SVM 모델을 이용한 중국어 장문 분할*

포항공대 정보통신대학원 전자컴퓨터공학부,[1] 첨단기술연구 정보센터[2]
김미훈[1†] · 김미영[2] · 김동일[2**] · 이종혁[2]

Segmentation of Chinese Long Sentence Using Support Vector Machine

Meixun Jin,[1†] Mi-Young Kim,[2] Dongil Kim,[2] Jong-Hyeok Lee[2]
Dept. of Graduate School for Information and Technology,[1] POSTECH, Pohang
Div of Electrical and Computer Engineering,[2] POSTECH, Advanced Information Technology Research Center(AITrc), Seoul, Korea

요 약

문장이 길면 구문분석의 정확률이 크게 낮아진다. 따라서 장문을 분할하여 분석하면 구문분석의 복잡도를 크게 줄일 수 있어 정확률 향상에 크게 기여할 수 있다. 특히, 중국어는 고립어로서, 교착어나 융합어와 비교할 때 자연언어처리에 도움을 줄 수 있는 굴절이나 어미정보가 없어 구문분석에 어려움이 더욱 많다. 반면, 중국어 문장에서는 쉼표를 비교적 많이 사용하고 있고 또한 쉼표의 쓰임이 정확하므로 구문 분석에 도움을 줄 수 있다. 본 논문에서는 쉼표가 많이 쓰이고 있는 중국어 문장에서 해당 쉼표위치 문장 분할가능여부를 Support Vector Machine을 이용 판단하여 정확률 88.61%의 높은 분할 성능을 보였다.

서 론

대부분의 CFG구문분석은 문장 길이가 길어짐에 따라 복잡도가 기하급수로 늘어나 정확한 구문분석 결과를 내는데 어려움이 많다. 따라서 문장을 효과적으로 분할하여 분석하면, 구문분석의 복잡도를 줄이며 정확률을 향상시킬 수 있다.

문장에서 분할 위치를 정할 때는 언어에 의존적인 특징을 이용해야 한다. 중국어에서는 한 단어가 문장의 주어로 쓰이든 서술어로 쓰이든 단어나 품사의 변형이 전혀 이루어지지 않고 있어, 단어나 품사정보를 이용해 분할 여부를 판단하기에는 많은 어려움이 있다.

중국어 문장에도 한국어나 영어 등 다른 언어처럼 쉼표를 사용하고 있다. 하지만 중국어는 쉼표의 쓰임이 타 언어와 비교할 때 많이 다르고, 한국어에 비해 더욱 활발히 사용되는 경향이 있다. 중국어 문장에서의 쉼표를 사용된 문맥에 따라 분류함으로써 구문분석에 큰 도움을 줄 수 있다.

본 논문에서는 쉼표의 사용된 문맥에 따른 분류를 먼저 살펴보고 이들 분류에 따른 각각의 위치에 문장분할이 적절한지의 여부를 Support Vector Machine(이하 SVM으로 약칭)을 이용해 판단하는 방법을 제안한다.

논문의 구성은 아래와 같다. 2장에서는 문장분 할, 문장부호처리(punctuation) 및 SVM을 이용한 자연언어 처리 등 기존연구를 살펴보겠다. 3장에서 는 중국어 쉼표가 사용되는 문맥에 대하여 소개하 고, 4장에서는 SVM을 이용한 학습, 5장에서는 실험 결과로 이어진다. 최종적으로 6장에서는 결론을 맺는다.

기존연구

1. 문장분할

이미 서론에 언급된 것과 같이 문장분할은 구문분석의 복잡도를 줄이는 좋은 방법이고, 이에 대하여 다양한 선행연구가 있었다.

*본 연구는 첨단정보기술 연구센터를 통하여 과학재단의 지원을 받았음.
**중국 길림성 연길시 연변과학기술대학 부교수
†E-mail : meixunj@kle.postech.ac.kr
E-mail : colorful@kle.postech.ac.kr
E-mail : dongil@kle.postech.ac.kr
E-mail : jhlee@kle.postech.ac.kr

먼저 분할의 단위들은 크게 절이거나 구, 또는 절보다 작은 특정 패턴 및 기타 특정 구조들로 정의될 수 있다. 분할에 사용되는 방법은 크게 규칙에 의한 분할 또는 기계학습 등을 이용한 통계적인 방법이거나 혼합 방식이다.

