김일성종합대학학보
어문학

주체108(2019)년
제65권 제1호

자연언어처리를 위한 중국어단어가르기연구

박사 부교수 **박 명 철**

1. 서 론

위대한 령도자 **김정일**동지께서는 다음과 같이 교시하시였다.

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

세계에서 가장 우수한 언어로 평가받는 우리 조선어와 함께 영어와 중국어 기타 다른 언어들에 대한 연구를 심화시켜 능률적인 기계번역프로그람을 개발하는것은 우리의 과학기술을 최단기간내에 획기적으로 발전시키며 선진과학기술을 적극 받아들이기 위한 중요한 요구로 된다.

자연언어처리에 대한 연구는 사람들이 일상적으로 리용하는 언어를 콤퓨터로 하여금 정확히 인식하고 음성인식이나 기계번역과 같은 일정한 목적에 맞게 처리하기 위한 연구이다.

최근년간 콤퓨터기술의 급속한 발전과 방대한 량의 자료기지구축, 인터네트를 통한 세계적규모에서의 자료공유와 정보교환은 자연언어처리 특히 기계번역에 대한 연구를 심화시킬것을 요구하고있으며 여기에서도 영어나 중국어와 같이 그 사용인구가 많은 언어들에 대한 연구가 중요시되고있다. 특히 중국어를 입구어로 하든 출구어로 하든 중국어가 포함된 기계번역을 실현하는데서 제일먼저 제기되는 공정이 바로 중국어단어가르기이다.

중국어단어가르기는 각종 중국어정보처리 즉 음성처리, 단어빈도통계, 찾아보기, 초록생성, 정보검색, 문장해석 등 중국어를 대상으로 하는 자연언어처리연구에서 실천적으로 제기되는 가장 기초적인 공정이다. 분할표식이 없는 다시말하여 단어의 경계가 없는 중국어 글자렬에서 의미를 가진 단어들의 경계를 정확히 구분하는것이 바로 중국어단어가르기와 관련한 연구에서 해결하여야 할 몫이다.

사람은 중국어본문을 읽을 때 각종 언어지식을 바탕으로 단어들을 정확히 구별해낸다. 그러나 콤퓨터를 리용하여 중국어글자렬에 대하여 이러한 처리를 진행하려면 콤퓨터에 의한 중국어단어가르기를 진행할수 있는 연구가 선행되여야 한다.

일반적으로 자연언어처리의 견지에서는 단어를 의미를 가지는 최소처리단위로 본다.

론문에서는 중국어단어가르기연구의 중요성으로부터 다음과 같은 문제점들에 중심을 두고 해결하려고 한다.

중국어단어가르기에 대한 일반적리해 즉 중국어글자렬에서 어느 부분이 단어이고 어떻게 가르기단위를 설정하는가?

중국어단어가르기산법문제, 즉 실제의미에 부합되는 중국어단어의 경계를 만들기 위하

여 어떻게 단어의 가르기를 진행하는가?

중국어단어가르기에서 미등록어식별문제, 즉 사전에 없는 지명, 인명, 외국이름과 같은 미등록단어들을 어떻게 식별하겠는가?

중국어단어가르기에서의 모호성해소문제 즉 어떤 방법을 리용하여 가르기모호성을 해소할수 있는가?

론문에서는 이상의 문제들을 자연언어처리의 견지에서 언어학적으로 리론체계화하고 그 실천방도를 제기하려고 한다.

2. 본 론

무엇보다먼저 중국어단어가르기에 대한 일반적인 리해를 가지는것이 중요하다. 중국어단어가르기가 중국어언어리해와 정보처리의 기초라는데로부터 1980년대초부터 단어빈도통계를 시작점으로 하여 그에 대한 연구가 진행되였다. 중국어에서는 관습적으로 단어들사이에 공백이 없이 련달아 쓰기때문에 단어들의 경계에 대한 인식, 즉 어느것이 단어인지를 구별하기가 매우 어렵다. 그러므로 단어가르기규범의 형식화를 통하여 어떤 글자조합이 단어를 구성하는가를 확정하는것이 중요하다.

중국의 많은 연구단위들이 집체적으로 완성한《정보처리용 현대중국어단어가르기규범》(아래에서는 략칭《단어가르기규범》이라고 한다.)은 1992년 중국국가기술감독국으로부터 국가표준(GB13715)으로 인정되였다. 이 규범은 중국어정보처리의 규범화와 표준성을 실현하기 위한 목적으로 현대중국어의 단어가르기원칙을 규정하였다. 또한 연구가 진척됨에 따라 이 규범을 수정보충할수 있는 련관규정들도 제기하였다.

단어가르기규범은 콤퓨터처리의 요구를 만족시키며 동시에 언어학적으로도 단어에 관한 론쟁을 극복하고 단어가르기단위를 문장가르기의 기초단위로 규정하였다. 이 규범에서는 〈/〉기호를 단어가르기단위사이의 분할부호로 정하였다.

단어가르기규범이 규정하는 가르기원칙은 다음과 같다.

— 공백과 문장부호는 가르기단위의 분할표기이다.

— 结合紧密(결합긴밀), 使用稳定(사용안정)을 중국어문자렬가르기에서 가르기단위의 기본원칙으로 한다.

이것은 주관성이 매우 강한 론법이다. 비록 우의 원칙에 실제적으로 단어빈도 혹은 사용빈도에 관한 요인이 있어도 단어가르기규범에는 빈도를 나타내는 표기에 대해서는 언급하지 않았다. 사실상 중국어의 복잡성으로부터 많은 2자, 3자, 4자 또는 그 이상의 문자렬이 하나의 가르기단위를 만들수 있는 경우가 있을수 있다.

— 4자성구를 포함한 4자단어조를 일률적인 가르기단위로 한다. 5자 및 그 이상의 속담, 격언 등과 같이 가른 후 원래 조합의 의미를 잃게 되는 경우에는 가를 필요가 없으며 이것은 하나의 가르기단위이다. 만일 그렇지 않다면 가른다.

실례로 不管三七二十一(성구: 아무것도 아랑곳하지 않고)는 가르지 않으며 失败/是/成功/之/母(실패/이다/성공/의/어머니)는 가른다.

의미가 바뀐 단어나 단어덩어리는 일률적으로 하나의 가르기단위로 하며 그렇지 않은 경우에는 가른다.

실례로 《妇女/能/顶/半边天。(녀성들은 한쪽 수레바퀴를 맡는다.)》와 《半边/天/都/红了。(절반 하늘이 다 붉게 물들었다.)》를 들수 있다.

