쉼표의 자동분류에 따른 중국어 장문분할
(Segmentation of Long Chinese Sentences using Comma Classification)

김 미 훈 † 김 미 영 †† 이 종 혁 †††
(Meixun Jin) (Mi-Young Kim) (Jong-Hyeok Lee)

요 약 입력문장이 길어질수록 구문분석의 정확률은 크게 낮아진다. 따라서 긴 문장의 구문분석 정확률을 높이기 위해 장문분할 방법들이 많이 연구되었다. 중국어는 고립어로서 자연언어처리에 도움을 줄 수 있는 굴절이나 어미정보가 없는 대신 쉼표를 비교적 많이, 또 정확히 사용하고 있어서 이러한 쉼표사용이 장문분할에 도움을 줄 수 있다. 본 논문에서는 중국어 문장에서 쉼표 주변의 문맥을 파악하여 해당 쉼표 위치에 문장분할이 가능한지 Support Vector Machine을 이용해 판단하고자 한다. 쉼표의 분류의 정확률이 87.1%에 이르고, 이 분할모델을 적용한 후 구문분석한 결과, 의존트리의 정확률이 5.6% 증가했다.

키워드 : 장문분할, 구문분석, 쉼표, SVM

Abstract *The longer the input sentences, the worse the parsing results. To improve the parsing performance, many methods about long sentence segmentation have been reserarched. As an isolating language, Chinese sentence has fewer cues for sentence segmentation. However, the average frequency of comma usage in Chinese is higher than that of other languages. The syntactic information that the comma conveys can play an important role in long sentence segmentation of Chinese languages. This paper proposes a method for classifying commas in Chinese sentences according to the context where the comma occurs. Then, sentences are segmented using the classification result. The experimental results show that the accuracy of the comma classification reaches 87.1%, and with our segmentation model, the dependency parsing accuracy of our parser is improved by 5.6%.*

Key words : Long sentence segmentation, Syntactic analysis, SVM, Comma

1. 서 론

중국어는 고립어로서 굴절현상도 없고, 어미를 이용하여 격을 나타내지도 않는다. 중국어 문장에서 대등접속문이나 종속접속문은 연결어 없이 나타나기도 한다. 따라서 타 언어처리에서 쉽게 사용될 수 있는 굴절이나 어미 등을 이용한 언어처리 방법론들을 중국어 처리에 적용하기에 어려움이 있다. levy와 Manning[1]은 중국어 문장 분석에 있어서 타 언어와 다른 구문분석 애매성을 지적했다. 중국어 장문분할 또한 타 언어와 다른 점들이 많이 존재한다. 중국어는 단어와 단어 사이에 띄어쓰기가 없는 반면, 중국어 문장에서 쉼표의 쓰임은 타 언어에 비해서 빈번하다[2]. 영어문장에서 쉼표의 사용빈도는 0.869[3]에서 1.04[4]이지만, 중국어문장에서 쉼표의 사용 빈도는 1.79로서, 영어문장에서 사용빈도의 약 1.5배에서 2배가 된다. 한국어에서 쉼표의 사용은 영어보다도 적다[2]. 그리하여 한국어나 영어의 구문분석이나 문장분할 연구에서는 쉼표가 항상 유용한 정보를 줄 수는 없다고 생각하고, 쉼표정보를 생략해 버리는 경우가 많다. 고립어인 중국어의 경우 구문분석이나 문장분할에 활용할 정보가 부족한 상황에서 쉼표가 나타내는 정보는 매우 유용하고, 이러한 정보는 중국어 문장분할의 적절한 위치를 찾는 데 큰 도움이 된다.

중국어 쉼표의 쓰임은 문맥에 따라 10여 가지에서 20여 가지로 나열할 수 있다. 이런 사용법들은 크게 중국어 문장 중 절 끝에 쉼표가 사용되는 경우와 절 내의 단어들 사이에 사용되는 경우로 나뉠 수 있다. 절의 끝

· 본 연구는 첨단정보기술연구센터를 통하여 과학재단의 지원을 받았음
† 학생회원 : 포항공과대학교 컴퓨터공학과
meixunj@postech.ac.kr
†† 정 회 원 : 성신여자대학교 컴퓨터정보학부 교수
miykim@sungshin.ac.kr
††† 종신회원 : 포항공과대학교 컴퓨터공학과 교수
jhlee@postech.ac.kr
논문접수 : 2005년 5월 16일
심사완료 : 2005년 12월 26일

위치에 사용된 쉼표는 중국어에서 전체 쉼표 사용의 약 30%를 차지하고 있다. 이와 같이 절의 끝위치에 사용된 쉼표의 위치에서 문장분할을 할 경우, 문장을 절 단위로 분할을 하였기에 전체 문장의 정확한 구문분석에 도움을 줄 수 있다. 하지만 절 내의 단어들 사이에 사용된 쉼표위치에서 문장분할을 할 경우, 문장분할로 인한 구문분석의 오류가 발생되어 전체문장 구문분석에서 틀린 결과를 내는 원인이 된다. 그러므로 중국어에서 쉼표위치를 문장분할 대상으로 사용할 경우, 쉼표가 절의 끝에 사용되었는지를 확인하여야 한다.

본 논문에서는 쉼표가 사용된 문맥에 따라 그 사용을 분류하고, 쉼표의 분류에 따라 문장 분할 적정 위치를 구분하는 방법을 제안한다. 논문의 구성은 아래와 같다. 2장에서는 장문 분할과 자연언어처리에서의 문장기호처리 등 기존연구에 대하여 먼저 소개가 있고, 3장에는 쉼표의 사용법에 따라 쉼표 분류의 기준을 제시한다. 4장에서 쉼표의 분류를 위한 자질 설명을 하고, 5장에서 쉼표 분류 기준에 따라 Support Vector Machine(이하 SVM으로 간략)으로 분류한 후, 분류한 결과에 따라 문장을 분할한다. 마지막으로 본 논문에서 제안한 문장 분할을 실행한 후의 구문분석기 성능과 문장분할을 하지 않은 구문분석기의 성능간의 비교 및 결론으로 이어진다.

2. 기존연구

2.1 절 분할

문장이 길어짐에 따라 구문분석의 애매성은 급속히 증가한다. 이를 극복하기 위한 방법으로 문장 분할을 자연스럽게 택하게 된다. 이에 대하여 여러 학자들이 여러 언어를 대상으로 다양한 분할 단위와 방법을 시도하였다. 장문의 분할에 관한 많은 기존 연구들을 살펴보면 절을 분할의 단위로 선택하였다[5-7]. 또한 [8]에서는 저자가 적당한 분할 단위를 설정하여 문장을 분할하기도 하였다.

하지만 기존 문장분할의 연구에서는 모두 쉼표를 무시하거나, 쉼표를 한 개의 자질로 삼았을 뿐 문장분할에서 쉼표의 정보를 충분히 이용하지 못하고 있다.

2.2 문장기호 처리

문장기호에 관련된 연구는 아래와 같이 두 가지로 나눌 수 있다.

첫째로 말뭉치를 이용하여 쉼표가 문장에서 어떻게 쓰이고 있는지에 대한 기술이다. Jones[9]는 문장기호들의 구문적 기능에 대하여 연구하였고 Bayraktar와 Akman[10]는 영어문장에서 쉼표가 사용된 구문적 패턴에 따라 쉼표를 분류하는 연구를 했다. 이러한 연구들은 문장기호들이 어떤 언어환경에 쓰이고 있는지에 대한 서술에만 집중되어 있다.