[Leffa ' 98]에서는 규칙 기반 알고리즘을 이용해 복문(complex sentence)의 절(clause)을 분할하고 해당 절이 문장에서 체언이거나 부사어 기능을 하고 있는지 판단을 한다.

[Sang '02]에서는 MBL(memory-based) 학습을 이용하여 기본명사구 인식(base NP identification), 기본구 인식(base phrase recognation), 절 인식(clause identification), 명사구 분석(NP parsing) 등을 했다.

[Orasan ' 00]에서는 기계학습으로 복문 절들의 구간을 인식하고, 규칙으로 결과를 향상 시키는 혼합방식을 제안했다.

[Kim ' 00]에서는 먼저 한 문장에서 분할 가능 후보들을 선정하고, 최대 엔트로피 모델을 이용하여 해당 후보들의 분할 가능 여부를 판단한다.

각각의 실험에 대하여서는 분할된 단위 및 실험에 사용된 말뭉치, 실험한 문장 수 등이 다르기에 상대적인 성능 비교를 할 수 없다. CoNLL ' 2001에서의 결과를 본다면, Carreras and Marquez가 결정트리로 정확률 84.82%의 제일 좋은 결과를 보여주고 있다[Sang ' 01].

2. 문장기호처리

지금까지 자연언어처리에서 쉼표 등 문장기호에 대하여서 많은 연구가 진행되지 못하고 있다. 자연언어처리에서 문장기호 처리에 대한 연구는 [numberg ' 90]가 출판된 이후부터 진행되기 시작했지만, 아직도 많은 시스템에서는 문장부호들은 자연언어처리에 정보를 줄 수 없다는 판단하에 처리를 하지 않고 있다. [Jones ' 94]에서 문장기호가 포함되어 있는 문법을 사용하여 구문분석을 한 결과와 문장기호를 사용하지 않은 문법을 사용한 구문분석의 결과를 비교하여 전자가 후자보다 좋은 결과를 보여 주는 것을 알아내고 문장부호는 구문분석에 도움을 줄 수 있다는 것을 설명했다.

[Bayraktar ' 98]에서는 영어에서의 쉼표를 그 쉼표가 사용된 문맥에 따라 쉼표 사용된 구문패턴 데이터베이스를 만들고, 이 패턴중 약 80%를 7개 큰 부류 와 16개 소부류로 분류하는 시도를 했다.

3. SVM과 자연언어처리

1) SVM의 소개

SVM의 원리는 분류 문제를 해결하기 위한 최적 분리 경계면(separating hyperplane)을 제공한다. 학습데이터 $\{(x_i, d_i), i=1, \ldots, N\}$ 가 주어졌을 때, xi는 두 클래스 중 하나에 속하며, di $\{-1, 1\}$는 해당 클래스를 표시하는 레이블의 역할을 한다. SVM은 각 클래스를 구분하는 최적의 분리 경계면을 구하기 위해 분리 경계면과 가장 인접한 점(support vector)과의 거리를 최대화한다. 최적의 선형 분리 경계면을 $f(x)=wtx+b$로 놓으면, support vector와 $f(x)$의 거리를 $\frac{1}{\|w\|}$로 나타낼 수 있다. SVM은 $\|w\|^2$를 최소화하여 분리 간격을 최대화하도록 하여 최적 분리면을 찾아낸다.

이 문제는 다음과 같은 블록 최적화(convex optimization) 문제가 된다.

$$Min \quad \frac{1}{2} w'w + \quad C\sum_{i=1}^{N}\xi_i$$

$$s.t. \quad d_i(w' x_i + b) \geq 1-\xi_i \quad for\ i=1...N$$

$$\xi i \geq 0 \ for\ all\ i$$

결과적으로 최적의 분리 경계면은 아래와 같다.

$$f(x)=sign\left(\sum_{i}^{l}\alpha_i y_i K(x_i, x) + b\right)$$

구체적인 것은 [Vapnik ' 95, ' 98]을 참조 바란다.

2) SVM모델의 자연언어처리에서의 응용

SVM모델이 뛰어난 분류 능력을 보여줌으로써 여러 분야에서 많이 사용되고 있고 좋은 성능을 보여 주고 있다.

