한국어 서술어 구조의 확률적 정보

이승욱
한영석

수원대학교 정보미디어학과
windnc@hanmail.net yshan@suwon.ac.kr

Probabilistic Evidences for Korean Predicate Structures

Seung W. Lee
Young S. Han

Dept. of Information Media, University of Suwon

요 약

본 논문에서는 질의 응답 시스템에서 정답 추출을 위해 사용되는 표층 텍스트 패턴을 장거리 의존 문제에도 적용 가능하도록 확장하는 방법을 제안한다. 기존의 패턴 추출 시스템들의 패턴을 구성하고 있는 단어들간의 연속성과 불연속성에 대한 정보를 나타내도록 패턴 형태를 확장함으로써 장거리 의존 문제를 해결한다. 본 논문에서 제안한 형태의 패턴을 TREC-10의 질의를 이용해서 웹 데이터로 실험하여 정확도와 TREC의 평가 기준인 MRR을 사용해서 기존 시스템들과 성능을 비교했다.

1. 서 론

한 문장 내에서 주어와 목적어 등 범주를 파악하는 것은 내포문 등 다양한 모호성에 문제의 어려움이 있다. 문장의 의미이상의 계산모델이 연계되지 않는 상황에서는, 구문단계의 행위는 심하게 비선형적일 수 밖에 없으며, 모호성은 필연적이다. 구문단계의 언어처리는 여러 행태의 구문적 상관 관계를 반영한 다양한 구문문법과 확률문법을 포함한 휴리스틱으로 비선형지도를 작성하려고 해왔지만 넘을 수 없는 한계점이 존재한다. 여러 관점과 가정으로 접근하기 이전에 언어적 이벤트 자체를 그대로 관찰할 수 있는 시각에서 접근할 필요가 제기된다.

구문문법이나 형태소 모델 없이 최소한의 조사 휴리스틱만으로 용언의 하위범주 패턴의 현상을 설명하는 것을 연구의 목표로 삼고, 어절의 단순화, 어절위치의 상대적 정규화를 통해서 단순한 서술구조를 설정하고 여기에 코퍼스를 이용한 확률변수를 얻어 낸 후 서술구조가 갖는 특징에 대해서 연구하였다. 이 과정은 사람의 개입을 필요로 하지 않고 자동으로 대량의 자료들을 학습에 사용하며 문법을 생성해 낼 수 있다.

서술구조의 요소 중에서 주어와 목적어의 상대적 발생 어순을 조사하기 위해서, 어절을 주어(S), 목적어(O), 없음(N), 기타(X), 등 네 가지로 범주화 하고, 문장을 구성하는 어절의 개수 별로 네트워크를 각각 구성한다. 어절의 개수가 다섯 개가 넘는 경우는 어절 들을 다섯 개의 그룹으로 나누어서 각 그룹을 같은 네트워크의 단계에 넣어 네트워크를 구성을 한다. 각 노드의 전이 확률과 어휘적 정보를 가지고 마코프 모델을 구성하고, 주어와 목적어 등 4가지 범주의 패턴에 따른 베이지언 확률을 계산한다.

본 연구에서는 "하다"라는 용언을 기준으로 하여 실험하였다. 전처리 과정으로 코퍼스들의 문장의 분할 및 띄어 쓰기 교정 작업을 위한 연구[2,3,4,5]와 명사구 인식 연구[1] 등을 참조하였다. 구축된 네트워크에 대한 마코프 모델에 비터비 알고리듬을 이용하여 순위화 한 결과 중 10%안에 정확한 결과가 있을 확률이 약 72%를 나타내었고, 베이지언 네트워크로 계산한 패턴 확률은 무작위로 주어, 목적어를 선택했을 때 보다 168%높은 정확률을 보였다. 언어적 사전들이 구문적 패턴을 형성하는데 있어서 주요 기능어가 상대적인 위치에 스

스로 자리 잡는 점은 언어의 재귀적이고 내포적인 특성을 증명하는 부분이다. 베이지언 네트워크는 앞선 어절 그룹의 패턴이 다음 그룹의 어절 선택에 유효한 영향을 주는지를 나타내 준다. 실제로 마코프모델의 단순한 선행 사건에 대한 의존에서 보다 정교한 의존정보를 얻을 수 있었다.