— 줄임말, 비한자부호렬, 음역외래어는 일률적으로 가르지 않고 하나의 가르기단위로 한다.

— 가르기단위에 儿음이 붙은것은 하나의 가르기단위로 한다.

실례로 花儿(꽃), 悄悄儿(살그머니) 등을 들수 있다.

— 가르기단위에는 또한 단어보다 작은 단위인 어휘소가 존재한다. 실례로 新華社/四月/五/日/讯(신화사/4월/5/일/소식)에서 讯자이다.

가르기단위를 결정하는데서 품사정보를 함께 고려하여야 하므로 단어가르기규범에서는 품사가르기에 따라 각종 단어의 가르기에 대하여 구체적으로 설명하고있다.

단어가르기규범에서는 전통적인 품사가르기를 리용하였으며 단어를 명사, 동사, 형용사, 대명사, 수사, 단위사, 부사, 전치사, 접속사, 조사, 감탄사, 상성사(소리본딴말) 등 12개 부류로 나눈다. 단어가르기규범에서의 《결합긴밀, 사용안정》원칙은 일정한 불확정성을 가진다. 례하면 두글자동사의 구성에는 지배적구조, 보충적구조, 종속적구조가 속하는데 이런 두 글자단어가 하나의 가르기단위로 되는가 하는것은 명확하지 않다. 아래의 실례에서 어떤 두 글자조합은 하나의 단어 혹은 가르기로 될수 있다.

念书	写/信 ｜ 写信	打球 ｜ 打/球	玩球
打倒	打垮 ｜ 打垮	打掉 ｜ 打掉	打/碎
胡说	乱说 ｜ 乱/说	重/说	

단어가르기규범에서는 자연언어가 복잡한 개방모임이라는데로부터 《하나가 아니면 둘이고 둘이 아니면 하나라는 서술방식을 쓰는것은 많은 경우에 통할수 없다.》고 하였다.

정량원칙은 사용할수 없고 일반적으로 잘 쓰이지 않는 특수한 실례를 드는것은 《응용에 리로운것이 못된다.》고 하였으며 이것이 바로 동일한 한자렬에 대하여 두가지 구분형식이 존재하는 원인이라고 보았다.

단어가르기규범에 대해 실험을 진행한 후 어떤 구조들은 단어가르기규범에 따라 가르기경계를 판정하는것이 비교적 힘들다는것이 판명되였다. 여기에는 일부 뒤붙이의 구분, 지배적구조, 보충적구조, 종속적구조의 가르기 등 7가지 부류가 포함된다.

초기에 일부 학자들은 이 규범에 많은 부족점이 존재한다고 하였다. 일부 전문가들은 단어가르기규범이 부닥친 문제를 분석하고 그 난점이 가르기단위확정에 있다고 지적하였다. 그들은 그 원인이 우선 문법전문가와 일반사람들의 단어인식에서의 차이에 있고 또한 단어가르기규범자체에 모호성이 존재하며 많은 경우 표준성이 없다는데 있으며 각이한 응용이 가르기단위에 대한 각이한 인식과 처리를 조성하므로 중국어단어가르기규범의 완성된 가르기표준을 구축할수 없다는데 있다고 하면서 다른 수단으로 중국어자동가르기의 련관문제를 해결해야 한다고 주장하였다. 그러나 단어가르기규범은 실천적인 중국어단어가르기를 위한 연구를 진행함에 있어서 아주 좋은 출발점으로 된다고 볼수 있다.

사전을 리용하는 가르기방법에 대하여 말할 때 한자렬이 하나의 가르기단위로 될수 있는가 없는가 하는것은 우선 그것이 사전에 등록되였는가 아닌가 하는데 의거한다. 콤퓨터능력의 발전속도가 급속히 높아짐에 따라 현재 수십만단어로부터 수백만단어의 중

국어전자사전이 계속 개발되고있으며 이것은 중국어자동가르기를 포함하여 내적언어정보처리의 쓸모있는 자원으로 리용되고있다.

사전규모가 부단히 확대되면서 단어가르기규범의《결합긴밀, 사용안정》원칙과 일치하지 않는 단어들이 대량적으로 사전에 반영되게 된다.

사전의 단어수가 증가하면 한 측면으로 많은 단어가 더 작은 가르기단위로 갈라지는것으로 하여 사람들의 일반적인 언어사용관습과 맞지 않게 된다. 실례로 从上到下(우에서 아래로), 从左到右(왼쪽에서 오른쪽으로), 从头到脚(머리에서 발끝까지), 从南到北(남에서 북으로)를 들수 있는데 이러한 단어들은 일상언어생활에서 하나의 의미로 사용된다.

다른 한편으로는 자연언어처리를 위한 연구에 도움이 되기도 한다. 중국에서 개발된 중영기계번역체계만 보아도 중국어와 영어사이의 번역등가 및 각 언어의 관습적인 용법을 고려하여야 하기때문에 반드시 영어의 단어나 단어결합에 중국어의 여러개의 가르기단위가 대응하는 경우가 생기게 된다. 이러한 가르기단위의 대응은 규범에 맞는 단어는 아니지만 해당 기계번역체계에서의 처리를 위해서는 필요한 공정인것이다.

다음으로 중국어단어들의 실제적인 의미에 맞게 중국어단어의 경계를 구분하기 위한 중국어단어가르기산법문제에 대하여 보기로 한다.

중국어자동가르기산법은 각이한 기준에 근거하여 분류할수 있다. 우선 가르기사전을 리용하는가 리용하지 않는가에 따라 사전을 리용한 가르기와 사전을 리용하지 않는 가르기로 가를수 있다. 또한 가르기과정에 사용되는 지식자원에 따라 규칙에 기초한 방법과 통계에 기초한 방법 등으로 가를수 있다. 사전을 리용한 가르기는 중국어자동가르기의 주류로 되고있으며 이 가르기의 기본산법은 최대정합법이다. 최근에는 통계자료를 기본으로 하는 연구가 더 많아지고있으며 여기에 규칙에 기초한 방법을 서로 결합한 혼합방법이 제기되고있다.

