김일성종합대학학보(어문학)
주체106(2017)년 제63권 제1호

자연언어의 콤퓨터처리에서 나서는 몇가지 언어학적문제

안성득

위대한 수령 **김일성**동지께서는 다음과 같이 교시하시였다.

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

사회주의강국건설을 힘있게 다그쳐나가고있는 오늘 우리앞에는 지식경제시대의 요구에 맞게 최신과학기술분야를 개척하며 인민경제를 현대화, 정보화하여야 할 절박한 과업이 나서고있다.

지식경제시대의 요구에 맞게 인민경제의 모든 부문을 정보화하기 위하여서는 콤퓨터에 의한 언어처리문제가 원만히 해결되여야 한다.

콤퓨터언어학은 콤퓨터로 언어를 자동분석하고 언어자료들을 자동처리하며 서로 다른 두 언어를 번역하는데서 나서는 언어학적문제들을 연구한다.

콤퓨터언어학연구에서 기본은 자연언어에 대한 기계적분석방법을 확립하는것이며 또한 최신과학기술의 성과에 토대하여 언어연구에 수학, 론리학, 통보공학, 기호학을 비롯한 다른 과학분야들에서 쓰이는 합리적인 분석수법들을 받아들이는것이다.

지난 시기에는 사람－콤퓨터－사람사이의 정보전달과 조종체계에서 주로 인공언어, 기계언어가 리용되여왔다.

그러나 현재 콤퓨터의 화면인식장치, 음성인식장치의 개발 등 콤퓨터기술이 급속히 발전함에 따라 프로그람언어가 아닌 자연언어가 직접 정보교환 및 조종체계에서 쓰이고있다.

자연언어의 콤퓨터처리를 위하여서는 무엇보다먼저 자연언어자료들에 대한 계량적조사와 분석을 하여야 한다.

자연언어자료란 임의의 언어로 기록된 정보를 가진 본문 다시말하여 글자와 음성언어로 기록된 정보자료를 말한다.

언어학에서는 음운, 형태부, 단어, 문장, 본문 등이 가지고있는 언어정보들을 수량적으로 정확히 규정함으로써 자연언어의 콤퓨터처리를 위한 과학적토대를 마련하여야 한다.

일반적으로 량이란 셀수 있거나 잴수 있는 대상이나 현상을 수로 표현한것을 말한다. 원래 수에 대한 개념은 어떤 대상, 현상을 연구함에 있어서 그것을 량적으로 측정하여야 할 필요성으로부터 나왔다. 량의 가장 중요한 특성은 그것이 척도단위로 주어진 어떤 량과 비교될수 있다는것이다.

연구하려는 량과 척도단위로 되는 량과의 비교과정을 측정이라고 하며 이 측정결과가 바로 수로 된다.

수량으로 나타나는 대상, 현상들은 무게, 길이, 부피, 세기 등으로 표현되며 단위화되거나 정량화할수 있는것들은 개수, 회수, 특성값 등으로 표현된다.

언어에는 말소리와 같이 크기, 세기를 가지거나 음절, 형태부, 단어, 문장 등과 같이 단

위화된것 즉 개수로 표현되는것도 있으며 단어의 의미구조와 결합관계, 문장성분, 문법규범 등과 같이 일정한 규칙을 가지는 특성들도 있다. 그러므로 문자, 음절, 토, 단어, 단어결합들의 개수, 정보무게와 그 사용정도, 형태단어의 길이, 단어, 문장, 본문이 가지는 의미정보를 수값으로 정확히 계산하고 규칙들을 종합적으로 조사분석하여 얻어진 수자와 모형들을 콤퓨터에 미리 넣어주어야 한다.

자연언어의 콤퓨터처리를 위하여서는 다음으로 자연언어를 형식화하여야 한다.

자연언어를 형식화한다는것은 언어기호를 콤퓨터가 인식하도록 만든다는것이다. 다시 말하여 수학이나 물리, 화학을 비롯한 자연과학부문들에서 만들어 쓰는 부호, 도식, 공식 등을 도입하여 언어의 량적, 질적특성들을 반영한다는것이다.

언어기호는 약속된것이며 그것은 자의적이고 다의성을 가진다.

