 문헌정

보는 글말로 기록되여있는 정보이다.

정보는 표현형태에 따라서 수량정보와 생체정보, 의미론적정보로 갈라볼수 있다.

이외에도 정보는 시간적순서에 따라 구정보와 신정보로; 정보의 량의 특성에 따라 확실한 정보와 모호정보로; 정보의 전달방식에 따라 문자정보와 장면정보로; 정보의 주요성 정도에 따라 주요정보, 차요정보, 잉여정보로; 정보의 표현방식에 따라 론리적정보, 형상적정보, 예술적정보로 나누어볼수 있다

본문이 가지는 정보는 사실정보와 평가정보로 구성된다.

사실정보는 본문의 정보적내용에서 사실, 사건, 시간, 장소, 원인의 주체와 객체에 대한 부분을 말한다.

평가정보는 정보적내용에서 사실정보에 대한 본문작성자의 견해, 평가, 측정, 예측 등에 대한 부분을 말한다.

만일 사실정보부분이 없거나 적고 평가정보부분이 확대된다면 그 본문은 정보적자료로서의 가치 즉 정보전달의 견지에서의 자료로서는 가치가 적다. 그렇다고 하여 문체론적 특성, 언어환경의 요구를 무시하고 사실정보를 자료적으로 소개하는 본문은 메마르고 딱딱하며 표현적효과가 낮아진다.

정보를 량적으로 측정하려면 다음의 두가지 출발개념에 의거해야 한다.

우선 우연적인 언어사건들의 출현확률이 얼마인가를 알아야 하며 또한 시행전에 내포하고있는 언어적사건의 엔트로피(불확정성)가 얼마인가를 미리 규정하여야 한다.

우리는 이미전에 언어적사건들의 출현확률이 같은 경우 단어와 본문의 정보를 량적으로 평가하였다. 그러나 그것은 등확률이라는 강한 조건하에서 론의된것으로 하여 정확한 정보량을 구할수 없었으며 단지 본문의 정보가 어떻게 분포될것이라는 예측을 모형화했을뿐이다.

2개의 사건이 등확률일 때와 등확률이 아닌 경우에 어느 경우의 엔트로피가 더 크겠는가 하는 문제가 제기된다. 통보론에서 증명한데 의하면 등확률인 경우가 등확률이 아닌 경우에 비하여 많은 엔트로피를 가진다.

본문의 엔트로피와 정보를 계산할 때 매 사건들의 출현이 등확률이라는 가정은 성립하지 않는다.

그것은 언어규범이 매우 다양하고 본문의 주제가 다르며 저자마다 언어적요소들을 각이한 확률을 가지고 리용하기때문이다.

현실생활에서 볼수 있는것처럼 응당 일어날 사건이 일어났거나 누구나 다 알고있는 사실을 전달할 때에 거기에 들어있는 정보량은 큰것이 못된다.

그러나 일어날수 없는 일이 일어나거나 아직 그 누구도 짐작할수 없는 내용을 알게 됨으로써 얻게 되는 지식의 량은 훨씬 큰것이다.

이런 경우에 정보량에 대한 평가는 사건이 일어난 후 알게 되여 얻어지는 지식의 량으로 계산하는것이 아니라 사건이 일어나기 전에 그 사건에 대하여 알지 못하는 정도 즉 엔트로피의 크기로써 계산한다. 즉 만일 시행후에 출현의 가능한 상태개수가 s이고 등확률이라면 그 매개 사건의 정보량은 $I=-\log(1/s)$로 표시된다.

만일 매개 사건의 출현확률이 서로 다르다면 i째 사건의 정보량 I_i는 다음과 같이 표시된다.

$$I_i = -\log p_i$$

여기서 P_i는 i째 사건의 출현확률이다.

여러번의 시행을 진행하면 i째 사건의 평균정보량은 $\hat{I}_i = -p_i \log p_i$ 이다.

결국 총 정보량은 다음의 식으로 표시된다.

$$I = -\sum p_i \log p_i$$

여기서 s는 시행후 사건출현의 가능한 상태개수이다.

결국 언어적시행에서 시행결과들이 확률 p 또는 그의 빈도수 f를 아는 경우에 확률—통계적방법에 의하여 정보량을 측정할수 있다.

여기에 기초하여 본문에서 매 자모(문자)들이 가지고있는 정보량을 계산하게 된다. 즉 본문에서 쓰인 조선어자모들이 어떤 빈도수를 가지며 그 순위는 어떠한가, 매 자모들의 사용빈도수가 알려진 조건에서 그의 정보량이 얼마인가를 측정할수 있다.