다른 학자들은 문장기호가 주는 정보를 구문분석 때 이용하고, 이렇게 함으로써 보다 나은 구문분석의 결과를 낼 수 있었다는 것을 보여주었다. Jones[11]는 문장기호가 포함이 된 문법을 이용하여 구문분석을 하는 것이 문장기호를 사용하지 않은 것보다 좋은 결과를 낼 수 있다고 했다. Collins[12]는 자신의 통계적 구문분석기에 쉼표를 중요한 자질로 삼았고, Briscoe와 Carroll [13]은 문장기호가 구문분석의 애매성 해소에 도움이 된다는 주장을 했다. Shiuan와 Ann[14]은 쉼표의 위치를 영어 복합문 분할에서 분할 후보로 정하고 후보들에 대한 분석을 하여 문장 분할 적정 위치인지를 판단한 후, 이로써 구문분석의 오류를 21% 줄일 수 있었다.

구문분석 외에 자연언어처리의 다른 분야에서도 문장기호 정보를 이용하여 좋은 결과를 낼 수 있다. Say[15]에서 이에 대하여 더욱 자세하게 설명하고 있다.

이러한 연구들은 문장기호가 자연언어처리에서 중요한 정보를 주고 있다는 것을 보여주고 있다. 특히 복합문 분할에서 쉼표사용에 대한 분석은 더욱 필요하다는 것을 말해주고 있다.

3. 구문분석을 위한 쉼표의 분류

중국어에서 쉼표는 가장 많이 사용되고 있는 문장기호로서 구문분석에 미치는 영향도 가장 크다. 중국어 문장의 구성성분들 중 분할후보를 찾기 어려운 상황에서 쉼표 위치에서 문장분할을 생각하는 것은 당연하다. 예문1을 쉼표위치에서 분할하면 他们上午上课와 下午做试验 두 부분으로 되고 나뉘어진 두 부분을 세그멘트라고 본 논문에서는 칭한다. 그러므로 n개 쉼표가 있는 문장을 쉼표위치에서 분할하면 n+1개의 세그멘트가 형성된다. 한 쉼표의 왼쪽에 위치한 세그멘트를 왼쪽 세그멘트, 오른쪽에 위치한 세그멘트를 오른쪽 세그멘트라고 본 논문에서 칭한다.

쉼표분할과정이 포함된 구문분석의 순서는 아래와 같다.

1. 쉼표 위치에서의 문장분할
2. 쉼표 왼쪽 세그멘트와 오른쪽 세그멘트 내부의 의존구문분석
3. 각 세그멘트들 간의 의존관계 설정

아래의 예문 1를 위 3단계로 의존문법 구문분석 한다면 첫 단계로 쉼표위치에서 문장 분할을 하고, 두 번째 단계로 각 세그멘트에 대하여 각각 의존구문분석을 한다. 왼쪽 세그멘트 부분에서는 上课(수업하다)라는 단어가 왼쪽 세그멘트의 루트(root)가 되고 오른쪽 세그멘트 부분에서 做(하다)라는 단어가 루트(root)가 된다. 각 세그멘트에서 루트(root)만이 헤드가 없는 단어이어야 하고, 이러한 단어를 그 세그멘트의 *메인헤드*라고 부르겠

다. 3번째 단계는 두 세그멘트의 메인헤드 사이의 의존관계 설정함으로써 쉼표의 좌우 분할 사이의 의존관계를 설정하게 되고 그림 1과 같이 의존구문분석이 완성된다.

예문[1] 1: 他们/ 上午/ 上课,/ 下午/ 做/ 试验。
그들/ 오전/ 수업,/ 오후/ 하다/ 실험.
(그들은 오전에 수업하고 오후에 실험한다.)

그림 1 문장분할을 이용한 의존구문분석

위의 과정에서 보여준 것과 같은 방법으로 구문분석할 경우, 분할과정 없이 구문분석하는 것보다 구문분석의 복잡도를 줄일 수 있어 전체문장의 의존구문분석이 더욱 효율적으로 이루어진다. 그러나 중국어 문장에서 모든 쉼표위치에서 분할이 가능한 것은 아니다. 아래 예문 2와 3의 경우는 쉼표위치에서 문장을 분할하고 각 세그멘트들을 단독적으로 의존구문분석을 할 경우, 쉼표위치 문장분할로 인한 의존구문분석 오류가 발생하여 전체문장의 의존구문분석에 나쁜 영향을 끼치게 된다.

예문 2를 쉼표 위치에서 문장 분할 후, 분할된 구간내에서 구문분석 수행시 헤드가 없는 단어는 오직 하나여야 하는데, 왼쪽 세그멘트 부분에 헤드로 설정가능한 용언이 없으므로 '北海'와 '在' 둘 다 헤드가 없게 되어 구문분석이 실패한다. 이 두 단어의 헤드는 왼쪽 세그멘트 부분에 있는 것이 아니고, 오른쪽 세그멘트 부분의 단어 '是'이다. 이러한 문장의 경우 쉼표위치에서 문장분할하게 되면, 정확한 의존구문분석 결과를 낼 수 없다.

예문 2: 北海/ 在/ 数年前,/ 是/ 一个/ 默默无闻/ 的/ 小/ 渔村。
북해/ ( ) / 몇 년 전,/ 이다/ 한/ 이름없다/ ~한/ 작은/ 어촌
(몇 년 전에, 북해는 이름 없는 작은 어촌이었다.)

예문 3과 같은 경우도 마찬가지다. 쉼표 위치에서 문장 분할 후 왼쪽 세그멘트 부분 '学生们比较喜欢年轻'을 독립적으로 의존구문분석을 하면 단어 '年轻'의 헤드가 '喜欢'으로 잘못 설정된다. 실제로 단어 '年轻'와 '喜欢' 둘 다 오른쪽 세그멘트 부분에 헤드를 가지고 있으므로 전체 문장의 구문분석이 실패하게 된다.

예문 3: 学生们/ 比较/ 喜欢/ 年轻,/ 美丽/ 的/ 教师。
학생들/ 더욱/ 좋아하다/ 젊다,/ 아름답다/ ~한/ 선생님
(학생들은 젊고 예쁜 선생님을 더욱 좋아한다.)

예문 2와 같이 한 세그멘트에 헤드가 없는 단어가 여러 개일 경우, 다음과 같은 보완방법으로 전체 의존구문분석의 결과를 낼 수도 있다. 한 세그멘트에 헤드가 없는 단어들을 다른 쪽 세그멘트의 메인 헤드와 의존관계 설정을 하게 되면, 전체 의존구문분석이 이루어진다.

하지만 예문 3과 같은 경우는, 이러한 보완 방법으로도 정확한 의존구문결과를 내는 데 부족하다. 왜냐하면 예문 3과 같은 경우는 비록 잘못된 구문분석 결과일지라도 왼쪽 세그멘트가 메인헤드를 하나만 가진 채 구문분석이 되기 때문에, (물론 이 구문분석 결과는 오류가 있지만, 구문분석기는 결과를 내게 된다), 자동적으로 오류여부를 판단하기 어렵다. 따라서 결국 잘못된 구문분석 결과가 도출된다. 문장의 길이가 길어지고, 쉼표의 개수가 늘어나게 되면 그 어려움은 더 크다.

예문 2와 예문 3에서 보여준 것과 같이 중국어 문장의 모든 쉼표위치에서 문장분할을 하게 되면, 적절하지 않은 문장분할로 인한 구문분석 오류가 발생할 수 있다.