규칙에 기초한 방법은 일반적으로 수동으로 작성한 가르기사전을 요구한다. 가르기를 할 때 가르기대상으로 되는 본문에 대하여 정합을 진행하고 가르기사전과 정합한 문자렬이 바로 가르기결과로 된다. 주로 정방향최대정합법과 역방향최대정합법, 쌍방향정합법, 축사편력정합법, 설립구분표시법, 정방향최량정합법, 역방향최량정합법 등이 있다. 이런 방법은 인공적으로 구축한 가르기사전의 범위가 제한된것으로 하여 모든 어휘를 적용할수 없으며 따라서 그 국한성을 피할수 없다. 이와 함께 규칙에 기초한 가르기방법은 사람이 제공하는 대량의 가르기지식을 필요로 하므로 정보기술의 발전과 정보처리범위가 확대됨에 따라 이 방법에 대한 높아지는 사용자들의 요구를 만족시키지 못하게 되였으며 결국 규칙에 기초한 자동가르기체계의 제한성이 나타나게 되였다.

통계에 기초한 방법은 글자와 글자, 단어와 단어사이의 동시출현확률에 의거하여 가르기를 진행하는 방법으로서 가르기사전을 리용하지 않는다. 이 방법의 우점은 그것이 응용령역의 제한을 받지 않으며 또한 미리 만든 가르기사전에 제한되지 않는다는것이다. 통계에 기초한 방법은 대규모의 현실에서 쓰이는 본문을 요구하며 출현빈도점수를 써서 해당 단어가 출현하는 본문에 대하여 실제적인 가르기를 하므로 일반적으로 비교적 큰 계산량을 요구한다. 한편 훈련에 리용되는 본문의 선택 역시 가르기결과에 대하여 뚜렷한 영향을 준다. 최근에 중국에서는 통계방법을 리용한 가르기체계가 점차 증가하고 통계정보 역시 간단한 단어빈도정보에 제한되지 않으며 품사 등 층차가 비교적 높은 정보

를 리용하여 구분모호성을 해소한다.

우에서 언급된 중국어가르기방법들에서는 최대정합산법, 완전구분산법, 통계적구분산법의 3가지 주요산법을 리용한다.

— 최대정합산법

이 산법은 콤퓨터에 의한 가르기산법에서 가장 대표적인 산법으로서 임의의 단어와 단어사이의 련립가능성을 고려하지 않고 오직 길이에 따라 사전에서 출현하는 문자렬을 찾는 산법이다. 최대정합산법에서 기본은 길이를 기본으로 하여 입구렬중에서 고정된 방향에 따라 매번 가장 긴 가능한 단어를 잘라내는것이다. 이 방법은 간단하지만 정확도가 낮다.

— 완전구분산법

이 산법은 구분점이 존재하지 않는 산법이다. 완전구분산법은 형식상 사전에 부합되는 모든 구분형식을 찾는 구분산법이다. 례하면 우리가 사용하는 사전에서는 련관된 홑글자단어를 내놓고 组合(조합), 合成(합성), 成分(성분)은 사전안의 단어이다. 그러므로 他的组合成分의 가능한 구분형식은 다음과 같다.

(1) 他/的/组/合/成/分/

(2) 他/的/组/合成/分/

(3) 他/的/组/合/成分/

(4) 他/的/组合/成/分/

(5) 他/的/组合/成分/

우의 실례에서 (5)가 정확하다. 그 표시는 방향그라프방법을 써서 표시할수 있다. 우의 구분식을 보면 역시 완전구분산법이 불필요하게 많은 구분식을 산생한다는것을 알수 있다. 이 구분식에 대하여 구분과정에 일정한 가지자르기산법, 일명《가지자르기》를 진행할수 있다. 즉 구분과정에 산생되는 불필요한 구분식에 대하여서는 수시로 배제를 진행하여 중복계산을 피할수 있게 한다. 우의 완전구분산법은 많은 구분식을 만드는데 이러한 구분식에서 가장 최량적인 구분식을 찾아내는것은 규칙을 리용하는 방법으로 해결할수 있고 통계적방법을 써서 해결할수도 있다.

— 통계적구분산법

통계에 기초한 방법은 글자와 글자사이, 단어와 단어사이의 동시출현확률을 리용하여 구분을 진행하는 방법이다. 이 방법의 우점은 그것이 응용령역의 제한을 받지 않으며 먼저 구축한 가르기사전에도 국한되지 않는것이다. 통계에 기초한 방법은 모형파라메터를 학습시키는데 대규모의 학습본문을 요구한다. 물론 학습이든 실제구분이든간에 일반적으로 비교적 큰 계산량을 요구한다. 그외에 훈련본문의 선택 역시 가르기결과에 대하여 뚜렷한 영향을 준다.

규칙에 기초한 방법과 통계에 기초한 방법이 각각 발전함에 따라 현재 많은 학자들이 이 두 방법을 결합하여 두가지 방법의 우점은 살리고 단점은 극복하면서 보다 좋은 결과를 얻기 위한데로 나가고있다. 순서를 보면 먼저 사전에 기초한 최대정합법을 써서 초보적으로 문장을 가른 다음 사전에 기초한 방법을 써서 여러 글자단어안에서 구분모호성을 발견하고 마지막에 통계에 기초한 방법을 써서 구분모호성과 품사모호성을 해소한다.

중국어단어가르기에서 다음으로 언급하여야 할 문제는 미등록어식별문제 즉 사전에 없는 단어들을 어떻게 식별하겠는가 하는 문제이다.

콤퓨터에 의한 중국어단어가르기를 실현하는데서는 보통 2가지 기본적인 장애가 있다. 하나는 구분모호성문제이고 다른 하나는 미등록어들에 대한 식별문제이다. 미등록단어라는 의미는 단어가르기를 위한 체계의 사전에 속해있지 않은 단어를 가리키는것이다. 중국어단어는 그 결합가능성이 다양한것으로 하여 아무리 포괄적인 대규모사전을 만든다고 하여도 모든 단어를 다 담을수는 없다. 또한 시간이 흐르는데 따라 대량적으로 출현하는 새 단어들을 모두 실시간적으로 사전에 반영한다는것은 불가능한것이다. 일부 사전에 빠진 상용어라든지, 아직 등록되지 않은 새 단어와 같이 사전에 받아들여야 할 아직 등록되지 않은 단어들에 대하여서는 반드시 미등록어식별과정을 거쳐야 한다. 미등록어식별은 그자체는 단어로 될수 있지만 인명이나 지명과 같이 사전에만 의거해서는 할수 없는 식별을 말한다.

