김일성종합대학학보
어문학

주체107(2018)년
제64권 제3호

본문의 언어학적특성에 기초한 영어본문구조모형화

장 충 혁

경애하는 최고령도자 **김정은**동지께서는 다음과 같이 말씀하시였다.

《과학기술의 종합적발전추세와 사회경제발전의 요구에 맞게 새로운 경계과학을 개척하고 발전시키는데 큰 힘을 넣어야 합니다.》(《조선로동당 제7차대회에서 한 중앙위원회사업총화보고》 단행본 40페지)

정보기술이 급속히 발전하고있는 오늘의 현실은 콤퓨터에 의한 자연언어처리에 대한 연구를 심화시켜 언어학리론을 더욱 풍부화하고 여러가지 언어처리체계들을 더 높은 수준에서 개발할것을 요구하고있다.

지난 시기 자연언어처리에서는 문장을 주되는 연구대상으로 하여 문장의 구조와 의미를 해석하기 위한 연구를 여러 측면에서 심도있게 진행하여 일정한 성과를 거두었다. 대표적으로 현재 국내에서 개발되여 널리 리용되고있는 영조기계번역체계 《룡남산》을 비롯한 기계번역체계들을 들수 있다.

그러나 인간의 사고능력을 모방한 자연언어처리기술을 도입하여 언어처리체계들의 정확성을 높이기 위하여서는 문장이상의 언어단위 즉 본문의 구조와 특성, 정보적내용에 대한 분석이 필수적이다.

론문에서는 본문의 중요한 언어학적특성들인 구조문법적통일성에 기초한 본문토막화방법과 론리의미적련관성에 기초한 본문구조모형화방법을 분석하고 각이한 자연언어처리체계개발에서 본문구조모형화실현을 위한 문제들을 론하려고 한다.

자연언어처리에서 본문이 가지고있는 정보내용을 리용하기 위하여서는 본문의 구조에 대한 모형화가 선행되여야 한다.

본문구조모형은 본문을 이루는 문장들사이의 구조적련관관계와 론리의미적관계를 반영하는 언어학적모형이다. 본문구조모형화는 본문을 이루는 문장을 정보내용에 따라 일정한 토막들로 분할하고 토막들사이의 의미적련관관계를 결정하여 전체 본문의 구조를 반영하는 모형을 얻는 과정으로 된다.

모형화에 의하여 얻어지는 본문토막들이나 그것들사이의 론리의미적련관관계는 대명사지시해석과 본문요약을 비롯한 자연언어처리의 여러 과제들에 효과적으로 리용될수 있다. 실례로 본문을 부분주제로 분할하는 경우 동일지시해석을 보다 쉽게 진행할수 있다. 그것은 주제가 일정하게 고정된 하나의 본문토막내에서 대명사들과 명사구들이 가리키는 대상들이 동일할 가능성이 높기때문이다.

례: Victoria, Chief Financial Officer of Megabucks Banking Corp since 2004, saw her pay jump 20%, as the 37-year-old also became the Denver-based financial-services company's president. It has been ten years since she came to Megabucks from rival Lotsabucks.

(2004년부터 메가박스은행업무회사의 재정부장으로 있는 빅토리아는 37살의

녀성으로서 덴버에 본부를 둔 투자봉사회사의 사장으로도 된것으로 하여 자기의 로임이 20% 늘어난것을 알게 되였다. 그 녀자가 경쟁자인 로쯔박스회사에서 메가박스로 온지도 10년이 되였다.)

우의 본문에서 밑줄을 그은 명사구들은 모두 Victoria라는 한 사람을 가리키고있다. 이 본문에 대한 구조모형화를 통하여 본문이 일정한 하나의 주제를 취급하는 본문토막이라는것을 알수 있다면 밑줄을 그은 명사구들의 지시해석에 도움이 될것이다. 그리고 본문구조모형을 통하여 본문을 이루는 문장들사이의 론리의미적련관관계를 추출해내면 본문의 기본정보를 담고있는 핵심적인 문장과 그것을 구체화하여 설명하는 부차적인 정보를 담고있는 문장을 구별하여 본문요약에도 리용할수 있다.