그렇다면 어떤 쉼표의 위치가 분할 가능한 위치인가? 중국어 문장 중에 쉼표가 나타났을 경우, 그 쉼표가 문장 중의 쓰임 방식에 따라 분할가능 쉼표인지 분할불가능 쉼표인지에 대한 판단이 정확히 이루어진다면, 구문분석에 큰 도움이 된다. 이를 위하여 쉼표의 쓰임새에 대한 분석이 필요하다. 먼저 언어학자는 쉼표에 대하여 어떻게 분석하고 있는지를 살펴보겠다.

3.1 종속분할자 쉼표(Delimiter comma)와 대등분할자 쉼표(separator comma)

Nunberg[16]은 쉼표의 사용방식을 두 가지로 분류했다. 쉼표가 비슷한 성분[2]들 사이에 사용되었는지 아니면 다른 성분 사이에 사용되었는지를 분류기준으로 정하고 판단했다. 비슷한 성분들 사이에 사용된 쉼표는 대등연결 기능을 가지고 있고 이를 대등분할자 쉼표(이하 SC로 칭함)로 칭했다. 예문 3과 예문 4에서 쉼표는 대등관계인 두 절 혹은 두 단어의 접속사 기능을 하고 있어 Nunberg의 분류에 따르면 SC로 볼 수 있다. 만약 쉼표가 한 문장 중에 서로 다른 성분 사이에 사용된다면 서로 다른 성분들의 경계를 나타낸다. 이러한 쉼표를

1) 중국어 문장의 단어들 사이에는 공백이 없다. 중국어를 모르는 독자들을 위하여, 그리고 한글 매핑을 위하여 단어들 사이에 공백을 넣었다. 이하 예문 중 단어들 사이 공백은 모두 이러한 경우이다.

2) 본 논문에서 뜻하는 같은 성분은 문장에서 같은 구문적 기능을 지닌 두 단어나 구를 말한다. 이러한 두 단어나 구는 대등접속어 또는 대등접속구를 이룬다.

종속분할자 쉼표(delimiter comma, 이하 DC로 칭함)로 분류했다. 예문 2에서의 쉼표는 주어와 서술어 사이에 사용되어 주어와 서술어 사이의 경계를 나타내고, 예문 5에서 쉼표는 종속절과 주절 사이에 사용되어 종속절과 주절의 경계를 나타내고 있어, Nunberg의 분류 방법에 따라 이러한 쉼표들은 DC이다.

예문 4: 小明/ 在/ 写/ 作业,/ 妈妈/ 在/ 打/ 毛衣。
시오밍/ ( )/ 하다/ 숙제,/ 엄마/ ( )/ 있다/ 스웨터
(시오밍은 숙제하고, 엄마는 스웨터를 짜고 있다.)

예문 5: 尽管/ 他/ 很/ 努力,/ 但/ 成绩/ 不/ 理想。
~이지만/ 그는/ 매우/ 노력,/ 그러나/ 성적/ 아니다/ 이상적이다
(그는 매우 노력하지만 성적은 별로다.)

Nunberg의 분류방식은 나름대로의 의미가 있고, 언어학적으로 쉼표 쓰임이 어떻게 다른가를 보여주고 있지만, 실제 자연언어처리, 특히 문장분할에서 이 분류방식을 그대로 적용할 수 없다. 예문 2와 예문 5의 경우 둘 다 쉼표의 분류가 DC에 속하지만, 예문 2의 쉼표위치는 분할불가능 위치이고 예문 5의 쉼표위치는 분할가능하다. 즉 DC로 분류된 쉼표 내에 분할가능 쉼표, 분할불가능 쉼표가 모두 포함되어 있다. SC로 분류된 예문 3의 쉼표위치는 분할 불가능 위치이지만, 같은 SC로 분류된 예문 4의 쉼표위치는 분할가능 위치이다. DC와 마찬가지로 SC로 분류된 쉼표에도 분할가능 쉼표와 분할불가능 쉼표가 함께 포함되어 있다. 그러므로 쉼표를 SC와 DC로 분류하는 방법은 중국어 문장을 적절히 분할하는 데 도움을 줄 수 없다.

3.2 절 내의 쉼표(intra-clause comma)와 절절 간의 쉼표 (inter-clause comma)

본 논문에서는 쉼표가 절내에 사용되었는지, 절절 간에 사용되었는지에 따라 절내의 쉼표(intra-clause comma, 이하 절내 쉼표로 약칭)와 절절 간(inter-clause comma, 이하 절절 쉼표로 약칭)의 쉼표로 분류한다.

쉼표는 문장 중에서 쉼(pause)의 위치를 나타내고[2], 두 세그멘트를 연결하는 기능을 한다. 한 세그멘트는 한 개 구나 여러 개의 구 또는 절이 될 수 있다. 쉼표가 연결한 양쪽 세그멘트가 절인지 구인지에 따라 본 논문에서는 (c,c), (p,p), (p,c), (c,p)로 해당 쉼표에게 값을 할당하고자 한다. 여기서 c는 절(clause)을 의미하고, p는 구(phrase)를 의미한다.

예문 1의 좌우 세그멘트는 모두 절이다. 이럴 경우 해당 쉼표의 쓰임을 (c,c)로 표현한다. 예문 4와 5에 사용된 쉼표도 모두 (c,c)에 해당된다. 쉼표가 (c,c)일 경우, 두 절 사이의 관계는 대등관계(예문 4)이거나 종속관계(예문 5)가 될 수 있다.

예문 2의 경우, 쉼표의 왼쪽 세그멘트는 명사구 北海와 전치사구 在数年前로 이루어졌고, 오른쪽 세그멘트는 동사구 하나로 구성되었다. 따라서 해당 쉼표의 쓰임을 (p,p)로 표현한다. 예문 3과 아래쪽의 예문 6,7에 사용된 쉼표들은 모두 (p,p)에 해당된다.

쉼표가 나타내는 정보는 (c,c)와 (p,p)외에 (p,c) (c,p) 두가지가 더 있다. 예문 8의 두 번째 쉼표와 예문 9,10의 쉼표의 값은 (c,p)에 해당된다. 이 때 쉼표 왼쪽 세그멘트는 한 개의 절이고, 오른쪽 세그멘트는 한 개의 구나 여러 개의 서로 겹치지 않은 구로 구성이 되어 있는 경우이다.

쉼표의 값이 (p,c)일 때는 왼쪽 세그멘트가 구이고 오른쪽 세그멘트가 절이다. 예문 11의 두번째 쉼표와 예문 12, 13의 쉼표들은 이 경우에 해당한다.

예문 6: 俄罗斯/ 国内/ 经济/ 的/ 发展/ 变化/, 促进了/ 两/ 国/ 之间/ 的/ 贸易/ 往来。
러시아/ 국내/ 경제/ 의/ 발전/ 변화/, 촉진하다/ 두/ 나라/ 사이/ 의/ 무역/ 왕래
(러시아 국내 경제의 발전변화는 두 나라 사이의 무역왕래를 촉진하였다.)

예문 7: 中国银行/ 在/ 去年/ 十月/, 聘请/ 日/ 某/ 公司/ 做/ 顾问/。
중국은행/ ( )/ 작년/ 10월/, 청빈/ 일본/ 모/ 회사/ 하다/ 자문/
(작년 10월에 중국은행에서는 일본 모 회사를 자문으로 청빈했다.)

예문 8: 在/ 单位/ 里/, 他/ 是/ 好/ 领导/, 在/ 家/ 里/, 他/ 是/ 好/ 爸爸。
( )/ 회사/ 에서/, 그/ 이다/ 좋은/ 지도자/, ( )/ 집/ 에서/, 그/ 이다/ 좋은/ 아빠
회사에서 그는 좋은 지도자이고, 집에서는 좋은 아빠다.