중국어의 미등록어로는 중국인명과 중국지명, 외국인명과 외국지명 등을 들수 있다. 이러한 단어들의 조성방식에는 규칙성이 부족하며 더우기 이러한 단어들에 대한 자동식별에 도움이 될수 있는 구별표기 역시 부족하다. 동시에 미등록어로 되는 단어의 글자는 대부분 많은 한음절단어들로 이루어진다. 그러므로 미등록어에 대한 자동식별을 실현하는것은 매우 어려운 과제이며 또 실제한 언어환경에서 미등록어들을 어떻게 식별하고 구분하는가 하는것이 단어가르기전반체계에 미치는 영향은 매우 크다. 그러므로 미등록어를 잘 판단해내지 못한다면 중국어의 자동가르기체계는 실지 기계번역체계를 포함한 자연언어처리체계의 요구에 부합될수 없게 된다.

미등록어문제는 개방형체계에 특유한것으로서 최근년간 중시되고있는 연구문제이다. 이 문제를 해결하기 위해서는 미등록어를 구별하는데 필요한 지식을 각이한 수단과 방법들을 리용하여 얻어내는것과 함께 특정한 류형의 미등록어들을 예측하는 기능을 갖추어야 한다.

우선 중국인명에 대한 식별문제이다. 중국사람들의 이름은 성과 이름의 두 부분으로 구성되여있으며 알려진 성만 하여도 수천개이상이다. 현재《중화대사전》(중국민족출판사, 2013)에 1 942개가 수록되여있고《중국성씨집》(중국민족출판사, 1998)에는 5 544개가 수록되여있다. 또한 중국사람들은 이름의 선택을 제 마음대로 정하고있으며 어떤 똑똑한 규칙이 없이 완전히 개인의 기호에 따르고있다. 그러므로 임의의 한자나 한자렬로도 이름을 만들수 있다. 이로부터 모든 중문이름을 가르기사전에 등록하는것은 불가능하며 이것은 곧 단어가르기체계가 중국사람이름을 자동식별하는 능력을 갖출것을 요구한다.

중국어본문에서 인명의 식별은 복잡한 문제이다. 중국어본문에는 영어와 같은 대문자형태가 없고 이름량이 아주 많으며 시간이 감에 따라 부단히 변하는것으로 하여 그 어려움이 계속 커지고있다. 중문이름의 구조는 복잡하고 표현형식이 다양하며 이름에 리용되는 글자자체가 단어일수 있을뿐아니라 그와 린접한 글자와 함께 단어를 구성할수 있다. 례하면 马는 명사로서 가축을 가리키는 명사일수도 있고 성씨일수도 있다. 이것은 모두 인명식별의 난도를 증가시킨다.

현재 대다수 인명식별방법은 모두 인명의 분포규칙, 인명용글자규범, 성씨용글자, 인명용글자의 출현빈도와 확률값 및 인명의 전후약속용단어 등의 정보를 종합적으로 리용하여 추리판단을 진행하는 방법이다. 일부 연구자료들은 언어자료류형의 차이에 따라 인명이 각이한 류형의 언어자료에서 출현하는 차수도 크게 구별된다는것을 보여주었다. 여기에서

신문과 공식문건 등에서 사람이름출현의 차수는 기타 류형의 언어자료에 비하여 많다. 베이징의 어느 구역에서 사는 12만명의 사람들의 이름에 대해 통계를 진행한 결과 12만명의 사람이름에서 성이 836개였으며 여기에서 400개 성이 99%를 차지하였다. 이름용글자수는 12만명중에서 1 668개, 여기에서 앞의 3개(军, 伟, 静)가 4%를 차지한다. 한글자이름의 99%를 차지하는 글자는 1 313자로서 전체 이름용글자의 79%를 차지한다. 한글자이름과 두글자이름의 사용글자범위가 다를뿐아니라 한글자와 두글자이름, 머리글자범위도 각이하다. 례하면 小가 두글자이름의 머리글자로 쓰이는것은 8번째 자리를 차지하지만 한글자이름에서는 오히려 대단히 적게 사용된다.

아래에 코퍼스와 규칙에 기초한 중국인명식별방법에 대하여 개괄해보기로 한다.

(1) 중국인명코퍼스를 구축하고 통계를 진행하여 두글자이름의 글자와 그것이 두글자이름에서 차지하는 지위를 도출하고 한글자이름을 목록화한다.

(2) 최대정합법을 리용하여 자동가르기를 진행한다.

(3) 만일 문장에 사전에 있는 성씨의 단어가 출현하면 성씨글자 혹은 단어후에 한글자가 함께 있는것을 두글자이름글자로 리용할수 있는가, 혹은 두글자이름의 한개 머리글자가 따르는가, 혹은 두글자이름 마지막글자가 있는가, 혹은 뒤에 하나의 한글자이름이 있는가와 같은 조건을 따진다.

(4) 규칙을 리용하여 특정의 중국인명에 대한 조정을 진행한다. 이 규칙은 사실 사람이름판정에 쓸수 있는 식별정보이다. 여기에는 女士, 先生과 같은 호칭들과 지명과 단위가 속한다. 사람이름앞에 단위이름과 지명을 써서 소재지와 단위를 표시한다. 사람이름앞에 的가 붙으면 규정어이다. 실례로 年过五十的王珊(50살이 넘은 왕산)을 들수 있다.

(5) 단어가르기규범에서 小李, 老王, 张总 등 호칭과 존칭은 모두 가르기단위로 한다. 이러한 인명을 표시하는 가르기단위는 규칙을 리용하여 식별할수 있다.

미등록어식별에서 중요한것은 또한 중국지명에 대한 식별문제이다. 중국어본문에 만일 식별되지 않은 중국지명이 있다면 이것은 엄중한 가르기오유를 야기시킬수 있다.

실례로: 这是蓬莱县小门家乡政府所在地。

만일 지명식별이 없다면 아래와 같이 구분된다.

这/是/蓬/莱/县/小/门/家乡/政府/所在地。

지명은 사람이름과 달리 복잡하기때문에 하나하나 렬거할수 없다. 현재 일부 중국어자동가르기체계에서는 사전에서 지명을 선택하여 실현한다. 이것은 실제본문의 가르기체계에 대하여서는 그리 적용할만 한것이 못된다. 지명 즉 성, 시, 현, 향, 촌 혹은 이름난 강이나 큰 호수, 산간벽지 등은 리론적으로는 렬거할수 있지만 실제적으로 다 렬거한다는것은 불가능하다. 뿐만아니라 비록 렬거한다고 해도 현실적으로 지명의 수가 대단히 많기때문에 모든 지명들을 가르기사전에 넣는다면 우선 가르기사전의 규모가 기하급수적으로 커져 체계의 운영효률이 떨어지게 된다. 또한 각종 모호현상이 출현할 확률이 증가하여 가르기정확도에 영향을 미친다. 실례로 지명을 나타내는 于山(여산)을 사전에 넣으면 由于山区에 대한 가르기가 사슬길이를 2로 하는 교차모호성으로 변한다. 그러므로 중국지명의 특징을 연구하고 글자사용규범, 단어사용규범, 단어조성규범과 지명의 문맥규범을 연구하여 실제한 본문에서 중국지명의 자동식별을 실현하여야 한다.