[Ganapathiraju ' 98], [Sun ' 03]은 SVM모델을 음성인식이나, 정보추출 등에 적용한 예이다.

[Yamada ' 03]에서는 SVM을 이용하여 의존(dependency) 구문분석을 했다. [Yamada ' 03]에서 인접한 두 단어 사이에 의존관계가 있을 경우, 지배소가 오른쪽이면 'r' 로, 지배소가 왼쪽이면 'l' 로, 의존관계가 없을 경우에는 'none', 이렇게 3개 부류로 분류했다. Penn Treebank Section 02와 21로 학습을 시켰고, Section 23으로 실험한 결과 90.3%의 좋은 성능을 냈다.

중국어 쉼표사용 및 사용패턴

1. 중국어 쉼표 및 사용패턴

중국어 쉼표의 간단한 사용법은 Table 1과 같다. 더 구체적인 사용법에 대하여서는 [Hu ' 01]을 참조하기 바란다.

Table 1. 중국어 쉼표 사용법

1. 서술어나 절 사이
예) 他为了到美国留学，念了一年的英语 그는 미국에 유학하기 위하여, 영어를 1년 공부했다.
예) 虽然他努力学习，成绩却不理想 그는 비록 열심히 공부하지만, 성적은 좋지 못하다.
2. 주어나 주어절 뒤에
예) 他昨天晚上通宵，是因为今天有考试 그가 어제 저녁에 밤샌 것은, 오늘 시험 때문이었다.
3. 대등접속문 또는 대등접속구의 연결
예) 他昨天晚上通宵，是因为今天有考试 여자아이는 3살이고, 앵두 입술은 예뻤다.
예) 去北海公园看看花儿，划划船儿 북해공원에 가서 꽃도 보고, 배도 탄다.
4. 부사어(절) 뒤에
예) 按照法律规定，他无权继承这一财产 법에 따라, 그는 이 재산을 상속받을 권리가 없다.
5. 삽입어 뒤에

중국어 쉼표의 사용은 Table 1에서 보여 준 것 과 같이 크게 절과 절의 사이이거나, 절과 구(phrase), 구와 구 사이에 사용되고 있다. 쉼표가 사용된 패턴을 Table 2와 같이 정리했다.

Table 2. 중국어 쉼표의 사용 패턴

Phrase와 Phrase 사이	1. [sub]	NP(Subject) , VP 教过他的老师，都喜欢他 그를 가르쳤던 선생님들은 모두 그를 좋아한다.
	2. [obj]	VP , NP(Object)
	3. [top]	NP(Topic), NP
	4. [ad]	PP(Adverbial), NP PP(Adverbial), VP
	5. [ad_p]	VP, PP(Adverbial)
	6. [c_NP]	Conjunction NP, conjunction NP 不论年轻人，还是老年人，都喜欢他 젊은 사람이든 연세 있으신 분이든 모두 그를 좋아한다.
	7. [conj_np]	, conjunction, NP
	8. [CP]	NP, NP VP, VP
Clause와 Phrase 사이	9. [subj_cl]	Clause(Subject Clause), VP 实现南北韩的统一，是我们的愿望 남북한의 통일을 실현하는 것은, 우리의 바람이다.
	10. [obj_cl]	VP, Clause(Object Clause) 他了解到，现在的状况的确不好 그는 현재 상황이 좋지 못함을 알았다.
Clause와 Clause 사이	11. [Coor_cl]	Clause, Clause
	12. [subor_cl]	(Subordinate conjunction) Clause, (Subordinate conjunction) Clause 因为他平时不努力，所以成绩不理想 그는 평소에 열심히 하지 않아서, 좋은 성적을 내지 못했다.

2. 분할 가능 쉼표와 분할 불가 쉼표

Table 2에서 보여준 것과 같이 중국어 구문분석에서 쉼표위치로부터 구와 구의 구분은 자연히 이루어 지고 있다. 또한 절과 절 사이에 쉼표를 사용하는 것이 대부분인 경우이기에 본 논문에서는 중국문장에서 쉼표위치를 분할 가능 위치로 정하는 것이 바람직하다고 생각한다.