예문 9: 科研/ 成果/ 迅速/ 转化/ 为/ 生产力/, 是/ 这个/ 开发区/ 的/ 特点。
과학연구/ 성과/ 신속/ 전환/ 하다/ 생산력/, 이다/ 이/ 개발구/ 의/ 특징
(과학연구성과를 생산력으로 신속하게 전환한 것은 이 개발구의 특징이다.)

예문 10: 学生们/ 来到了/ 操场/, 高高兴兴地。
학생들/ 오다/ 운동장/, 즐겁게
(학생들은 즐겁게 운동장으로 모였다.)

예문 11: 美国/ 总统/ 布什/ 上/ 周三/ 回应/ 了/, 罗夫/ 泄密/ 事件/,
미국/ 대통령/ 부시/ 지난/ 수요일/ 응하다/ ( )/ , 루어프/ 비밀루설/ 사건/,
他/ 说/ 在/ 调查/ 结束/ 前/ 不会/ 评价/ 罗夫。

그는/ 말하다/ ( )/ 조사/ 끝나다/ 전/ 안하다/ 평가하다/ 루어프
(미국 대통령 부시는 지난 수요일에 루이프 비밀루설 사건에 대하여 반응을 보였고, 그가 말하기를 조사가 끝나기 전에 루어프의 평가를 삼갈 것이라고 했다.)

예문 12: 统计/ 资料/ 表明/, 大连/ 对/ 韩/ 出口/ 达/ 一亿/ 多/ 美元。
통계/ 자료/ 나타내다/, 대련/ 대/ 한국/ 수출/ 달하다/ 1억/ 여/ 달러
(통계자료에 따르면 대련의 한국으로의 수출은 1억달러에 달한다.)

예문 13: 一九九四年/, 通用/ 在/ 中国/ 购买了/ 四千多万/ 美元/ 的/ 东西。
1994년/, 통용/ ( )/ 중국/ 구입하다/ 4천만/ 달러/ 의/ 물건
(1994년에 통용은 중국에서 4천만 달러의 물건을 구입했다.)

예문 1처럼 쉼표의 값이 (c,c)일 때, 이러한 절-절 간의 쉼표에서의 문장 분할 후 구문분석을 수행하면 보다 효율적이다. 이 쉼표를 분할 가능 쉼표라고 본 논문에서는 칭한다. 하지만 예문 2와 3처럼 쉼표의 값이 (p,p)일 때, 해당 쉼표의 좌우 세그멘트는 같은 절에 속한다고 볼 수 있다. 이러한 쉼표 위치에서의 문장 분할은 구문분석의 오류를 유발할 수 있다. 이 쉼표는 분할 불가 쉼표라고 칭하도록 하겠다.

쉼표의 값이 (p,c)와 (c,p) 일 경우, 양쪽 세그멘트의 구문적 관련성에 따라 해당 쉼표는 분할 가능할 수 있고, 분할 불가능할 수도 있다. 분할 가능한 (p,c)를 (p,c)-I으로, 불가능한 것은 (p,c)-II로 표기한다. (c,p)일 경우도 마찬가지로, 분할 가능한 (c,p)를 (c,p)-I로, 불가능한 것은 (c,p)-II로 표기한다.

(p,c)와 (c,p)에 해당하는 쉼표의 경우, 왼쪽 세그멘트의 구성성분(들)과 오른쪽 세그멘트의 구성성분(들)간에 의존관계가 있을 수도 있고 없을 수도 있다. 이러한 두 세그멘트의 구성성분 간의 의존관계를 왼쪽과 오른쪽 세그멘트 간의 구문적 관련성으로 본 논문에서는 정의한다. 쉼표의 두 세그멘트 사이에 구문적 관련성이 있는지 여부를 나타내기 위하여 아래 표 1과 같이 함수 Rel(left, right)와 Dom(left, right)를 정의한다.

예문 8의 두 번째 쉼표의 값은 (c,p)이고 두 세그멘트 사이에는 구문적인 의존관계가 없다. 즉 Rel(left,right)=0이다. 이 때 양쪽 세그멘트는 다른 절에 속한다고 볼 수 있고, 이 쉼표는 분할가능 쉼표, 즉 (c,p)-I에 속한다.

쉼표의 값이 (c,p)이고 쉼표가 연결하고 있는 왼쪽 세그멘트와 오른쪽 세그멘트 사이에 구문적인 의존관계가 있을 경우, 즉 Rel(left,right)=1 일 때, Dom(left,right)의 값에 따라 예문 9와 예문 10의 예와 같이 두 가지 경우로 나눌 수 있다. 예문 9에서의 쉼표는 Rel(left, right)=1이고 Dom(left,right)=right이다. 왼쪽 세그멘트 科研成果迅速转化为生产力(과학연구성과를 생산력으로 신속하게 전환하는 것)는 주어절(subject clause)이고, 문장의 메인헤드는 오른쪽 세그멘트에 포함되어 있는 是(~이다)이다. 절 세그멘트가 구 세그멘트의 의존소인 경우이고, 이러한 경우 두 세그먼트는 각각 다른 절에 속한다고 볼 수 있다. 따라서 이 쉼표위치는 분할 가능 위치이고, (c,p)-I에 속한다.

예문 10의 경우, 쉼표의 값은 (c,p)이고, Rel(left, right)=1이며 Dom(left,right)=left이다. 이와 같은 경우에는 오른쪽 세그멘트 高高兴兴地(즐겁게)는 왼쪽 세그멘트의 学生们来到了操场(학생들은 운동장으로 모였다.)을 수식해 주는 문법적 기능을 가지고 있다. 구 세그멘트가 절 세그멘트의 의존소인 경우이므로, 이 두 세그먼트는 하나의 같은 절 내에 속한다고 볼 수 있고, 이러한 쉼표위치는 분할 불가능한 위치이다. 즉 (c,p)-II에 속한다.

쉼표의 값이 (p,c)일 경우, (c,p)인 경우와 비슷한 현상을 나타내고 있다. 예문 11의 두 번째 쉼표, 예문 12의 쉼표, 예문 13의 쉼표의 값은 모두 (p,c)이다. 예문 11의 두 번째 쉼표 경우, Rel(left,right)=0이어서 의존관계가 없으므로 분할가능하다. 예문 12의 경우, Rel(left,

표 1 함수 Rel(left,right), Dom(left, right)의 정의

Rel(left, right)
• 쉼표의 왼쪽 세그멘트의 구성성분(들)과 오른쪽 세그멘트의 구성성분(들) 사이에 구문적 의존관계(한 쪽이 지배하고 다른 한 쪽이 지배를 받는 관계)가 있는지 여부를 체크한다. • 구문적 의존관계가 있으면 함수 Rel(left, right)값은 1이고, 구문적 의존관계가 없으면 함수 Rel(left, right)값은 0이다.
Dom(left, right)
• 쉼표의 왼쪽 세그멘트의 구성성분(들)과 오른쪽의 구성성분(들) 간에 구문적 의존관계가 있을 경우, 즉 Rel(left, right)=1일 때, 왼쪽이나 오른쪽 중에 어느 쪽이 헤드 단어를 포함하고 있는지를 보여준다. • 헤드 단어가 왼쪽 세그멘트에 포함되었다면 Dom(left, right)값은 left이고, 반대로 오른쪽 세그멘트에 포함되었다면 Dom(left, right)값은 right이다.