중국지명의 특징은 일부 식별에 대하여 일정한 어려움을 조성한다는것이다. 실례로

중국지명의 길이에는 일정한 제한이 없고 한글자부터 여러 글자까지 다양하다. 실례로 京, 津 등의 략칭은 길이가 1, 또한 北京은 길이가 2, 內蒙古는 길이가 3인것 등을 들수 있다. 이외에 중국어의 상용글자가 자주 지명에 출현한다. 실례로 大直街, 马象沟의 한자들은 모두 상용글자들이다. 이외에 지명에 포함된 여러 글자단어들도 역시 지명의 식별에 불리한데 실례로 黃果树瀑布에서 果树자체가 바로 하나의 단어이다.

그러나 중국지명식별과정에 리용할수 있는 정보도 있다. 실례로 乡(향), 村(촌), 市(시), 县(현)등과 같이 지명식별에 리용할수 있는 접미사들을 들수 있다.

지명식별에도 역시 통계와 규칙을 결합한 방법을 리용할수 있다. 아래에 그에 대하여 보기로 한다.

(1) 먼저 1개 성, 자치구, 직할시, 구, 및 산줄기, 강, 호수, 만, 반도 등을 반영한 중국지명자료기지를 구축하고 지명을 가능한껏 많이 수집한다.

(2) 지명에서 지명용글자 및 지명의 머리글자, 중간글자, 끝글자규칙 및 빈도에 대한 통계를 장악하고 지명용글자자료기지에서 각 글자가 지명의 머리글자, 증간글자, 끝글자로 쓰이는 확률을 통계낸다.

(3) 가르기방법을 리용하여 가르기를 진행한다.

(4) 한글자로 된 지명에서 그 글자가 지명용자료기지에 속하면 지명식별과정에 들어간다. 우선 지명용앞글자의 단어 혹은 글자를 찾고 다시 중간 및 끝글자 혹은 단어를 찾는다. 글자빈도와 지명용글자의 통계규칙을 리용하여 지명의 경계를 확정한다.

(5) 글자에 대하여 1차식별을 진행한 후 그 정확도를 다시 조정할수 있다. 실례로 지명 黑龙江省哈尔滨(흑룡강성 할빈시)는 응당 둘로 갈라야 한다. 련속지명의 가르기에 대하여서는 아래의 두가지 방법을 리용할수 있다. 첫째로, 규칙을 사용하여 식별을 진행한다. 중문지명의 끝글자 성, 시, 현, 향 등은 뚜렷한 특징을 가지고 지명의 경계를 찾는데 쓸수 있으므로 그것들을 열쇠로 하여 규칙방법을 써서 식별할수 있다. 물론 다른 규범을 가지고 규칙을 표시하는데 쓸수 있을뿐아니라 지명식별과정에 쓸수도 있다. 둘째로, 확률합을 리용한다.

또한 미등록어의 중요한 부류로서 외국이름을 들수 있다. 외국이름이라고 할 때 그것은 중국어가 아닌 다른 언어로 된 사람이름이나 지명과 같은 고유명사들을 들수 있다.

중국어단어가르기과정에 만일 식별되지 못한 외국이름이 있으면 문장이 갈라지던가 보다 치명적인 오유가 생길수 있다.

례: 埃及/总理/穆/巴/拉/克/访问/叙利亚
에짚트/총리/무/바/라/크/방문/수리아
埃及/总理/穆巴拉克/访问/叙利亚
(에짚트총리 무바라크가 수리아를 방문)
国际/田联/取/消费/尔/南/多/参赛/的/资格
국제/륙상련맹/취하다/소비/尔/南/多/경기참가/의/자격
国际/田联/取消/费尔南多/参赛/的/资格
(국제륙상련맹에서는 페르난도의 경기참가자격을 취소하였다.)

우의 실례는 외국이름의 식별 역시 중국어단어가르기에서 중요한 자리를 차지한다는 것을 보여준다.

외국이름의 식별은 중국지명의 식별과 일부 류사한 점도 있지만 자기의 특징을 가지고있다. 그 특징은 우선 외국이름은 글자가 중국지명에 쓰이는 글자들에 비해 비교적 규범성이 있고 쓰는 글자가 중국지명보다 량적으로 작으며 또한 외국이름을 번역한것들중에는 威廉明娜(윌헬름미너)와 같이 여러 글자단어도 있다는것이다.

외국이름식별에 대부분 통계적방법을 리용하지만 식별과정에 문맥정보도 리용할수 있다. 아래에 외국이름식별의 한가지 방법을 준다. 우선 외국이름에 대한 자료기지를 준비하고 자료기지속에 있는 외국이름에 대한 통계를 진행한다. 외국이름에 쓰는 글자표 및 각 외국이름용글자로 쓰이는 앞, 중간, 끝글자들의 확률을 계산한다. 다음 일정한 가르기방법을 리용하여 가르기를 진행한다.

련속글자렬에 대하여 외국이름의 대략적인 한계를 규정한다. 즉 련속글자에 대하여 매개 글자가 모두 외국이름에 속하면 초보적으로 그 렬의 글자를 외국이름으로 한다. 다음 문맥을 써서 식별을 진행한다. 실례 印度总理尼夫人来访.(인디아총리 네루의 부인이 방문하였다.)에서 《네루》라는 이름에 호칭이 붙어 쉽게 식별되는데 이로부터 직위나 칭호에 속하는 단어들을 표로 작성하거나 사전안에 있는 직위, 칭호를 나타내는 단어들을 장악하여 리용하는것이 좋다고 본다.

례: 约翰·史密斯来中国。

(죤·스미스가 중국을 방문한다.)

중국어에서는 습관상 외국인명사이에 《·》를 쓰는 경향이 있는데 련속한자렬중에 《·》이 있으면 론의할 여지없이 외국이름으로 식별할수 있다.

이밖에도 외국이름을 식별하는데 문맥정보를 쓸수도 있는데 실례로 외국이름의 앞뒤에 자주 쓰이는 동사 등을 리용하는 방법도 있다.