콤퓨터에 의한 자연언어처리에서 자연언어가 가지고있는 다의성을 극복하고 필요한 정보만을 간단명료하게 표현하며 전달하기 위하여서는 그것을 형식화하여야 한다.

자연언어의 형식화를 실현할수 있는것은 자연언어가 수학언어처럼 정보전달의 기호체계로 되기때문이다.

언어기호를 콤퓨터처리와 수학적모형화의 견지에서만 본다면 그때의 언어기호는 순수기호로 된다.

기호의 형성은 의미되는것(대상)을 반영하는 련계의 심리적과정이라고 말할수 있다. 기호는 형식과 내용의 두 측면으로 이루어지는데 이 둘사이에는 아무런 련관관계도 없다.

언어에서 단어의 어음과 의미, 단어와 대상사이의 호상관계 역시 기호론에 의하여 설명된다. 즉 이름은 대상의 본질적속성을 반영하지 않는다.

이러한 련계의 설정으로 하여 사람들은 외부세계의 어떤 대상을 연구할 때 자신을 위하여(때로는 콤퓨터, 로보트용으로) 이미 잘 알려진 기호들을 리용한다.

기호란 어떤 대상이나 현상, 관계를 대신하여 나타내는 미리 약속된 물리적표식이다. 기호는 문자에 의하여 표시될수도 있고 빛이나 음파, 전자기파에 의하여 표시될수도 있다.

기호는 어떤 체계속에서만 약속된 의미를 나타낸다. 실례로 기호《ㅇ》는 조선어자모체계에서《이응》, 영어나 로어글자체계에서《오우》, 아라비아수자체계에서는《령》, 화학에서는 산소원자를 의미한다.

언어에서 어음과 의미의 호상관계를 표시하는것(기호)과 표시되는것(대상)과의 관계로 볼수 있다.

언어가 기호적성질을 가지고있는것만큼 언어를 기호로 전환시킬수 있다. 일찌기 언어를 전신부호로 전환시켰고 오늘에 와서는 그것을 기호화하여 콤퓨터를 비롯한 현대적인 기술수단들에 입력시키고있다.

론리학에서는《표시하는것 A》와《표시되는것 B》로 이루어진 전일체 AB에서 A를 기호라고 한다. 이때 A(기호)는 표시하는것(형식)이고 B(대상)는 표시되는것(내용)으로 된다.

기호는 표시하는것과 표시되는것사이의 자의성을 본질적속성으로 한다.

기호가 자의적이라고 하는것은 대상과 의미사이에 아무런 련관관계도 없고 다만 일정한 대상, 현상을 가리킬뿐이라는것이다.

실례로 교통신호로서의《푸른색》이《통과하라》라는 의미를 나타내는것은 조건적인 약

속에 의한것이지 《푸른색》속에 《통과하라》는 속성이 이미 주어져있거나 《통과하라》는 의미가 《푸른색》에 의해 미리 주어진것은 아니다.

언어가 기호성을 가진다고 하여 곧 언어가 기호로 되거나 기호의 일종으로 되는것은 아니다. 다시말하여 언어의 기호성에 대하여 말할 때 언어가 그 본성으로 보아 기호와 동일시될수 없으며 따라서 기호학의 테두리속에서 언어학을 다룰수 있다고 간주하여서는 절대로 안된다.

그러나 언어와 기호가 본질상 같은것이 아니기때문에 언어현상에 대한 기호학적, 기호론리학적연구를 할수 없다고 볼수는 없다. 그것은 언어와 기호가 비록 본질상 같은것은 아니지만 언어에 기호적인 측면, 요소가 있기때문이다.

바로 이 점을 언어의 기호성이라고 한다.

언어의 기호성을 비록 언어의 본질적속성으로 간주할수는 없어도 언어에 기호적측면이 있다는 점을 무시하여서는 안된다.

그것은 우선 언어현상을 기호의 체계로서 정연하게 분석할수 있다는데서 뚜렷이 찾아볼수 있다.

언어는 정보를 가진 기호의 체계로서 분석되며 나아가서 정보를 지닌 단위들은 다시 기호성을 띤 최소성분들로 분석된다.

그것은 또한 언어의 매개 단위가 《의미하는것》과 《의미되는것》의 두 요소들의 자의적련관으로 이루어지는 성질을 가지는데서 찾아볼수 있다.