표 2 (c,p)값을 가진 쉼표의 분류

Rel값	Dom의 값 및 해당예문	(c,p) 분류	분할여부
Rel(left,right)=0	예문(8)의 두번째 쉼표	(c,p)-I	쉼표에서 분할가능
Rel(left,right)=1	Dom(left,right)=right 예문(9)		
	Dom(left,right)=left 예문(10)	(c,p)-II	쉼표에서 분할불가

표 3 (p,c)값을 가진 쉼표의 분류

Rel값	Dom의 값 및 해당예문	(p,c) 분류	분할여부
Rel(left,right)=0	예문(11)의 두번째 쉼표	(p,c)-I	쉼표에서 분할가능
Rel(left,right)=1	Dom(left,right)=left 예문(12)		
	Dom(left,right)=right 예문(13)	(p,c)-II	쉼표에서 분할불가

right)=1, Dom(left,right)=left이다. 절 세그멘트가 구 세그멘트의 의존소인 경우이므로, 두 세그멘트는 각기 다른 절에 속한다고 볼 수 있다. 따라서 분할가능 쉼표이며 (p,c)-I인 경우다. 예문 13의 경우는, Rel(left,right)=1, Dom(left,right)=right이다. 구 세그멘트가 절 세그멘트의 의존소인 경우이므로 이 쉼표는 분할불가능하며 (p,c)-II 에 속한다.

표 2와 표 3에서 쉼표의 분류를 정리해서 보여주고 있고, 예문 8에서 예문 13까지를 예제로 들었다.

쉼표의 값이 (c,c), (c,p)-I, (p,c)-I일 때, 해당 쉼표는 서로 다른 절 사이에 사용되므로 이러한 쉼표를 절-절 간의 쉼표(inter-clause comma, 이하 절절쉼표로 약칭)라고 하겠다. 쉼표의 값이 (p,p), (c,p)-II, (p,c)-II 일 때, 해당 쉼표는 같은 절 내에 사용된 경우이고 이러한 쉼표를 절 내의 쉼표(intra-clause comma, 이하 절내쉼표로 약칭)라고 하겠다. 예문 8의 두 번째 쉼표, 예문 9의 쉼표, 예문 11의 두 번째 쉼표, 예문 12번의 쉼표는 절절쉼표인 경우이고, 예문 10과 예문 13의 쉼표는 절내 쉼표이다.

4. 쉼표의 분류를 위한 자질 추출

자질 추출을 위하여 절절쉼표와 절내쉼표가 각각 어떤 유형들이고, 어떤 자질들이 필요한지 먼저 분석해 보고자 한다.

쉼표가 절절쉼표인지를 판단하기 위해서는 각 세그멘트가 절을 이루고 있는지 여부를 추정해야 한다. 절절쉼표 또는 절내쉼표를 구성하고 있는 쉼표의 특징은 표 4와 표 5에 기술되어 있다. 절이 되기 위해서는 서술어가 있어야 하고 서술어가 될 수 있는 것은 용언이므로, 쉼표 양쪽 세그멘트를 용언 출현여부에 따라 아래와 같이 분류할 수 있다.

표 4 절절쉼표를 구성하고 있는 쉼표의 특징

	왼쪽 세그멘트	오른쪽 세그멘트
(c,c)	-절	-절
(c,p)-I	-절	-동사구1[3]
	-절	-동사구가 아닌 구
(p,c)-I	-구	-절

표 5 절내쉼표를 구성하고 있는 쉼표의 특징

	왼쪽 세그멘트	오른쪽 세그멘트
(p,p)	-구	-구
(c,p)-II	-절	-구
(p,c)-II	-구	-절

- 쉼표 양쪽 세그멘트에 용언이 모두 출현하지 않음.
- 쉼표 한쪽 세그멘트에만 용언 출현.
- 쉼표 양쪽 세그멘트에 용언이 모두 출현.

쉼표가 절내쉼표인지 판단하기 위해서는 세그멘트가 구가 되는지 판단할 수 있어야 한다. 해당 세그멘트에 용언이 없는 경우는 세그멘트가 구임을 쉽게 판단할 수 있다.

중국어에서 절인지 여부를 판단하기 위해서는 용언의 자질을 확인해야 하고, 용언으로 사용될 수 있는 품사는 동사(VV, VC, VE)와 형용사(VA)이다(표 6 참조). 용

표 6 직접적 관련 자질 집합(용언으로 사용될 수 있는 품사)

자질명	직접적 관련 자질 집합
VC	문자열 내에 是(copula)가 등장할 때
VA	문자열 내에 형용사가 등장할 때
VE	문자열 내에 有There is..)가 등장할 때
VV	문자열 내에 동사가 등장할 때
CS	문자열 내에 종속접속사가 등장할 때

3) 예문 9의 오른쪽 세그멘트와 같이 '동사+목적어'의 형식을 취하는 동사구를 칭한다.

언 외에 종속접속사(CS)도 해당 세그멘트가 절인지 여부를 판단하는 데 중요한 정보를 줄 수 있다. 따라서 세그멘트가 절인지 판단하는 데 사용되는 직접적인 자질로서 표 6의 다섯 가지 자질을 사용한다.

세그멘트 내에 용언이 출현했을 경우, 해당 용언이 서술어인지 판단해야 한다. 예제 3을 예로 들어 설명하면, 쉼표의 양쪽 세그멘트에 존재하는 喜欢(좋아하다), 年轻(젊다), 美丽(예쁘다) 각각 VV(동사), VA(형용사), VA로서 직접적 관련 자질이지만, 모두 서술어로 기능을 하지는 않는다. 美丽는 뒤에 오는 조사 的(~한)과 연결되어 뒤에 오는 教师(교사)를 수식하고 있다. 중국어문장에서는 같은 단어가 문장 중에서 서로 다른 구문적 기능을 하더라도 형태적인 변형이 없기 때문에 품사정보만으로 서술어로 쓰였는지 명사의 수식어로 사용되었는지를 알 수 없다. 예로 중국어 毁灭(파멸하다)라는 단어는 문장 중 쓰임에 따라 진행형(파멸되고_있다)/과거형(파멸되었다)/관형어(파멸한)/목적어(파멸을)... 등으로 나누어질 수 있지만, 모두 동사 품사로 태깅된다.

중국어 문장에서 동사 자체는 형태적인 변형이 없지만 동사가 서술어로 사용될 때 자주 함께 쓰는 기능어들이 있다. 이러한 기능어가 동사가 서술어인지 판단하는 데 도움을 줄 수 있다.

그 예로 동사가 조사[4] '的'이나 '地'와 함께 쓰여 조사 오른쪽의 명사나 동사를 수식할 경우가 있다[18]. 이런 조사 앞에 나타난 동사는 서술어 기능을 하지 않는다. 조사 외에도 부사(副词), 전치사(介词), 연사(连词)들도 동사가 서술어로 쓰이는지 판단하거나 해당 세그멘트가 절인지 여부를 판단하는 데 도움을 줄 수 있다. 이러한 단어들을 모아서 간접적인 자질 집합을 구성했다. 실험에 사용된 간접적인 자질은 표 7과 같다.

정리하면 본 논문에서 사용한 자질은 아래와 같다.

4) 본 논문에서의 조사는 중국어 助词를 말하는 것이다. 한국어 조사와는 다른 뜻을 나타낸다.