문맥이 불명확한 련속한자렬의 외국어종의 식별에 대하여서는 한글자단어빈도와 번역이름용글자표의 빈도를 리용하여 판정할수도 있다. 이 경우에 해당 글자에 대한 등급비교를 진행하여 외국이름에 쓰이는 글자를 식별해내거나 해당 글자를 앞, 중간, 끝글자와의 빈도를 리용하여 직접 통계를 내기도 한다. 즉 외국이름으로 번역된 글자들의 앞에만 출현하는 글자표, 외국이름을 나타내는 글자렬의 시작에 출현할수 없는 글자표, 끝부분에만 출현하는 글자표, 끝부분에는 출현할수 없는 글자표를 장악한다. 다음 외국이름의 대략적인 경계규정의 기초우에 후보번역명에 대하여 《시작끝근사법》을 리용하여 시작끝경계를 식별해낸다. 이러한 공정을 거쳐 외국이름을 얻고 단어가르기가 진행된다.

끝으로 중국어자동가르기에서 제기되는 가르기모호성과 그 해소를 위한 방법에 대하여 보기로 한다.

사람들이 중국어를 리해한다고 할 때 그것은 곧 중국어글자렬에 대한 가르기과정으로 되며 이 과정에 형태부, 문법, 의미 등 각종 언어학적정보를 결합하게 된다. 리상적인 중국어가르기체계도 역시 이 정보들을 종합적으로 응용해야 한다. 그러나 콤퓨터로 처리하는 경우 이러한 언어학적정보를 추출하는것도 역시 단어가르기를 전제로 한다. 그러므로 중국어자동가르기와 언어학적정보의 응용은 호상련계하고 제약하는 관계이다. 순수한 기계적인 구분(최대정합법과 같은)은 반드시 구분모호성을 가져온다.

구분모호성은 중국어문장의 어느 한 글자단에서 순수 단어표에 근거하여 글자렬정합을 할 때 여러가지로 구분되는 경우를 말한다. 구분모호성이 포함된 중국어글자렬을

모호성을 가진 글자단이라고 한다. 새 단어로 하여 생기는 모호성을 내놓고 구분모호성의 형태에는 주로 다음의 3가지가 있다.

▪ 교차모호성: 중국어글자렬 ABC가 AB/C형식 또는 A/BC형식으로 구분될수 있다. 즉 AB가 단어이며 BC역시 단어이다.

▪ 조합모호성: 글자렬이 AB로, 또는 A/B로 구분될수 있다. 즉 AB가 단어, A, B역시 단어이다.

▪ 혼합모호성: 앞의 2가지 모호성형식의 개별적인 반복 혹은 나란히 존재하여 모호성이 생긴다.

구분모호성문제는 중국어자동가르기체계의 구분정확도에 직접적인 영향을 미친다. 이러한 구분모호성의 해결방법은 대체로 규칙에 기초한 해결방법과 통계에 기초한 해결방법으로 나눌수 있다. 초기에는 규칙에 기초한 방법이 비교적 많이 리용되였으나 연구가 심화되여 사람들의 가르기체계에 대한 개방성요구에 따라 수동작업인 모호성속성규칙방법이 점차 통계적방법으로 교체되였다. 그러나 규칙에 기초한 방법에도 일정한 우점은 있으며 일부 특수한 형태의 구분모호성을 해결하는데는 효과가 비교적 뚜렷하다.

아래에 모호성류형에 따라 그 해소방법에 대하여 보기로 한다.

► 교차모호성과 그 해소

교차모호성은 단어와 단어사이에 교차조합으로 산생된다. 즉 단어《ABC》가《AB/C》혹은《A/BC》로 갈라질수 있기때문이다.

실례로 不合理(불합리)는 不合/理로 구분될수 있고 또는 不/合理로 구분될수 있다. 不满意(불만족)도 不满/意, 또는 不/满意으로 구분될수 있다.

중국어의 단어와 구분후의 자료를 통하여 교차모호성글자단이 아래와 같은 특징을 가진다는것을 알수 있다.

(1) 교차모호성을 가진 글자단의 구분에서는 오직 한가지형식만 가질수 있다. 즉 그 글자단만으로는 구분모호성을 발견할수 없으며 정방향최대구분법을 리용하는 경우《AB/C》의 한가지 구분형식으로만 가를수 있다.

실례로《他在海上游泳。(그는 바다에서 수영을 한다.)》에서 모호성을 가진 글자단 上游泳은 上游/泳로만 갈라질수 있다.

(2) 교차모호성글자단은 확정적인 문맥에서 오직《A/BC》의 형식으로만 구분된다.

실례로《他在海上游泳。(그는 바다에서 수영을 한다.)》에서 모호성을 가진 글자단 上游泳은 上/游泳로만 갈라질수 있다.

(3) 교차모호성글자단에서 두가지 구분형식의 공통적인 부분의 글자개수를 사슬길이라고 한다. 실례로《{上[游}泳]》의 사슬길이는 1이다. 서로 다른 사슬길이의 교차모호성글자단의 구분은 서로 다른 특징을 가진다. 통계자료에 의하면 사슬길이가 1과 2인 모호성글자단의 개수는 각각 모호성글자단총수의 55%와 41%를 차지하며 그 출현의 차수는 모호성출현총수의 64%와 34%이다. 이것들은 합쳐서 모호성글자단총수의 96%를 차지하며 차수는 모호성출현차수의 98%이다. 만일 사슬길이가 1과 2인 교차모호성의 구분문제를 잘 해결하면 모호성구분의 정확도를 크게 높일수 있다.

아래에 서로 다른 사슬길이모호성구분의 특징을 설명한다.

1) 사슬길이가 1인 교차모호성글자단 ABC

(1)《ABC》는 언어학적지식의 리용에 적합하지 않다. 즉 중국어형태부, 문법 및 의미형식을 리용하기에 적합하지 않다.

실례: 增长(증가하다), 长汉(은하수)은 단독으로 단어를 이루지만 增长汉은 중국어에서 련이어 사용되지 않는다.

(2)《ABC》는 오직《AB/C》형식으로만 갈라질수 있다. 즉《ABC》는 중국어의 문법, 의미에 부합되지 않는다. 실례로 모호성글자단 考核对(검토조)는 오직 考核/对로만 구분될수 있다.

(3)《ABC》가 오직《A/BC》형식으로만 갈라질수 있다. 즉《AB/C》가 중국어의 문법, 의미에 부합되지 않는다.

실례로 上游泳는 오직 上/游泳로만 갈라질수 있다.

(4)《AB/C》,《A/BC》가 중국어문법, 의미층에서 모두 성립한다.