자연언어처리에서는 각이한 류형의 본문을 대상으로 여러가지 언어처리체계들이 개발리용된다. 기계번역체계나 본문요약체계와 같이 글말본문을 기본처리대상으로 하면서 본문의 주제적내용이나 형식적분류에는 크게 의존하지 않는 언어처리체계들도 있으며 음성인식체계와 같이 입말본문을 기본처리대상으로 하는 경우도 있다. 정보검색체계와 같이 글말본문과 입말본문을 다 처리대상으로 하면서 본문의 론리의미적련관관계보다도 일정한 부분주제를 가진 본문토막들로 분할하는 토막화과제를 중요한 처리과제로 내세우는 언어처리체계도 있다. 그러므로 본문의 류형에 관계없이 모든 본문에 보편적으로 존재하는 언어학적특성을 본문모형화의 기준으로 삼고 해당 언어처리체계의 특성에 맞는 모형화방법으로 구조모형을 얻어내는것이 합리적이다.

언어학적특성에 기초한 영어본문구조모형화에는 구조문법적통일성에 기초한 본문토막화와 론리의미적련관성에 기초한 구조모형화가 있다. 이러한 본문모형들은 본문이 가지고있는 언어학적특성에 기초하고있는것으로 하여 본문의 구조를 언어학적측면에서 분석하여 일반화한 대표적인 언어모형으로 된다.

무엇보다먼저 구조문법적통일성에 기초한 본문토막화방법에 대하여 보기로 한다.

본문토막화는 하나의 본문을 부분주제들로 분할하여 선형적인 렬을 얻어내는 과정이다. 본문토막화로는 본문의 구조에 대한 정교한 계층구조모형을 얻어낼수는 없지만 선형적인 구조를 얻을수 있다.

본문토막화의 목적은 주어진 본문을 여러개의 부분주제모임(subtopic group)으로 분할하는것이다. 전일적인 체계를 가진 하나의 본문을 부분주제들로 분할하는것은 본문토막들이 전체적인 본문구조모형화의 단위로 되며 이에 기초하여 보다 정교한 계층구조를 얻어낼수 있기때문에 콤퓨터에 의한 본문처리의 가장 기초적인 과제로 된다.

본문토막화는 본문의 주요한 특성의 하나인 구조문법적통일성에 기초하여 진행된다.

구조문법적통일성을 실현하는데 리용되는 어휘적수단들에는 우선 같은 단어의 반복리용이 있다.

례: Before winter I built a chimney, and shingled the sides of my house. … I have thus a tight shingled and plastered house.

(겨울이 오기 전에 나는 굴뚝을 세우고 집옆벽에 널판을 댔다. … 그래서 나에게는 단단한 널판을 대고 벽토를 바른 집이 생기게 되였다.)

구조문법적통일성을 실현하는데 리용되는 어휘적수단들에는 또한 상위어(hypernym)와 하위어(hyponym)의 련결수단이 있다.

례: Peel, core and slice the pears and the apples. Add the fruit to the skillet.
(배와 사과의 껍질을 벗기고 속을 파낸 다음 얇게 자르시오. 과일을 료리용남비에 넣으시오.)

우의 문장에서 pear와 apple은 fruit의 하위개념을 나타내는 단어들로서 두 문장을 구조문법적으로 련결시켜주는 수단으로 되고있다.

본문의 구조문법적통일성을 실현하는 문법적수단으로서는 지시어의 리용, 접속어의 리용, 대용어의 리용, 생략수법의 리용 등이 있다.

례: The Woodhouses were first in consequence there. All looked up to them.
(우드하우스팀은 결과적으로 거기에서 첫자리를 차지하였다. 모두가 그들을 찬양하였다.) (지시어의 리용)

They have what it takes to exist for millions of years. That is why they are still here today. (그것들은 수백만년동안 존재하는데 필요한것을 가지고있다. 그렇기때문에 오늘까지도 여기에 있는것이다.) (접속어의 리용)

They are closing this factory next month. Their plan is to build two new ones in the city. (그들은 다음달에 이 공장의 문을 닫을것이다. 그들의 계획은 도시에 두개의 새 공장을 세우는것이다.)(대용어의 리용)

구조문법적통일성에 기초한 본문토막화에는 우선 비감독본문토막화(Unsupervised Text Segmentation)가 있다.

구조문법적통일성에 기초한 비감독본문토막화는 하나의 부분주제에 속하는 문장들이나 단락들은 서로 구조문법적으로 통일되여있으며 다른 부분주제의 문장들이나 단락들과는 구조문법적통일성을 가지지 않는다는 특성에 기초한다. 그러므로 이웃하고있는 모든 문장들의 구조문법적통일성을 측정한다면 부분주제의 경계선에 《틈》이 생기게 되는것이다. 문장들사이의 구조문법적통일성을 평가하는데는 우에서 언급한 어휘적수단들이 특징으로 리용된다. 즉 문장을 이루는 단어들사이에 존재하는 반복출현, 상하위어관계, 반의어관계 등을 리용하여 구조문법적통일성을 평가한다.