그러나 모든 쉼표위치에서 분할하는 것이 다 합리적인 것은 아니다. 분할 가능 위치를 적절 분할(safe segment [Kim '00])로 정의하고, 분할 불가능한 위치에 있는 쉼표는 부적절 분할(non-safe segment)로 정한다. 적절분할은 그 분할된 절 이나 구들 중에서 하나만 헤드가 없고, 나머지 분할 된 절이나 구에는 헤드가 모두 존재 경우를 말한다.

예) 我们认为按水平分班，更便于教师因材施教
(저희들은 능력에 따라 반을 나누는 것이, 선생님이 학생들 수준에 맞추어서 가르치기에 더 편리하다고 생각한다.)

이 예제에서 쉼표 앞쪽에 위치하고 있는 按水平(능력에 따라)와 分班(반을 나누는 것이)의 헤드는 쉼표 뒤쪽에 있는 便于(편리하다고)이다. 만약 위의 예제를 쉼표 위치에서 분할할 경우, 정확한 헤드를 찾아 주는 데 어려움이 있다. 문장이 길어지고 한 문장에 쉼표가 많이 사용될 경우, 여러 가지 조합 가능성이 있어, 정확한 쉼표위치 분할이 없을 경우 구문분석이 올바른 결과를 내는데 어려움이 따르게 된다.

Table 2에서 쉼표가 구와 구를 연결할 때 이러한 문제가 발생할 가능성이 있기에, 이러한 쉼표에 대하여서는 연결한 위치에서의 분할이 적절분할인지에 대한 판단이 필요하다.

중국어 장문 분할에 필요한 자질(Feature) 추출

장문 분할에서 쉼표의 위치가 적절 분할에 해당하는 패턴으로는 Table 2에서 1, 2, 3, 8, 9, 10, 11, 12에 해당 하고, 부적절 분할에 해당하는 패턴은 Table 2에서 4, 5, 6,

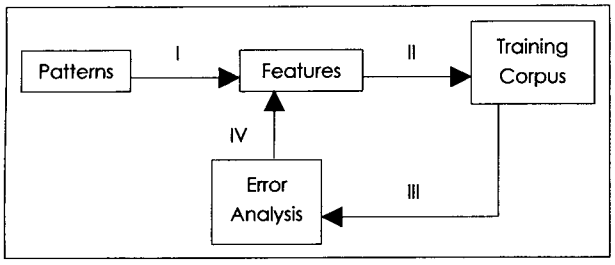

Fig. 1. 자질 추출 과정.

Table 3. Negative 자질

Pattern 4에서 뽑은 자질
Left
1)当|在|按照|根据|通过|除了|除|关于|为了|经过|作为|趁|通过|按|随着|自从|对|除|跟
2) Frame Structure[Zhou] chunk
Left 끝 단어
1) 的时候|时|后|以后|之后
2) 품사정보가 f(중국어 방향사)
Pattern 5에서 뽑은 자질
Right 첫 단어
1) 以|跟
Pattern 7에서 뽑은 자질
1) Conjunction, NP
Pattern 6 에서 뽑은 자질
1) Conjunction NP, conjunction NP

Table 4. Positive 자질

Pattern 1과 Pattern 9에서 뽑은 자질
Right 첫 단어
1)是|就是|predicate verb| 好不好
Left 중간 단어들
1) 能|會
Pattern 2과 Pattern 10에서 뽑은 자질
Left 마지막 단어
1) 是|predicate verb
Patten 3에서 뽑은 자질
1) word count of(Left) < 3
Right 첫 단어
1) 其中|部分
Pattern 12에서 뽑은 자질
Left, right 동시에 만족
1) subordinate conjunction, subordinate conjunction
2) 着|了|过 등 조사(助词)
Right 첫단어
1) subordinate conjunction
2) 免得|非|行不行

7의 경우이다.

자질 추출과정은 Fig. 1과 같은 단계로 진행된다.

Ⅰ. Table 2의 패턴에 따라, 각각 패턴판단에 도움을 줄 수 있는 자질들을 사람이 추출.

Ⅱ. 추출한 자질로 학습말뭉치에 적용하여 분류한다.

Ⅲ. 분류결과에 대하여 에러 분석한다.

Ⅳ. 에러분석 결과로 자질들을 upgrade한다.

Ⅱ, Ⅲ, Ⅳ단계는 반복적으로 자질에 더 변화가 없을 때까지 진행한다.