표 7 간접적인 자질 집합

자질명	간접적 자질 집합
AD	문자열 내에 부사가 등장할 때
AS	문자열 내에 양상표시단어가 등장할 때
P	문자열 내에 전치사가 등장할 때
DE	문자열 내에 的이 등장할 때
DEV	문자열 내에 地가 등장할 때
DER	문자열 내에 得가 등장할 때
BA_BEI	문자열 내에 把 또는 被가 등장할 때
LC	문자열 내에 장소사가 등장할 때
FIR_PR	문자열 첫 번째 단어가 대명사일 때
LAS_LO	문자열 마지막 단어가 장소사일 때
LAS_T	문자열 마지막 단어가 시간사일 때
LAS_DE_N	문자열 마지막이 的+명사로 끝날 때
No_word	문자열 내 단어의 개수가 5보다 클 때
no_verb	문자열 내에 동사, 형용사가 없을 때
DEC	문자열이 관계절(relative clause)일 때
ONE	문자열 내에 단어의 개수가 단지 하나일 때

1. 절 인식에 직접적으로 관련성이 있는 자질: 서술어 관련 자질
2. 절 인식에 간접적으로 관련성이 있는 자질: 조사, 부사, 전치사, 연결사 등 서술어 판단에 도움을 줄 수 있는 자질

표 6과 표 7에서 사용된 자질들의 라벨은 중국어 CTB4.0[18]에서 사용된 태그들이다. 이러한 자질이 왼쪽 세그먼트에 나타나면 자질 라벨 앞에 'L_'을 붙이고, 오른쪽에 나타나면 'R_'을 붙여서 구분한다. 문자열에 출현한 자질타입의 값은 1로, 나타나지 않은 자질타입의 값은 0으로 설정을 한다. 표 8은 예문 13의 쉼표를 분류하기 위해 추출된 자질벡터를 보여주고 있다. 자질을 추출하기 위해 고려한 문자열 사이즈는, 쉼표 좌우 세그멘트 전체이다. 이렇게 한 문장에 대응하는 자질 벡터를 얻을 수 있고 이는 쉼표의 분류를 위한 기계학습에 사용된다.

표 8 예문 13의 자질벡터

L_VC = 0	L_VA = 0	L_VE = 0	L_VV = 0
R_VC = 0	R_VA = 0	R_VE = 0	R_VV = 1
L_CS = 0	L_AD = 0	L_AS = 0	L_P = 0
R_CS = 0	R_AD = 0	R_AS = 1	R_P = 0
L_DE = 0	L_DEV = 0	L_DER = 0	L_BA_BEI = 0
R_DE = 1	R_DEV = 0	R_DER = 0	R_BA_BEI = 0
L_LC = 0	L_DEC = 0	L_FIR_PR = 0	L_LAS_LO = 0
R_LC = 0	R_DEC = 0	R_FIR_PR = 0	R_LAS_LO = 0
L_LAS_T = 1	L_LAS_DE_N = 0	L_No_word = 0	L_no_verb = 1
R_LAS_T = 0	R_LAS_DE_N = 1	R_No_word = 1	R_no_verb = 0
L_ONE = 0			
R_ONE = 0			

5. 실 험

실험의 목적은 본 논문이 제시한 쉼표분류를 실행함으로써 중국어 구문분석에 미치는 영향을 알아보기 위함이다.

연구실에서 사용된 구문분석기는 중국어 의존관계 구문분석기로서 중국어 문장이 들어오면 <단어분할(segmentation) → 품사태깅 → 구묶음 → 구문분석> 등 4단계를 걸쳐 구문트리가 생성된다. 단어분할 단계는 입력된 중국어 문장을 단어 단위로 띄어쓰기하는 단계이고, 품사태깅 단계에서는 해당 단어의 품사정보를 설정해준다. 구묶음 단계에서는 NP(명사구), VP(동사구), PP(전치사구) 등 단어들이 묶여서 구를 형성하게 되고, 최종적으로 구문분석 단계에서는 각 단어들 사이의 의존관계를 규칙에 의거하여 설정하게 된다.

실험에 학습과 평가로 사용된 말뭉치는 Penn Chinese Treebank(이하 CTB) 4.0이다. CTB4.0은 펜실바니아 대학(University of Pennsylvania)에서 만든 중국어 트리뱅크로서 664,633한자, 404,156단어, 15,162문장으로 구성되었다. 중국어 트리뱅크로는 비교적 큰 규모를 가지고 있고, 중국어 구문분석시스템에 사용되어 구문분석기의 성능을 비교 평가하는 데 많이 사용되고 있다[23,24].

쉼표분류를 위한 학습 및 평가모델을 만들기 위해, CTB4.0 문장의 쉼표마다 절절쉼표인지 절내쉼표인지를 알아내어 태그를 달아줘야 한다. CTB4.0 말뭉치는 구문분석이 되어 있기 때문에, 쉼표의 분류를 구문분석정보를 이용하여 자동적으로 판단하였다.

쉼표의 분류를 위한 기계학습방법으로 SVM을 사용한다. 자질 추출 및 실험결과에 대한 자세한 설명은 아래와 같은 순서로 설명하도록 한다. 5.1에서는 정확한 쉼표 분류를 위한 SVM 커널함수 선택, 자질추출을 위한 문자열의 입력범위 등에 관련된 실험을 보여준다. 5.2에서는 본 논문에서 제안하는 쉼표분류과정을 먼저 실행하고 중국어 구문분석을 한 결과와 중국어 구문분석기만 실행한 결과를 비교하여 보여주고자 한다.

5.1 쉼표분할 실험

아래 실험들은 CTB를 10개 그룹으로 나누어, 10중 교차타당성(10-fold validation)으로 실험 평가한 결과이다. 실험에서 자질을 추출하기 위해 고려한 문자열의 윈도우 사이즈는 쉼표를 기준으로 왼쪽과 오른쪽의 전체 문자열이다.

평가는 절내쉼표 정확률(Intra-P)과 재현율(Intra-R), 절절쉼표의 정확률(Inter-P)과 재현율(Inter-R), 절내쉼표의 F-값(Intra-F), 절절쉼표의 F-값(Inter-F), 전체 쉼표의 F값(Total-F)으로 실험의 결과를 평가했다. Fβ=1/2의 값을 주었다.

5.1.1 커널함수 선택에 관한 실험

먼저 쉼표의 왼쪽 세그멘트와 오른쪽 세그멘트의 모든 단어들을 자질 추출을 위한 입력문자열로 설정하고, SVM의 커널함수로 Linear, Polynomial, RBF로 선택하여 각각 실험을 했다. Polynomial 커널함수를 사용했을 때는 파라미터 d의 값을 2와 3으로 했고, RBF 커널함수를 사용했을 때는 파라미터의 값은 0.5, 1.5, 2.5, 3.5로 하여 각각 실험을 하였다. 실험 결과는 표 9와 같다.

표 9에서 보여준 것과 같이 RBF 커널함수의 파라미터 값을 1.5로 했을 경우, 제일 좋은 성능을 보여 주었다. 아래에서 보여지는 실험은 모두 이 커널함수에 파라미터의 값을 1.5로 한 결과이다.