이러한 원리에 기초한 본문토막화방법에서 대표적인것은 TextTiling알고리듬이다. 알고리듬은 세 단계 즉 형태부분할, 어휘적점수결정, 경계판정으로 이루어진다.

형태부분할단계에서는 공백을 경계로 하여 입력문의 단어들을 모두 소문자로 변환하고 기능어들로 이루어진 금지어목록에 있는 단어들을 제외시킨다. 다음 나머지 단어들에 대한 형태부해석을 진행하여 원형단어를 찾는다. 다음 원형단어들을 단어개수 w=20인 가상적인 문장들로 무리를 짓는다. 즉 실지 문장이 아니라 길이가 같은 가상적인 문장들로 분할한다.

이렇게 되면 가상적인 문장들사이에 틈이 생기게 되는데 그 량쪽에서 어휘적응집성점수를 계산한다. 응집성점수는 틈의 뒤에 있는 가상적인 문장의 단어들과 앞에 놓이는 가상적인 문장의 단어들의 평균류사도로 정의된다. 일반적으로 틈의 량쪽에서 10개의 단어로 이루어지는 가상문장토막을 리용한다. 류사도를 계산하기 위하여 틈의 앞에 있는 토막에서 단어벡토르 b를, 뒤에 있는 토막에서 벡토르 a를 취하는데 이 벡토르들은 길이 N(본문에서 기능어를 제외한 모든 단어들의 수)의 벡토르이며 단어벡토르의 i번째 요소는 단어 w_i의 빈도수이다. 다음 아래의 공식에 따라 류사도를 계산한다.

$$Sim_{\cos ine}(\vec{b},\vec{a}) = \frac{\vec{b}\cdot\vec{a}}{\left|\vec{b}\right|\left|\vec{a}\right|} = \frac{\sum_{i=1}^{N} b_i \times a_i}{\sqrt{\sum_{i=1}^{N} b_i^2}\sqrt{\sum_{i=1}^{N} a_i^2}}$$

가상적인 문장들사이에 있는 모든 틈 i에 대하여 i-k로부터 i까지의 가상적인 문장들과 i+1로부터 i+k+1까지의 가상적인 문장들사이의 류사도를 나타내는 류사도점수를 계산한다. 아래의 그림을 실례로 하여 류사도점수를 계산해보자.

아래의 그림은 4개의 가상적인 문장들을 도식으로 나타내고있다. 20개의 단어로 이루어진 가상적인 문장들은 여러개의 실지 문장들을 포함하고있을수도 있다. 그림에서는 매개의 가상적인 문장들이 2개의 실지 문장을 포함하고있는 경우 즉 k=2인 경우를 보여준다. 그림은 또한 이웃하고있는 가상문장들사이에서의 스칼라적의 계산을 보여준다. 그러므로 문장 1과 2로 이루어진 첫번째 가상문장에서 단어 A가 두번, B가 한번, C가 두번 출현했다고 하면 처음 두 가상문장사이의 스칼라적은 2×1+1×1+2×1+1×1+2×1=8로 된다. 출현하지 않은 모든 단어들의 빈도를 0이라고 하면 두 문장의 코시누스값은 0으로 된다.