학습에 사용된 말뭉치는 중국 할빈공대대학(哈工大) 자연언어 연구실(http : //mtlab.hit.edu.cn)의 말뭉치와 중국 인민 일보(人民日報) 98년 말뭉치중 쉼표가 한 개씩 포함된 문장 575개를 선택했다.

이 문장들에서 먼저 쉼표위치 분할이 적절분할 인지 여부의 판단이 필요한, 구와 구를 연결한, 즉 Table 2에서 4, 5, 6, 7 패턴에 해당되는 문장들에서 아래 Table 3과 같이 negative 자질들을 확정했다. 표 중에서 Left는 쉼표 좌측 문맥을 의미하고, Right는 쉼표 우측부분의 문맥을 말한다.

Table 3에서 보여준 것 같이 negative 자질 중, Pattern 4의 경우는 쉼표 앞에 위치한 문맥과 연관이 많고, 쉼표 뒤쪽에 위치한 문맥의 영향은 적게 받는다. Pattern 5일 경우는 쉼표 뒤쪽 문장의 첫 단어로 판단을 할 수 있다.

Positive 자질은 쉼표로 연결된 앞뒤쪽 문장의 문맥에 함께 영향을 받게 된다. 특히 구문분석의 전 단계에서 진행을 하기에, 쉼표 양쪽에 절(clause)로 판단이 될 수 있는 용언이나 중국어 성구(成語, idiom)들을 중요한 자질로 선정 됐다.

Table 3, 4는 negative 또는 positive 특징을 강하게 보여주는 자질 들이다. 말뭉치 분석으로부터 적절 분할에 해당되는 쉼표가 80%이상임을 감안하여, 만약 문장에서 위의 자질 중 하나도 나타나지 않았을 경우, 초기값을 positive로 주었다.

실 험

실험으로 사용된 말뭉치는 중국 인민일보(人民日報) '98년 말뭉치 중에서 쉼표가 하나인 임의의 문장 920개를 선택했다. 먼저 본 연구실의 품사태깅(tagging)과 구묶음(chunking) 시스템을 이용하여 품사태깅 및 구묶음을 했다.

한 문장에 쉼표가 한 개인 경우, 쉼표가 사용된 문맥은 크게 종속어절, 부사어, 대등접속문 등이다. 한 문장에 쉼표가 여러 개일 경우에 쉼표가 사용된 문맥은 Table 2에서 Pattern 7과 Pattern 8 등이 비교적 많았다.

그래서, 실험의 정확률을 높이고 좋은 자질을 추출하기 위하여, 우선 한 문장에 쉼표가 1개인 경우 실험을 먼저 했다. 다음으로 쉼표가 2개, 3개, 4개인 경우로 하여 총 4

회의 실험을 수행 하였다.

1. 쉼표가 한 개인 경우

실험용 문장 920개 중에 positive가 721개 이고, negative가 199개이다. 실험에 SVM ^{light}를 사용했다([Joachimas '99]).

커널 함수로는 linear, polynomial, RBF로 각각 실험 했다. 실험의 평가방법은 다음과 같다.

$$\text{전체정확률} = \frac{(positive \text{ 로 바르게 인식된 개수} + negative \text{ 로 바르게 인식된 개수})}{\text{총 개수}}$$

$$precision = \frac{(Positive \text{ 또는 } Negative\text{로 바르게 인식된 개수})}{(Positive \text{ 또는 } Negative\text{로 인식된 개수})}$$

$$recall = \frac{(Positive \text{ 또는 } Negative\text{로 바르게 인식된 개수})}{(\text{문장중 } Positive \text{ 또는 } Negative\text{의 개수})}$$

Table 5. 쉼표가 1개일 때 실험결과 1(Polynomial)

Linear		
	Positive=721	Negative=199
Correct #	702	191
Detected #	710	210
Precision	98.87%	90.95%
Recall	97.36%	95.98%
전체정확률	97.07%	
Polynomial d=2		
Correct #	709	191
Detected #	717	203
Precision	98.88%	94.09%
Recall	98.33%	95.98%
전체정확률	97.83%	
Polynomial d=3		
Correct #	711	190
Detected #	720	200
Precision	98.89%	95%
Recall	98.61%	95.48%
전체정확률	97.93%	
Polynomial d=4		
Correct #	712	190
Detected #	721	199
Precision	98.75%	95.48%
Recall	98.75%	95.48%
전체정확률	98.04%	
Polynomial d=5		
Correct #	713	188
Detected #	724	196
Precision	98.48%	95.92%
Recall	98.89%	94.47%
전체정확률	97.93%	