5.1.2 자질추출을 위한 입력 문자열 범위(window size) 설정 및 입력 문자열 위치 설정

자질추출을 위한 문자열의 범위를 설정하기 위하여 쉼표와 가까운 좌우 단어 1개씩, 2개씩, 3개씩, 4개씩, 5개씩 선택하여 각각 실험했다. 표 10에서 Win n-n은 자질 추출시 쉼표의 좌우로 쉼표와 가까운 단어 n개씩의 단어를 선택한 실험을 뜻한다. Win3-3일 때, 제일

표 9 커널함수 및 파라미터 선택에 관한 실험 결과

Kernel function	Intra-P	Intra-R	Inter-P	Inter-R	Intra-F	Inter-F	Total-F
Linear	74.22%	77.87%	72.52%	70.61%	76.00%	71.56%	73.14%
Polynomial d=2	79.84%	81.15%	84.51%	83.77%	80.49%	84.14%	82.86%
Polynomial d=3	78.57%	81.15%	88.39%	86.84%	79.84%	87.61%	84.86%
RBF γ = 0.5	78.46%	83.61%	88.64%	85.53%	80.95%	87.05%	84.86%
RBF γ = 1.5	78.69%	78.69%	**89.04%**	**89.04%**	78.69%	**89.04%**	**85.43%**
RBF γ = 2.5	80.62%	85.25%	88.24%	85.53%	82.87%	86.86%	**85.43%**
RBF γ = 3.5	79.41%	**88.52%**	85.05%	79.82%	**83.72%**	82.35%	82.86%

표 10 자질추출을 위한 입력 문자열 범위설정 실험 결과

	Win1-1	Win2-2	Win3-3	Win4-4	Win5-5
Total-F	66.01%	75.36%	82.86%	80.98%	78.93%

좋은 성능을 보였다.

다음으로 쉼표에서 거리가 먼 단어들을 자질 추출 대상으로 설정해서 실험했다. 표 11에서 Win m-n은 쉼표의 왼쪽 세그멘트의 시작 단어 m개를 선택하고, 쉼표의 오른쪽 세그멘트의 끝 단어 n개를 선택한 경우이다.

표 11에서 보여준 것 같이 Win2-3에서의 성능은 Win3-3인 경우보다 좋았다. 이로부터, 쉼표와 인접하고 있는 세그멘트에서 왼쪽 세그멘트의 첫 2단어, 오른쪽 세그멘트의 마지막 3단어가 도움을 줄 수 있음을 알 수 있다. 또한 이 결과는 표 10에서 쉼표좌우의 전체 문자열을 대상으로 실험한 결과보다 더 좋은 성능을 보이고 있다. 다시 정리하면, 쉼표위치에서의 분할여부를 선택하는 데 있어서, 쉼표의 좌우 가까운 단어들보다는, 왼쪽 세그먼트의 처음단어들과 오른쪽 세그멘트의 마지막 단어들이 더 유용한 정보를 주고 있음을 알 수 있다.

5.1.3 품사정보만을 이용한 실험

마지막으로 품사정보만을 자질정보로 사용한 경우 실험을 해보았다. 이 실험에서 자질추출을 위해 쉼표 좌우 각각의 전체 문자열을 대상으로 하였다. 결과는 표 12와 같고, win2-3의 결과보다 훨씬 나쁜 성능을 보이고 있다. 즉 품사정보 자체만으로는 쉼표의 분류를 판단하는 데 충분하지 못함을 보여주고 있다.

5.2 쉼표 분류 후 구문분석 성능 평가

쉼표분류를 통한 문장 분할의 목적은 구문분석에 도움을 주고자 하는 것이다. 다음 실험으로는 본 논문에서 제안한 문장 분할 방식을 실제 구문분석기에 적용했을 때, 구문분석기 성능에 얼마나 도움이 되는지 실험했다. CTB 말뭉치 중 앞서 쉼표분할 테스트용으로 사용된 문장들을 구문분석기 입력문장으로 사용했다. 실험 평가방법은 아래와 같다[17].

의존관계 정확률 = 정확하게 분석된 의존관계 수 / 전체 의존관계 수

문장분할 후 구문분석은 앞서 3절에서 소개했던 3단계 구문분석 과정을 거친다. 표 13에서 문장분할 후 구문분석기를 실행한 성능이 구문분석기만 실행한 성능보다 5.6%의 향상이 있음을 보여주고 있다. 또한 절절쉼표 위치에서 문장을 분할함으로써 구문분석의 복잡도를 크게 줄일 수 있어, 구문분석기의 성능을 향상시킬 수 있었다. 쉼표는 장문에서 특히 많이 등장하므로, 장문의 구문분석에서 쉼표위치에서의 문장분할을 수행하면 성능향상에 더욱 효과적이다.

표 13 기존구문분석기와 분할을 적용한 구문분석기의 성능비교

	기존 구문분석기	문장분할 + 구문분석기
의존관계 정확률	73.8%	79.4%

6. 기존연구와의 비교 및 결론

6.1 기존연구와의 비교

자연언어처리 커뮤니티에서 문장기호에 큰 관심을 기울이지 않아, 관련된 연구가 많지 않다. Shiuan과 Ann [22]은 영어복합문의 분할을 대상으로 하여 본 논문과 비슷한 방법을 시도했다. Shiuan 과 Ann[22]은 쉼표를 포함한 주변단어들이 문장에서 어떻게 사용되었는지 파악하여 쉼표위치에서의 분할 여부를 판단한다. 이 방법은 본 논문에서 제안하는 방법과 유사하지만, 아래와 같은 두 가지 차이점이 있다.

Shiuan과 Ann[22]은 쉼표의 여러 사용법 중에서 2가지 유형만을 판단하고 있지만, 본 논문에서는 쉼표의 전반 사용법에 대하여 분석을 했다.

Shiuan과 Ann[22]은 쉼표 외에 기타 접속사와 종속전치사 또한 문장 분할의 후보위치로 선정했다.

비록 언어가 다르고 실험한 데이타가 서로 다르지만, Shiuan과 Ann[22]과 본 논문에서는 둘 다 장문분할을 통하여 구문분석의 애매성을 해소하고자 하였다. Shiuan과 Ann의 시스템은 문장분할을 수행한 결과, 기존 영어 구문분석기의 성능보다 4% 향상된 구문분석 결과를 보였다[22]. 또한 본 논문에서 제안한 문장분할 방법은, 문장분할을 수행하지 않은 본 연구실의 중국어 구문분석기보다 5.6%의 성능 향상을 보였다. 대상언어도 다르고 구문분석방법도 다르기 때문에 [22]의 문장분할 방법과 객관적인 비교는 불가능하나, 본 논문에서 제안한 문장분할 방법에 의한 중국어 구문분석의 성능향상은 기존의 문장분할 방법에 의한 영어구문분석기의 성능향상

표 11 자질벡터 추출 window size 설정에 관한 실험 결과

Word Window	Intra-P	Intra-R	Inter-P	Inter-R	Intra-F	Inter-F	Total-F
Win3-3	80.45%	87.70%	84.33%	80.26%	83.92%	82.25%	82.86%
Win2-3	85.60%	87.70%	88.00%	86.84%	86.64%	87.42%	**87.14%**

표 12 품사정보만을 자질로 사용한 실험결과

성능 / 자질벡터	Intra-P	Intra-R	Inter-P	Inter-R	Intra-F	Inter-F	Total-F
품사정보	75.42%	72.95%	80.60%	82.02%	74.17%	81.30%	78.86%

수치보다 높았다.

구문분석을 위한 쉼표의 자동분류 과정이 추가되었지만, 구문분석의 수행시간이 길어지지는 않았다. SVM분류기는 분류결과를 빨리 도출하는 기계학습기이고, 분류를 위한 학습은 구문분석 이전에 미리 수행하기 때문이다. 오히려 쉼표분류 모델을 도입함으로써 문장이 분할되어 구문분석이 수행되므로 구문분석의 복잡도가 줄어들기 때문에, 계산적 부담을 줄인다. 따라서 본 논문에서 제안하는 쉼표분류 모델을 구문분석 전단계에 도입함으로써 발생하는 악영향은 없다.