례: ① 他的确切了一块肉。(그는 정말 고기덩어리를 잘랐다.)

② 他的确切地址我知道。(그의 확실한 주소를 우리는 안다.)

여기에서 모호성글자단 的确切은 실례 ①에서 的确/切(정말 자르다)로 구분되였으며 실례 ②에서는 的/确切(의 확실한)으로 구분되였다.

이상의 분석으로부터 단어가르기에 의한 구분후의 모호성글자단형식이 보통《AB/C》로 되며 모호성은 (3), (4)의 두 경우에 발생하였다는것을 알수 있다. 또한 2만여개 문장의 언어자료에 대한 분석을 통하여 대다수의 모호현상은 모두 (3)의 경우에 속한다는것을 알수 있다. 그러므로 사슬길이 1의 구분난점을 해결하는것은《A/BC》형구분문제의 해결로 귀착된다.

2) 사슬길이가 2인 교차모호성글자단 ABCD

《ABCD》는《A/BC/D》,《A/B/CD》,《AB/CD》,《ABC/D》,《AB/BCD》,《AB/C/D》,《ABCD》의 형식으로 갈라질수 있다. 통계자료는《AB/CD》형이 98%를 차지한다는것을 보여준다. 실례로 已经过去는 已经/过去로 구분된다.

3) 사슬길이가 3인 교차모호성글자단 ABCDE

글자단《ABCDE》중에서 AB, BC, CD, DE는 각각 단어이다. 경험은 사슬길이 3의 교차모호성이 주로 앞에 있는 3개 글자의 구분을 잘해야 한다는것을 보여준다. 실례로 为人民服务(인민을 위하여 복무함)은 다만 为人民을 为/人民로 구분하여야 为人民服务를 정확하게 구분할수 있다.

4) 사슬길이가 4인 교차모호성글자단 ABCDEF

교차모호성을 가진 글자단《ABCDEF》중에서 AB, BC, CD, DE, EF는 각각 단어이다. 일반적으로《AB/CD/EF》로 구분한다.

실례로 中国产品质量(중국제품의 질)은 中国/产品/质量으로 구분된다.

교차모호성은 규칙에 기초한 방법과 통계에 기초한 방법으로 해소할수 있다.

규칙에 기초한 방법을 위해서 구분모호성해소규칙을 아래와 같은 형식으로 규정한다.

례: @上游:1=荡|行|泳|玩$

단어 上游로 구분된 후 즉 이 규칙기지를 조사하고 만일 찾았다면 구분렬에서 游의 다음글자가 규칙에서의《=》와《$》사이의 렬에 있는가를 본다. 만일 있다면 곧

上/游로 구분한다.

또한 사슬길이가 서로 다른 교차모호성글자단에 대한 구분특징에 따라 서로 다른 사슬길이의 교차모호성글자단에 대하여 서로 다른 규칙기지에 구축한다. 례외적인 구분단위에 대하여서는 례외규칙기지를 조사하여 해소한다.

통계적방법에 기초한 가르기산법은 확률곱하기산법에서 이미 모호성의 해소에 대해 취급하였으므로 론의하지 않는다.

▶ 조합모호성과 그 해소

이 모호성은 한개 단어중의 일부 부분이 또 다른 단어를 구성하는것을 가리킨다. 즉《AB》로 될수도 있고《A/B》로 갈라질수 있는것을 말한다. 실례로 个人은 个人(개인)으로 구분할수 있고 또한 个/人(개/사람)으로 구분할수 있다. 또한 马上은 马上(당장)과 马/上(말/우)로 구분할수 있다.

조합모호성과 교차모호성은 류사한 점이 많은데 그 특징은 아래와 같다.

(1) 자동가르기시에 조합모호성글자단은 일종의 구분형식을 가질수 있다. 그자체로는 구분모호성을 발견할수 없다. 정방향최대정합법에 의하여 오직 한개의 형식 즉《AB》로만 갈라질수 있다. 실례로 모호성글자단 个人은 다만 个人(개인)으로만 구분할수 있다.

(2) 조합모호성글자단은 확정적인 문맥환경에서 오직 한개의 구분형식만 가질수 있다. 실례로《将来的产业会更加繁荣。(장래의 산업은 더욱 번영할것이다.)》에서 조합모호성을 가진 글자단 将来는 오직 将来(장래)로만 구분할수 있다.

(3) 조합모호성글자단의 구성형식은 아래의 정황에서 더 세분할수 있다.

정황 1:《AB》는 오직《AB》로만 갈라질수 있고《A/B》는 중국어의 형태부, 문법 및 의미에 부합되지 않거나 가능성이 극히 적다. 실례로 空白에서 空, 白는 모두 단어를 이루지만 空, 白는 단독으로 단어를 이룰 때 련이어 쓰이지 않는다. 만일 련속쓰기라면 오직 空白의 형식으로만 구분될수 있다.

정황 2:《AB》,《A/B》가 중국어문법의미층에서 모두 성립한다. 실례로《他学会了解方程。(그는 배워서 방정식을 풀줄 알게 되였다.)》에서 모호성글자단 了解는 了/解로 구분되여야 한다. 또한 실례《他很了解我。(그는 나를 잘 안다.)》에서 모호성글자단 了解는 오직 了解(료해하다)로 구분될수 있다.

조합모호성글자단은 교차모호성에 비해 찾기 쉽다. 그것은 자동가르기를 진행한 후 모호성글자단의 형식은 오직《AB》형식으로밖에는 될수 없기때문이다. 즉《AB》가 단어를 이루고《AB》자체가 곧 모호성글자단의 전부로서 정확한 구분형식이거나《AB》혹은《A/B》중에 포함되므로 사전에서 오직 어간《AB》만 모호성표식이 가능하게 해야 한다. 특징 (3)에서 정황 1을 모호성으로 보지 않으며 어간이 모호성표기를 가진다는것을 발견할 때 즉시 그에 대하여 모호성처리를 진행할수 있다.

그러나 조합모호성해소는 교차모호성에 비해 더 힘들다. 그것은 조합모호성글자단자체가 모호성있는 단어이며 모호성이 없는 단어가 모호성이 있는 단어속에 포함되기때문이다. 이러한 모호성은 반드시 문맥에 의거하여 해결해야 하며 모호성글자단자체 및 그 전후 한자단어(혹은 글자)를 통하여 추리판단해야 한다. 이런 방법으로 중국어문법에서 비교적 복잡한 구분문제를 해결하여야 한다. 조합모호성의 난점은 중국어자체의 특징으로부터 조성

되는것이다. 아래의 실례를 들어 시험해보기로 하자.