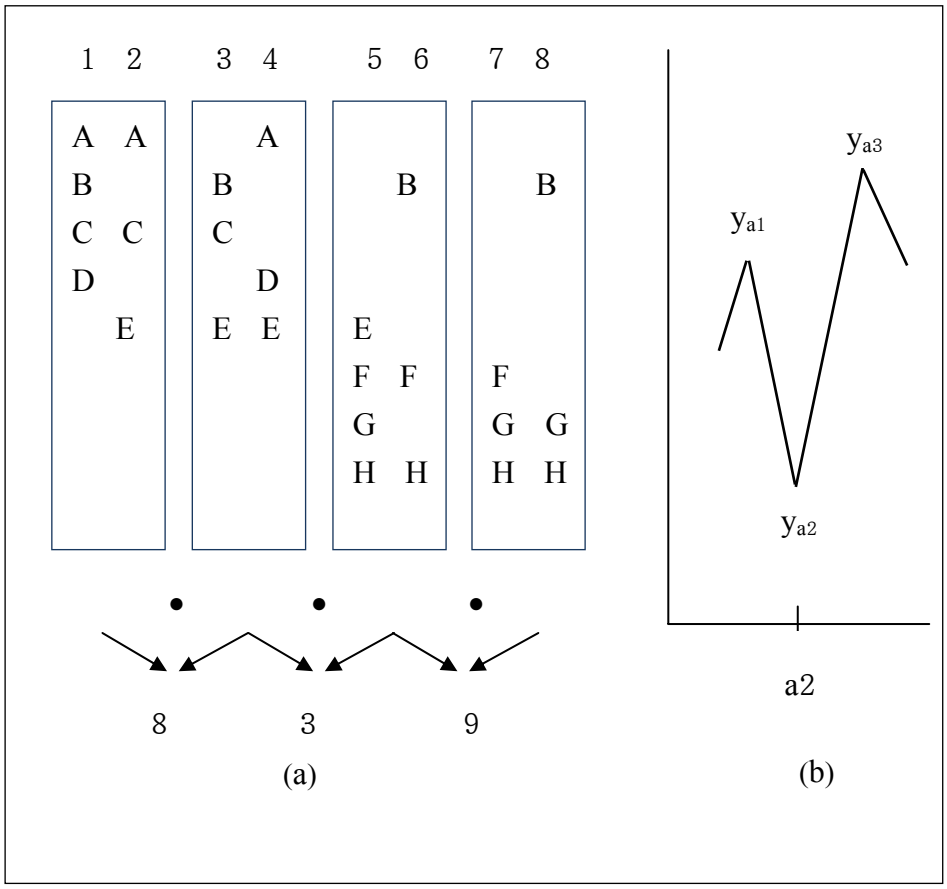

그림 1. TextTiling알고리듬

(a는 문장 1, 2와 문장 3, 4사이의 류사도의 스칼라적계산을 보여준다. A, B, C와 같

은 대문자들은 단어들의 출현을 보여준다. b는 골짜기의 깊이점수계산을 보여준다.)

다음 단계에서 틈에서의《류사도골짜기》의 깊이를 평가하여 깊이점수를 계산한다.

깊이점수는 골짜기의 량쪽 정점에서부터 바닥까지의 거리이다. 그림에서는 $(y_{a1}-y_{a2})+(y_{a3}-y_{a2})$로 된다. 깊이점수가 일정한 턱값보다 더 낮은 골짜기에 경계선을 할당하여 본문을 토막화한다.

구조문법적통일성에 기초한 본문토막화에는 또한 감독본문토막화(Supervised Text Segmentation)가 있다.

감독본문토막화는 본문토막경계가 표식되여있는 훈련코퍼스자료에서 담화표식어나 실마리어를 특징으로 리용하여 자동분류기를 훈련시켜 본문을 부분주제로 토막화하는 방법이다.

감독본문토막화에서는 우선 분류기가 특징으로 리용할수 있는 담화표식어나 실마리어들을 확정해야 한다.

글말에서 쓰이는 담화표식어들은 본문의 주제분야에 따라 서로 다르다. 즉 본문이 어떤 내용을 담고있는가에 따라 일반적으로 널리 쓰이는 담화표식어와 함께 특정한 단어나 표현들도 담화표식어로 쓰인다. 실례로 방송보도인 경우에 첫시작에 자주 쓰이는《Good evening, I'm <PERSON>.》과 같은 표현이나 일정한 토막의 시작에 쓰이는《Joining us now is <PERSON>》,《Correspondent <PERSON> has more…》와 같은 표현들은 방송보도토막화에서 쓰이는 중요한 담화표식어로 된다.

담화표식어나 실마리어들이 확정되면 그에 기초하여 분류기를 훈련코퍼스에서 실행한다. 수동으로 경계를 표시한 라지오나 텔레비죤방송자료들이나 기호 <p>로 단락이 표시된 본문자료들을 훈련코퍼스로 리용할수 있다. 감독토막화를 위한 분류기로는 지지벡토르기계(SVM)나 결정나무와 같은 2진분류기를 리용할수도 있고 숨은마르꼬브모형(HMM)과 같은 련속분류기를 리용할수도 있다.

다음으로 론리의미적련관성에 기초한 본문구조모형화에 대하여 보기로 한다.