6.2 에러분석 및 향후계획

쉼표를 구성하고 있는 양쪽 세그멘트에 모두 용언이 출현한 문장과 모두 출현하지 않은 문장에서의 절절쉼표 및 절내쉼표 판단은 정확했다. 하지만 한 쪽에서만 쉼표가 등장한 경우, 특히 쉼표의 값이 (c,p)-II(예문 10 참조)일 경우, 분류 오류가 많이 발생했다.

그 원인으로는 쉼표의 쓰임상 쉼표의 값이 (c,p)이거나 (p,c)일 때, 각각 (c,p)-I와 (c,p)-II, (p,c)-I와 (p,c)-II로 나눌 수 있지만, 문장의 표면적인 구성으로 보았을 때 각각의 특징을 알 수 없기 때문이다. 쉼표의 값이 (c,p)일 경우, 실제 (c,p)-I이어서 절절쉼표인 경우가 80% 이상이어서, 실제 (c,p)-II인 경우도 SVM이 기계학습 후 절절쉼표로 잘못 판단하는 경향이 크다.

향후계획으로는 문장 표면적인 구성으로 특징이 나타나지 않는 (c,p)-I와 (c,p)-II, (p,c)-I와 (p,c)-II를 분별할 수 있는 방법을 연구할 것이다.

6.3 결 론

중국어는 고립어로서 문장을 이루고 있는 단어들 사이에 띄어쓰기가 없어 문장분할의 적절한 위치를 찾는데 어려움이 있다. 하지만 중국어에서 쉼표가 타 언어보다 많이 사용되고 있으므로 쉼표가 장문분할에 중요한 정보를 줄 수 있다는 판단 아래 쉼표의 사용법을 알아보았고, 쉼표를 절절 쉼표와 절내쉼표 두 가지로 분류하는 기준을 설정했다 그리하여 본 논문에서는 중국어 문장에서 쉼표가 절절 쉼표인지 절내 쉼표인지 SVM에 의해 자동적으로 분류한 후, 판단결과에 따라 쉼표위치에서 문장 분할하는 방법을 제안했다.

제안한 쉼표 분류에 따라 문장분할한 경우 기존 구문분석의 성능을 5.6% 향상시킬 수 있었다.

참 고 문 헌

[1] Roger Levy and Christopher Manning, "Is it harder to parse Chinese, or the Chinese Treebank?," Proc. of the 41st meeting of the Association for Computational Linguistics, pages 439-446, 2003.

[2] Shui-fang Lin. "study and application of punctuation,"(*in Chinese*). People's Publisher, P.R.China. 2000.

[3] B. Jones. "What's the point? A(computational) theory of punctuation," PhD Thesis, Centre for Cognitive Science, University of Edinburgh, Edinburgh, UK, 1996.

[4] R.L. Hill. "A comma in parsing: A study into the influence of punctuation (commas) on contextually isolated "garden-path" sentences," M.Phil dissertation, Dundee University, 1996.

[5] X. Carreras, L. Marquez, V. Punyakanok, and D. Roth. "Learning and inference for clause identification," Proc. of 13th European Conference on Machine Learning, Finland, pages 35-47, 2002.

[6] V.J. Leffa. "clause processing in complex sentences," Proc. of 1st International Conference on Language Resources and Evaluation, Spain, pages 937-943, 1998.

[7] E.F.T.K. Sang and H.Dejean, "Introduction to the CoNLL-2001 shared task: clause identification," Proc. of 5th Conference on Computational Natural Language Learning, pages 53-57, 2001.

[8] S. Kim, B.Zhang and Y. Kim. "Learning-based intrasentence segmentation for efficient translation of long sentences," *Machine Translation*, Vol.16, no.3, pages 151-174, 2001.

[9] B. Jones. "Towards testing the syntax of punctuation," Proc. of 34th meeting of the Association for Computational Linguistics, pages 363-365, 1996.

[10] M. Bayparktar, B. Say and V. Akman, "An analysis of English punctuation: the special case of comma," *International Journal of Corpus Linguistics*, Vol.3, no.1, pages 33-57, 1998.

[11] B. Jones, "Exploring the role of punctuation in parsing natural text," Proc. of COLING-94, pages 421-425, 1994.

[12] M. J. Collins, "Head-driven Statistical Models for Natural Language Parsing," Ph.D. thesis, University of Pennsylvania, Philadelphia, 1999.

[13] Briscoe, E. and J. Carroll, "Developing and evaluating a probabilistic LR parser of part-of-speech and punctuation labels," Proc. of the 4th ACL/SIGPARSE International Workshop on Parsing Technologies, Prague, Czech Republic, pages.48-58, 1995.

[14] P.L. Shiuan and C.T.H. Ann. "A divide-and-conquer strategy for parsing," Proc. of the ACL/SIGPARSE 5th international workshop on parsing technologies, Santa Cruz, USA, pages 57-66, 1996.

[15] B. Say and V. Akman, "current approaches to punctuation in computational linguistics," Computers and the Humanities, Vol.30, no.6, pages 457-469, 1997.

[16] Geoffrey Nunberg. "the linguistics of punctuation," CSLI lecture notes. No. 18, University of Chicago Press, 1990.

[17] M.Y.Kim, S.J. Kang, J.H. Lee. "Resolving

ambiguity in Inter-chunk dependency parsing," Proceedings of the sixth Natural Language Processing Pacific Rim Symposium, Tokyo, Japan, 2001.
[18] N. Xue and F. Xia. "The bracketing Guidelines for the Penn Chinese Treebank(3.0)," Technical Report. 00-08, University of Pennsylvania, IRCS Report, 2000.
[19] V N. Vapnik. "The nature of statistical learning theory," Springer-Verlag New York, Inc., New York, NY, 1995.
[20] H. Yamada and Y. Matsumoto, "Statistical Dependency Analysis with Support Vector Machines," IWPT03, pages 195-206 2003.
[21] T.Joachims. "Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning," B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999.
[22] P.L. Shiuan and C.T.H. Ann. "A divide-and-conquer strategy for parsing," Proc. of the ACL/SIGPARSE 5th international workshop on parsing technologies, Santa Cruz, USA, pages 57-66, 1996.
[23] D.M.Bikel and D.Chiang. "Two statistical parsing models applied to the Chinese Treebank," Proc. of the NAACL-ANLP workshop of Second Chinese Language Processing Workshop, pages 1-6, 2000.
[24] R.Levy and C.D.Manning. "Is it Harder to Parse Chinese, or the Chinese Treebank?," Proc. of the ACL, 2003.

김 미 훈
1993년 중국연변대학교 수학과 학사 2004년 포항공과대학교 정보통신대학원 석사. 2005년~현재 포항공과대학교 컴퓨터공학과 박사과정. 관심분야는 자연언어처리, 중-한(한-중) 기계번역, 구문분석 등

김 미 영
1999년 포항공과대학교 컴퓨터공학과 학사. 2005년 포항공과대학교 컴퓨터공학과 박사. 2006년~현재 성신여자대학교 컴퓨터정보학부 전임강사. 관심분야는 자연언어처리, 구문분석, 기계번역, 정보검색 등

이 종 혁
1980년 서울대학교 수학교육학과 학사 1982년 한국과학기술원 전산학과 석사 1988년 한국과학기술원 전산학과 박사 1989년~1991년 일본전기(NEC) 중앙연구소 초청연구원. 1991년~현재 포항공과대학교 컴퓨터공학과 교수. 1998년~1999년 미국 CRL/NMSU(뉴멕시코주립대학) 방문교수. 관심분야는 자연언어처리, 기계번역, 정보검색 등