这个问题不是一个人所能解决的。

(이 문제는 한사람이 풀수 있는 문제가 아니다.)

기계적구분후에 얻어진 결과는 /这/个/问题/不/是/一/个人/所/能/解决/的/, 여기에서 글자단 个人에 모호성이 생기는데 이 한가지 규칙으로 个人에 대하여 쓸수 있다.

만일 个人의 앞단어가 수사나 혹은 지시대명사라면 个人은 응당 갈라야 한다. 만일 하나의 단어로 구분하지 않는다면 그때에는 우의 실례에 대하여 우선 个人의 모호성표시에 따라 모호성글자단 个人을 발견하고 다시 우의 규칙에 따라 이때의 个人이 个/人로 구분되여야 한다는것을 판단할수 있다.

조합모호성에 대한 처리는 형태부해석 및 문장구조해석을 하는 단계에서 동시에 진행할수 있다. 그것은 분석과정에 더 많은 단어의 정보 및 상관문맥정보가 얻어질수 있기때문이다.

► 혼합모호성과 그 해소

이 모호성은 앞의 두가지 모호성으로 하여 아래의 특징을 가진다.

(1) 자동가르기를 진행할 때 하나의 구분형식만 가지며 자체내에서는 구분모호성을 발견하지 못한다. 실례로 모호성글자단 充分地利用은 充分/地利/用으로 잘못 구분될수 있다.

(2) 혼합모호성글자단은 실지 문맥환경에서 오직 하나의 구분형식만 가질수 있다. 실례로《我们要充分地利用这个条件。(우리는 이 조건을 충분히 리용하여야 한다.)》에서 모호성글자단 充分地利用은 充分/地/利用으로만 구분된다.

(3) 혼합모호성글자단의 구성형식은 아래의 정황에서 세분할수 있다.

정황 1: 자동가르기를 진행할 때 여러가지 가능한 경우가 있지만 중국어의 형태부, 문법 혹은 의미에 부합되는것은 오직 하나이다. 실례로 要好好学习은 다만 要/好好/学习로만 구분할수 있다.

정황 2: 각이한 구체적언어환경에서는 여러가지 구분이 가능하다. 실례로 문장《政治局面对我方有利。(정치정세가 우리편에 유리하다.)》에서 모호성글자단 政治局面对은 政治/局面/对로 구분된다. 문장《政治局面对当前形势发出了新的号召。(정치국은 직면한 정세에 대처하여 새로운 호소를 하였다.)》에서는 政治局/面对로 구분된다.

혼합모호성글자단내부가 포함하는 정보는 풍부하고 외연은 작다. 그러므로 글자단내부가 서로 제약하여 정확한 구분이 이루어진다. 그러므로 정황 1이 큰 비률을 차지하며 정황 2는 상대적으로 매우 작다. 이렇게 하면 쉽게 해결방법을 찾을수 있으며 단지 정황 1을 만족하는 모호성글자단을 구분견본으로 해당한 규칙기지에 넣기만 하면 즉시 쓸수 있다. 다시말하여 사전을 거치지 않고 이 규칙기지를 리용하여 직접 이런 모호성을 정확히 구분할수 있다. 정황 2를 만족하는 모호성글자단에 대하여서는 먼저 규칙기지를 통과시켜 해소하며 만일 해소가 안되면 사람기계대면부 혹은 확률통계모형을 통하여 해소할수 있다.

이러한 모호성의 류형에 대하여 잘 알고 중국어자동가르기의 정확도를 제고하기 위한 방법을 연구하여야 한다.

중국어자동가르기는 중국어정보처리의 중요한 부분으로서 구분의 정확도를 어떻게 높이는가 하는것은 중국어가르기에서 관건적인 문제이다. 현재 개발리용되고있는 중국어자동가르기의 정확도를 높이는 방법을 몇가지 측면에서 제기하려고 한다.

먼저 중국어단어가르기규범을 사전과의 련관속에서 보다 완비하는 방법이다.

이를 위해서 단어가르기규범과 일치하는 기계사전을 구축하고 그 성능을 높이는 방법을 취할수 있는데 단어가르기는 오직 해당 사전상에서만 진행하며 일단 이 공정을 거쳐 사전이 확정되면 사전상의 구분규칙이 곧 단어가르기규범으로 될수 있다. 여러 자연언어처리체계들에서 리용할수 있는 사전을 구축하려면 사전의 올림말들 역시 단어가르기규범과 일치해야 한다. 그러나 사전상에서 바라는 구분결과를 완전히 얻을수 없는것이 문제로 된다. 따라서 가르기산법을 부단히 개선해나가는 방도를 취하여야 한다.

다음으로 가르기산법을 개선하는 방법이다. 여기에서 중요한것은 미등록단어의 식별과 그것들에 대한 구분모호성해소를 강화하는것이다. 중국어에는 미등록어의 종류가 대단히 많다. 중국인명, 외국인명, 중국지명, 수사 등을 실례로 들수 있다.

미등록어의 식별과 그 모호성해소는 통계적방법을 쓸수도 있고 규칙방법을 쓸수도 있으며 두 방법을 결합한 혼합방법을 쓸수도 있다.

미등록어에 대한 식별과 모호성해소에서는 긴 단어우선선택의 원칙을 리용한다. 그것은 긴 단어가 단어일 때의 확률이 긴 단어가 다시 짧은 단어로 될 때의 확률보다 훨씬 높은것과 관련된다.

3. 결 론

중국어단어가르기는 중국어를 입구어로 하는 각종 자연언어처리체계의 기초이다. 중국어단어가르기를 어떻게 하는가에 따라 전반적인 체계의 성능이 좌우된다고도 말할수 있다. 때문에 중국어단어가르기는 해당 체계개발에서 반드시 해결하여야 할 문제이다.

론문에서는 이것을 해결하는것을 기본과제로 제기하고 중국어단어가르기산법에 대하여 밝히고 중국인명과 지명, 외국인명들과 같은 미등록단어들을 식별하기 위한 언어학적연구를 진행하였으며 이와 함께 중국어자동가르기과정에 제기되는 모호성의 류형을 3가지로 가르고 그 해소 및 가르기정확도를 높일수 있는 방법론을 주려고 하였다.

우리는 앞으로도 이 분야에 대한 연구를 더욱 심화시켜 능률적인 자연언어처리체계개발에 적극 이바지하여야 할것이다.

실마리어 중국어단어가르기, 기계번역, 미등록어처리, 모호성해소