그런데 문제점으로 되는것은 비록 같은 자모일지라도 그 자모가 단어의 처음위치에서 쓰이였는가(초성) 중간위치에서 쓰이였는가(중성) 아니면 마감위치에서 쓰이였는가(종성)에 따라서 그 출현확률이 달라지는데 있다.

조선어에서 초성과 종성위치에는 반드시 자음자가 오며 중성위치에는 꼭 모음자만이 온다.

현대조선어본문에서 자모글자의 확률분포를 보면《ㄴ》은 그 순위에서 2번째 자리를 차지하지만 초성위치에서의 확률분포는 그 순위가 8로 떨어진다.《ㄹ》도 그 순위에서 4번째이지만 초성위치에서는 14번째로 떨어진다.

이로부터 언어적시행은 무조건확률에 의해서가 아니라 해당 본문구역의 문맥적환경에 의하여 결정되는 조건부확률에 의하여 특징지어진다.

사실 자모나 단어들, 기타 언어적단위들은 언어행위속에서 문맥상제한을 받기때문에 위치에 따라 출현확률이 각이하게 된다.

우리 말 본문의 한 글자를 결정하는 시행 α_1의 확률 $H_1 = H(\alpha_1)$을 계산할 때 모든 글자들은 서로 련관되여있다. 가령 우리 말 음절자의 구성을 놓고보더라도 초성에는 자음자, 중성에는 모음자 그리고 받침글자인 경우에는 종성위치에 자음자가 오게 되므로 사실상 어떤 임의의 위치에 어떤 글자가 오겠는가에 대하여 비교적 쉽게 결정할수가 있다.

이와 같이 글자들의 호상련관을 고려하는 경우 본문에서 한개의 글자가 가지는 엔트로피는 보다 더 감소하게 된다.

이 감소를 량적으로 특징짓기 위하여서는 해당 본문에서 선행한 글자를 결정하는 시행 α_1의 결과가 알려졌다는 조건하에서 다음글자를 결정하는 시행 α_2의 조건부엔트로피 $H_2 = H\alpha_1\ (\alpha_2)$를 계산하여야 한다.

본문과 그것을 이루는 단어, 단어결합, 문장 등에 포함되여있는 의미정보를 량적으로 평가하기 위해서는 문장론적인 정보와 문맥적인 제약성에 의거해야 한다.

알려지지 않은 본문의 문자들을 추측할 때 두가지 류형의 결합을 보게 된다. 그 하나는 문자나 음절의 결합인 모양결합이며 다른 하나는 형태부, 단어들의 결합인 기호결합에 대한 일정한 제한조건으로부터 출발하여 본문의 가장 확률적인 연장성을 고려해야 한다. 즉 본문을 이루자면 어떤 문자나 단어, 문장뒤에 어떤 문자나 단어, 문장이 올 가능성이 가장 큰

가를 따져보고 그에 맞게 가설을 세우고 론증해야 한다.

실험결과들은 3~4번째 위치에서 문자나 음절들의 결합이 형태부나 단어로 될 가능성이 크다는것을 보여준다.

따라서 본문의 앞부분에서 얻어지는 정보가 문자들의 분포와 통계에 대한 량적평가를 내릴수 있게 한다면 본문의 앞부분과 멀리 떨어진 곳에서 얻는 문장론적인 정보는 의미정보로 된다.

개별적단어에 포함되여있는 의미정보를 구해보자.

본문이 다음과 같은 단어사슬에 의하여 이루어졌다고 하자.

$$\omega_1,\ \omega_2,\cdots,\omega_n$$

이때 단어 ω_1에 포함되여있는 정보량이 얼마인가를 평가하자.

이를 위해 무엇보다먼저 본문토막 $\omega_2,\ \omega_k$에 내포하고있는 정보를 계산한다.

본문토막 $\omega_2,\ \omega_k$에 단어 ω_1은 일정한 정보를 준다. 따라서 본문토막 $\omega_2,\ \omega_k$에서 얻어진 정보 $I(\omega_2,\ \omega_k)$는 앞단어 ω_1을 알고있다는 조건하에서 얻은 정보 $I((\omega_2,\ \omega_k)/\omega_1)$에 비하여 크다.

이때 본문토막 $\omega_2,\ \omega_k$가 가지고있는 정보량과 단어 ω_1을 알고있다는 조건하에서 본문토막 $\omega_2,\ \omega_k$가 가지고있는 정보량의 차는 단어 ω_1에 포함되여있는 의미정보이다. 즉 $I(\omega_1)=I(\omega_2,\ \omega_k)-I((\omega_2,\ \omega_k)/\omega_1)$가 본문토막 $\omega_2,\ \omega_k$의 엔트로피를 감소시키며 따라서 두번째 예측을 쉽게 해준다.