구조문법적통일성에 기초한 문장구조모형화가 본문토막화에 귀착되며 결과적으로 본문의 선형구조모형을 얻어낸다면 론리의미적련관성에 기초한 모형화에서는 본문에 대한 보다 정교한 계층구조가 얻어진다. 그것은 본문을 이루는 개별적문장들사이의 론리의미적관계가 서로 련관되면서 보다 큰 본문단위를 이루며 종당에는 전체 본문의 전일적인 의미를 나타내기때문이다. 문장들사이에 존재할수 있는 론리의미적련관성은 매우 다양하지만 이러한 관계들을 대표적인 관계부류들로 일반화할수 있다. 아래에 그러한 관계들중 일부를 제시한다. 여기서 S_0과 S_1은 론리의미적련관성을 가지는 두 문장의 의미를 나타낸다.

결과(Result): S_0에 의하여 표현되는 상태나 사건이 S_1에 의하여 표현되는 상태나 사건을 일으키거나 일으킬수 있다는것을 나타낸다.

례: He was caught in the train. His joints rusted.

(그는 기차에 갇혀있었다. 그래서 관절이 상했다.)

설명(Explanation): S_1에 의하여 표현되는 상태나 사건이 S_0에 의하여 표현되는 상태나 사건을 일으키거나 일으킬수 있다는것을 나타낸다.

례: John hid Bill's car keys. He was drunk.

(죤은 빌의 차열쇠를 감추었다. 그는 취했었다.)

병렬(Parallel): S_0에 의하여 표현되는 내용에서 $p(a_1, a_2, \ldots)$가 S_1에 의하여 표현되는 내용의 $p(b_1, b_2, \ldots)$와 비슷하다는것을 나타낸다. 즉 모든 i에 대하여 a_i와 b_i는 비슷하다.

례: They wanted some brains. We wanted a heart.

(그들에게는 지혜가 필요했다. 우리에게는 용기가 필요했다.)

구체화(Elaboration): S_0과 S_1이 표현하는 내용에서 같은 전제 P를 나타낸다.

례: She likes fresh fruit. Her favourite fruit is apples.

(그 녀자는 신선한 과일을 좋아한다. 그가 좋아하는 과일은 사과이다.)

계기(Occasion): S_0에 의하여 표현된 상태가 S_1에 의하여 표현되는 상태로 변화되였거나 S_1에 의하여 표현되는 상태의 초기상태가 S_0에 의하여 표현된 상태라는것을 나타낸다.

례: She picked up the oil-can. She oiled the robot's joints.

(그 녀자는 기름통을 집어들었다. 그리고 로보트의 관절부위에 기름을 주었다.)

문장들사이에 맺어지는 이러한 론리의미적관계에 기초하여 전체 본문의 구조관계를 반영한 계층나무구조가 얻어진다. 즉 본문의 구조모형이 나무구조로 표현된다. 아래의 본문을 실례로 본문구조를 나무구조로 표현해보자.

John went to the bank to deposit his paycheck. (S_1)
He then took a train to Bill's car dealership. (S_2)
He needed to buy a car. (S_3)
The company he works for now isn't near any public transportation. (S_4)
He also wanted to talk to Bill about their football game. (S_5)
(죤은 자기의 로임을 예금하려고 은행에 갔다.
다음 그는 빌의 승용차판매소로 가는 기차를 탔다.
그는 승용차를 사려고 했다.
그가 일하는 회사가까이에는 공공운수수단이 없었다.
그는 또한 빌에게 자기들의 축구경기에 대해 이야기하고싶었다.)

실례본문을 읽어보면 본문의 구조가 선형구조가 아니라는것을 알수 있다. 이 본문은 문장 S_1과 S_2에 서술된 련속적인 사건에 대한것으로서 문장 S_3과 S_5는 S_2에 보다 직접적으로 련관되여있으며 S_4는 S_3에 직접적으로 련관되여있다. 문장들사이의 이러한 론리의미적관계를 아래의 그림과 같은 구조로 표현할수 있다.