실례로 6개의 자모로 된 유한자모렬 《청진》은 조선의 북동부에 위치하고있는 함경북도의 도소재지를 의미한다. 이때 신호사슬 《청진》은 언어기호적으로 또는 음성적으로 남게 된다.

자연언어를 콤퓨터나 로보트가 리해하도록 하기 위하여서는 언어구조와 언어체계를 모형화하고 기호화하여야 한다. 언어를 모형화하고 기호화한다고 하여 언어를 순수한 기호체계로 보거나 언어의 사회적속성을 부인해서는 안된다. 자연언어를 기호화하는것은 다만 0과 1밖에 리해하지 못하는 콤퓨터의 특성과 관련된다.

자연언어의 형식화는 매개 민족어들의 일반적인 특성들을 과학적으로 깊이 조사하고 언어적기호들의 다의성을 단의적인것으로 전환시켜야 가능하게 된다.

지난 시기에는 언어의 형식화에서 주로 의미는 제외시키고 그 구조적인 측면만을 취급하였다.

언어는 물론 구조로 이루어져있다. 그러나 의미를 떠나서는 그 구조를 정확히 밝힐수 없다.

실지 기계번역을 위한 분석과 합성을 위하여서도 순수 구조만 가지고서는 문장구조를 옳게 분석하고 합성할수 없으므로 의미정보에 의한 의미론적분석과 의미구조를 형식화, 모형화하여야 한다.

자연언어의 콤퓨터처리를 위하여서는 다음으로 언어의 모호성을 처리하여야 한다.

언어의 모호성이란 언어적대상이나 현상들의 의미관계를 정확히 규정할수 없는 복잡성과 밀접히 련관된 성질을 말한다.

실례로 《젊은이들이 일을 많이 하였다.》, 《키가 큰 사람이 소년과 함께 아침달리기를 하고있다.》라는 문장에서 《젊은이, 아침, 많다, 크다, 일, 소년》은 모두 강한 호소성을 띠는 단

어들이다. 그리고 《소설책을 사겠니?》라는 물음에 대한 《후에 보자.》라는 대답은 소설책을 사겠다거나 사지 않겠다는 명백한 립장이 아니라 당분간 사지 않고 앞으로 두고보겠다는 뜻을 나타내는 모호성을 가진 문장이다.

언어의 모호성은 언어환경과 교제하는 사람들의 언어능력(표현능력, 인식능력)에 따라 커지기도 하고 작아지기도 한다.

언어의 모호성은 언어의 다의성의 일반화이다.

언어의 모호성에 대한 연구는 모호수학에 의하여서도 진행한다. 모호수학은 사물현상의 모호성과 그 반영으로서의 모호개념을 연구하는 학문이다.

보통모임을 A, 모호모임을 □이라고 하면 주목하는 원소를 x라고 할 때 $x\in A$이면 x의 소속도는 A(x)=1, $x\square A$이면 x의 소속도는 A(x)=0이다.

그러나 x의 □에로의 소속도 □(x)는 0과 1사이의 모든 실수값을 다 취할수 있다.

즉 $0\leq\square(x)\leq 1$이다.

만일 □(x)가 0과 1값만을 취한다면 모호모임 □는 보통모임 A로 되며 바로 이때 언어의 모호성은 언어의 다의성과 같아진다.

자연언어의 콤퓨터처리를 위하여서는 이밖에도 언어적법칙과 규칙들을 콤퓨터가 알고 처리할수 있도록 지식기지를 조성하는 문제도 심화시켜야 한다.

이러한 문제들에 대답을 주는것이 바로 콤퓨터언어학앞에 나선 주요과제이다.

언어의 콤퓨터처리를 위하여서는 이외에도 언어학적으로나 기술공학적으로 많은 문제들이 제기될것이라고 본다.

우리는 과학기술강국건설에 박차를 가하여 짧은 기간에 나라의 과학기술발전에서 새로운 비약을 이룩할데 대한 경애하는 최고령도자 **김정은**동지의 말씀을 가슴깊이 새기고 주체적인 응용언어학을 건설하는데서 나서는 어렵고 복잡한 문제들을 하나하나 풀어나가야 할것이다.