각이한 언어들속에 있는 서로 다른 단어들과 각이한 언어들속에서 한 의미로 쓰이는 단어에 대한 정보를 엄밀하게 평가하자면 평균의미정보의 개념을 도입하여야 한다.

평균의미정보는 여러개의 서로 다른 문맥들을 택하고 그에 대한 의미정보를 평가한 다음 그것을 택해진 문맥개수로 평가하는 방법으로 얻는다.

연구결과들은 분석적구조를 가진 언어(프랑스어, 벌가리아어)의 단어가 종합적구조를 가진 언어(굴절어인 로어, 교착어인 에스또니아어)의 단어들보다 적은 의미정보를 가진다는 것을 보여준다.

이것은 매개 언어의 단어들이 의미정보를 가지는데 여기서 교착어들이 보다 많은 의미정보를 가진다는것을 의미한다.

이 문제에 대하여서는 테마와 레마에 관한 리론에서 교착어인 조선어가 굴절어인 로어나 영어보다 많은 정보무게를 가지지만 굴절어는 교착어에 비하여 정보를 빨리 알수 있는 우점이 있다는 연구결과가 이미 발표되였다.

본문에서 어떤 언어적실험 A에 의하여 얻어지는 정보를 계산하자.

실험 A는 본문의 n째 구역에서 s개의 출현가능성을 가지고 진행되는데 이미 이 구역앞에 있는 언어적요소 b_{n-1}은 알고있다.

사슬 b_{n-1}은 첨수번호 i를 가지는 우연사건이다.

주목하는 본문의 n째 위치에서 이러저러한 요소들의 출현은 값 $j_h(1\leqq h\leqq s)$를 취하는 우연량이다.

매 i에 대하여 A_n이 j_h를 취할 조건부확률은 $p(j_{ih}/b_{n-1})$와 같다.

따라서 정보 I_n과 크기가 같은 평균조건부엔트로피는 사슬 b_{n-1}의 매 첨수에 해당한

확률무게를 가지며 모든 값 b_{n-1}에 관계되는 평균엔트로피로 얻어진다.

즉 $H = I = -\sum p(b_{i\ n-1})\sum p(j_{ih}/b_{n-1}) \times \log p(j_{ih}/b_{i\ n-1})$ 이다.

우의 식은 사슬 b_{n-1}이 알려졌다는 조건하에서 n째 구역에서 불확정성의 평균값이 얼마이며 언어적요소들을 선택할 때 그 정보량이 얼마인가를 보여준다.

그런데 평균조건부엔트로피가 본문의 n째 구역에서 요소들의 확률분포와 b_{n-1}의 출현확률에 관계되므로 이 사실을 고려하면 $H_n = I_n = H(j/b_{n-1}) = H(b_{i\ n}) - H(b_{i\ n-1})$이 성립한다.

이 식은 2차정보(1개의 앞선 문자가 알려진 경우), 3차정보(2개의 앞선 문자가 알려진 경우) 등을 아는 조건에서 다음문자가 가질 정보량이 얼마인가를 규정한다.

음운, 음절, 형태부 등에 대한 1차, 2차,…, n차 정보를 계산할 때 그 작업량은 매우 방대하다.

실례로 음절의 2차정보를 계산할 때 H_2의 계산은 다음과 같이 진행한다.

$$H_2 = H_{d_1}(d_2) = H(d_1, d_2) - H(d_1)$$

단어, 형태단어에 대한 평가방법은 단어결합에 대해서도, 문장에 대해서도, 더 나아가서 본문에 대해서도 그대로 적용할수 있다.

그런데 음운개수가 5～6개 혹은 그 이상일 때 이 음운들로 이루어지는 결합은 여러개의 서로 다른 단어들을 형성할수 있다.

따라서 그런 경우에도 조건부엔트로피를 계산하자면 막대한 계산작업이 필요한것이다.

본문의 각이한 구역에서 엔트로피, 정보량을 정확히 그리고 빨리 결정하기 위하여 새로운 예측방법을 적용한다.

이 방법은 문자들의 조건부확률을 주관적으로 평가하는 방법이다.

이 방법을 적용하려면 언어소유자에게 언어규범의 확률적특성량들에 대한 충분한 지식이 있어야 한다.

주관적으로 어떤 언어적단위가 다른 단위에 비하여 더 자주 또는 드물게 출현한다는 결론을 내리게 되는데 그 정확도가 사람마다 다르다.

이때 주관적판단의 확률은 다음의 식으로 표시한다.

$$P_i = \frac{n_i}{N}$$

여기서 N은 총 예측개수이고 n_i은 주어진 문자에 대한 예측개수이다.

실마리어 확률, 본문, 정보, 문장, 엔트로피