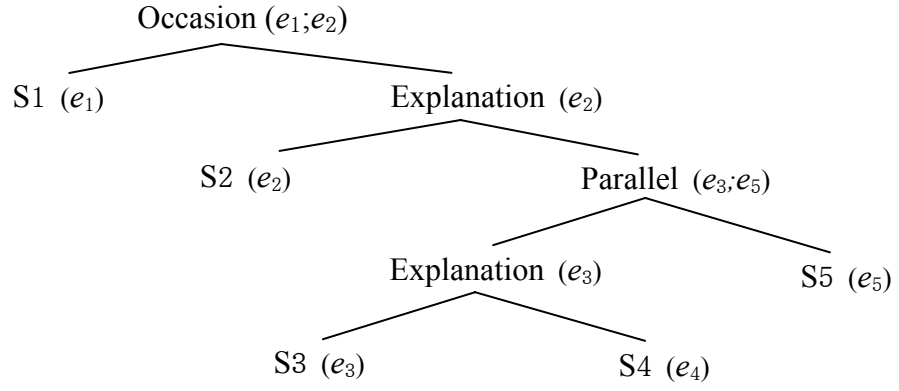

그림 2. 실례본문의 구조

우의 그림에서 매 마디는 부분적으로 맞물리는 절이나 문장모임 즉 본문토막을 나타낸다. 이 절의 앞부분에서 언급한 본문토막화에 의하여 얻어진 본문의 선형구조 즉 본문토막들이 계층구조를 이루는 구성요소들로 된다. 이러한 본문토막들은 문장구조모형에서 문장의 구성요소에 해당하는것으로 볼수 있다. 그러므로 구조문법적통일성에 기초한 본문구조모형화의 결과로 얻어지는 본문토막들은 론리의미적련관성에 기초한 본문구조모형화의 모형화단위로 된다.

우에서 언급한 본문구조모형화방법들은 자연언어처리체계개발에서 각이한 목적으로 리용될수 있다. 어떤 방법을 리용하여 본문의 구조모형화를 실현하겠는가는 언어처리체계의 목적과 특성에 따라 선택되며 하나의 언어처리체계에서도 해결하려는 과제에 따라 여러가지 구조모형들이 리용된다.

정보검색체계개발에서는 본문을 부분주제로 분할하는 본문토막화가 많이 리용된다. 정보검색에서는 사용자가 요구하는 정보내용을 전체 본문 또는 본문모임으로부터 탐색하여 검색어의 내용과 일치하거나 가장 류사한 본문 또는 본문부분을 결정하여야 한다. 그러므로 본문모임을 일정한 부분주제로 분할하여 해당 부분에 대한 주제를 확정하는것이 주요과제로 된다. 이 경우에는 TextTiling 알고리듬과 같은 본문구조모형화방법을 리용할수 있다.

많은 본문으로 이루어진 본문코퍼스를 일정한 주제의 본문모임으로 분류하거나 보도기사코퍼스에서 기사별로 내용을 분할하여 검색하기 위한 언어처리과제들을 수행하는 경우에도 마찬가지이다. 그러나 기계번역이나 본문요약과 같이 본문에 대한 보다 깊이있는 해석을 요구하는 언어처리체계개발에서는 선형토막화보다는 본문의 론리의미적련관성에 기초한 구조모형화가 기본으로 된다.

기계번역의 경우 대명사지시해석과 단어의미모호성해소를 비롯한 의미해석과제들을 수행하기 위해서는 우선 문장들사이에 존재하는 론리의미적련관관계를 확정하여야 하는데 이러한 문맥해석을 진행하기 위해서는 보다 높은 수준의 본문모형 즉 본문의 론리의미적구조모형을 리용하여야 한다. 문장들사이에 맺어지는 의미적련관관계를 고려한다면 대명사지시해석이나 단어의미모호성해소뿐아니라 번역문이 원문에 존재하는 론리의미적련관관계를 충분히 반영하고있는가를 평가하여 문장이상준위에서의 번역정확도도 판정할수 있다.

자동요약체계나 정보검색체계와 같이 본문을 대상으로 하는 자연언어처리체계들에 있어서도 론리의미적련관성에 기초한 본문구조모형화는 매우 중요하다.

본문의 언어학적특성에 기초한 본문구조모형화는 본문이 가지는 외적련결수단과 내적련결방식을 직접 표현하는 모형화방식이므로 표층구조분석에 크게 의존하는 자연언어처리에 효과적으로 쓰일수 있다.

우리는 본문의 특성에 대한 언어학적연구를 더욱 심화시키고 여러가지 본문모형들을 실현하여 자연언어처리에 리용함으로써 나라의 과학기술을 세계적수준에 올려세우는데 적극 이바지하여야 할것이다.

실마리어 본문구조모형, 구조문법적통일성, 론리의미적련관성