Linear, polynomial 커널함수의 degree값이 각각 2, 3, 4, 5일 때의 실험결과는 Table 5와 같다. Table 5에서 Positive 또는 Negative로 인식된 개수를 Detect#로 표시하고 그 중 바르게 인식된 개수를 Correct#로 표시한다. Positive 분류에는 비교적 만족스러운 결과를 보여주었고, polynomial 커널함수의 degree를 향상에 따라 정확률도 향상되는 추세를 보였으며 degree값이 5일 때는 감소되었다.

RBF의 degree 값이 각각 0.5, 1.0, 2.0, 2.5, 3.5, 4,5, 5.5 일 때의 실험결과는 Fig. 2에서 보여주고 있다.

RBF 커널함수는 γ값이 향상됨에 따라, polynomial 커널함수와 다르게 positive 및 negative가 모두 큰 향상을 보였다. γ값이 5.5이상일 때는 성능이 더 이상 향상되지 않았다. Fig. 2와 함께 비교해 보았을 때, RBF 커널함수에 γ값을 3.5로 할 때, 전체정확률뿐만 아니라, positive 및 negative의 precision과 recall 모두가 가장 좋은 결과를 보여주고 있다. γ값을 4.5일 때, 전체정확률이 조금 더 향상되었다.

γ값이 4.5일 때에 오류가 9개였는데, 이중 3개는 품사 태깅 오류였고, 나머지는 분류에 유용한 특정 자질이 나타나지 않은 경우였다.

2. 쉼표가 2개 이상인 경우

앞에서 쉼표가 1개일 때 학습한 결과로 인민 일보 '98 말뭉치에서 임의의 쉼표가 2개인 문장, 쉼표가 3개인 문장, 쉼표가 4개인 문장을 각각 500개 실험했다.

커널함수는 앞 실험에서 제일 좋은 결과를 보였던 RBF degree 값 4.5를 사용했고, 실험결과는 Table 6와 같다.

Table 6에서 보여 준 것처럼 한 문장에 사용된 쉼표수가 늘어남에 따라 negative값이 증가되고 있다. 이는 한 문장

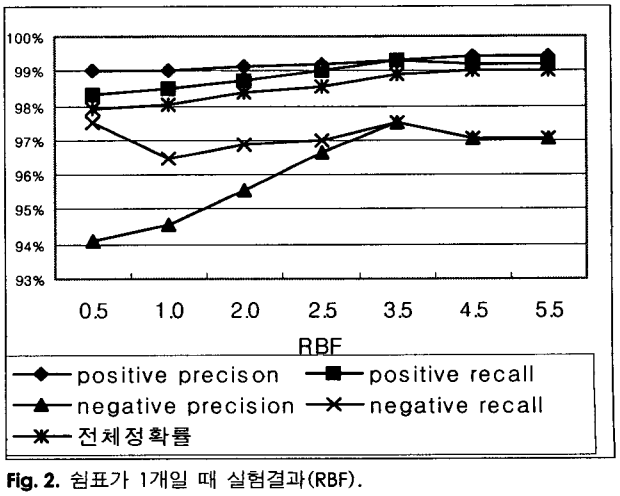

Fig. 2. 쉼표가 1개일 때 실험결과(RBF).

Table 6. 쉼표가 1개 이상인 실험결과

쉼표 개수	2	3	4
Positive로 인식된 개수	780	1020	1090
Negative로 인식된 개수	220	480	910
전체정확률	91.30%	89.20%	82.05%
문장내의 모든 쉼표가 정확하게 인식된 문장의 비율	84.60%	82.00%	53.00%

Table 7. 전체 실험 결과

쉼표 개수	전체정확률	문장내의 모든 쉼표가 정확하게 인식된 문장의 비율
5420	88.61%	83.14%

에 쉼표가 2개 이상일 때, 쉼표가 사용되는 문맥의 분포가 쉼표가 1개일 때와 다르기에, 결과도 큰 차이를 보이고 있다는 것을 설명한다.

Table 7에서는 전체 실험 데이터의 실험결과를 보여주고 있다.

3. 실험평가

기존연구들은 대부분이 영어를 대상으로 실험을 한 것이고, 선택된 실험 말뭉치도 많이 차이가 있어, 비교를 할 수 없다. CoNLL '01에서 제일 좋은 기록이 84.82%의 정확률이다. [Bayraktar '98]의 정확률도 80%였다. 정확률만을 비교해 보면, 본 논문의 제안 방법이 88.61%의 가장 높은 성능을 보였다.

결론 및 향후 계획

중국어 문장 분할에서 쉼표 위치를 분할 후보로 선정한다. 쉼표가 사용된 환경에 따라 분할 가능 쉼표와 분할 불가 쉼표로 SVM을 이용하여 시도했고, 정확률 88.61%의 높은 성능을 보였다.

쉼표가 1개 존재하는 문장들과 쉼표가 2개 이상 존재하는 문장들을 비교해 보면, Table 2에서 정의한 패턴들의 분포가 다르게 나타난다. 그리하여 쉼표가 1개 일 때 학습한 값으로 쉼표가 2개 이상 내포되어 있는 문장에서 쉼표 사용 문맥의 분류를 했을 때 큰 차이를 보여주었고, 쉼표의 개수가 늘어남에 따라 이러한 차이는 더욱 두드러지게 나타났다.

향후 계획으로는 쉼표가 2개 이상인 경우에 대한 정확률 향상에 중점을 두어 연구를 진행할 것이며 또한 한 문장에서 각 분할들 사이의 의존관계(dependency relation)를 통계적인 모델을 이용한 방법을 적용할 예정이다.

REFERENCES

Murat Bayraktar(1998) : *An Analysis of English Punctuation : The Special Case of Comma, International Journal of Corpus Linguistics 3(1) : 33-57*

Aravind Ganapathiraju(1998) : *Support Vector Machines for Speech Recognition, Proceedings of the International Conference on Spoken Language Processing, pp2923-2926, Sydney, Australia, November*

MingLiang Hu, DouHao, HanYuJiaoXueDe, YiGeNanDain, GuoJiHanYuJiaoXueXueShuYanTao HuiLunWenJi, *2001 (in Chinese)*

Joachims T(1999) : *Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning, B. Scholkopf and C. Burges and A. Smola(ed.), MIT-Press*

Bernard Jones(1994) : *Exploring the Role of Punctuation in parsing Real Text. In Preceedings of the 15th International Conference on Computational Linguistics(COLING-94) pp421-425, Kyoto, Japan*

SungDong Kim(2000) : *Reducing Parsing Complexity by Intra-Sentence Segmentation based on Maximum Entropy Model, 2000 joint SIGDAT conference on empirical methods in natural language processing and very large corpora*

Vilson J. Leffa(1998) : *Clause processing in complex sentences. In : "Proceedings of LREC '98", Granada, Espanha*

Geoffrey Nunberg(1990) : *The Linguistics of Punctuation, CSLI*

Constantin Orasan(2000) : *A hybrid method for clause splitting in unrestricted English texts, In : "Proceedings of ACIDCA '2000", Monastir, Tunisia*

Erik F.Tjong Kim Sang(2002) : *Memory-based Shallow Parsing, Journal of Machine Learning Research 2, pp559-594*

Erik F.Tjong Kim Sang(2001) : *Herve Dejean, Introduction to the CoNLL-2001 Shared Task : Clause Identification, Conference on Natural Language Learning*

Aisin Sun, Myo-Myo Naing, Ee-Peng Lim, Wai Lam(2003) : *Using Support Vector Machines for Terrorism Information Extraction, Proc. of the 1st NSF/NIJ Symposium on Intelligence and Security Informatics (ISI 2003), pp1-12, Tucson, Arizona, USA, Jun*

Vladimir N. Vapnik(1995) : *The Nature of Statistical Learning Theory. New York*

Vladimir N. Vapnik(1998) : *Statistical Learning Theory. A Wiley-Interscience Publication*

Hiroyasu Yamada(2003) : *Statistical Dependency Analysis with Support Vector Machines, International Workshop on Parsing Technologies*

MingZhou, *A Block-Based Robust Dependency Parser for Unrestricted Chinese Text*