2장에서 확률모델이 적용된 사례를 설명을 하고, 3장에서 베이지언 용언 모델을 다룬다. 4장에서 마코프 모델의 소개와 빈도 확률 모델의 이용을, 마지막 5장에서 결론을 짓는 것으로 구성되어 있다.

2. 확률모델의 적용 사례

언어 처리에 있어서 크게 규칙 기반에 기반한 방법과 확률에 근거를 둔 두 가지 방법으로 나눌 수 있다. 규칙 기반은 사람이 직접 정의한 규칙과 여러 사전 따위를 이용하는 방법이고, 일반적으로 높은 정확성을 보여준다. 규칙을 정의 하기가 쉽지 않은 경우도 있고, 한번 정의한 규칙을 관리 하는 것이 힘들다는 단점이 있다. 확률에 근거한 모델은 표층적으로 발생한 현상의 관찰이므로 특별히 사람의 개입을 필요로 하지 않지만 많은 양의 자료와 실험이 요구된다. 최대 엔트로피 모델, 은닉 마코프 모델, 베이지언 모델 등이 이에 해당한다.

특정 단어의 앞 뒤 특징에 따른 형태소 확률과 가중치를 계산하는 방법으로 문장을 파싱 하려는 연구가 있었다[6]. $W_1^n = w_1 \ldots w_n$ 단어가 주어졌을 때, 최적화된 묶음 태그 $P(C_1^n \mid W_1^{n)})$ 이 최대값을 가지게 하는 $C_1^n = c_1 \ldots c_n$ 의 순서를 찾는 과정에서 다음과 같은 식을 사용했다.

$$\arg\max_{c_i^n} P(C_1^n \mid W_1^n) \qquad \text{(수식2.1)}$$

$$\approx \arg\max_{c_i^n} \prod_{i=1}^{n} P(C_i \mid f_i) \qquad \text{(수식2.2)}$$

$$\approx \arg\max_{c_i^n} \prod_{i=1}^{n} \sum_{f_{ij} \in F, c_i \in C} \lambda_{ij} P(C_i \mid f_{ij}) \qquad \text{(수식2.3)}$$

f_i는 i번째 단어의 특성(feature)을 나타낸다. F는 모든 특성, f_{ij}는 i번째 단어의 묶음 태그C_j를 지지 해 주는 특성이며, λ_{ij}는 f_{ij}의 가중치이다.

3. 베이지언 용언 모델

베이지언 모델을사용하기 위해서 네트워크를 우선 구성한다. 대상 용언을 포함하고 있는 문장의 각 어절을 조사의 휴리스틱에 따라 범주화 한다. 범주는 크게 4가지로 나눈다. "은", "는", "이", "가" 등을 포함하고 있으면 주어 범주(S), "을", "를" 등을 포함하고 있으면 목적어 범주(O), 그 외에 기타 조사 범주(X), 조사 없음 범주(X)가 그것이며, 각 어절은 위의 네 범주 중 하나에 반드시 속한다. 범주화 된 어절들은 다섯 개의 영역으로 구분되고, 각 영역에서 다음 영역으로 전이 하는 분포를 구한다. 베이지언 모델은 순서에 상관없이 선행 사건에 대한 패턴적 의존관계를 모델 한다. 주어, 목적어가 연속된 어절에 위치 하지 않고 문장의 처음과 끝에 떨어져 있다고 해도 베이지언 모델에서는 이 확률을 구할 수 있다.

베이스의 정의는다음과 같다.

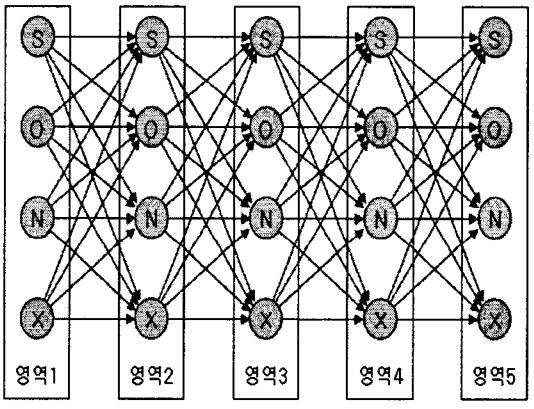
$$P(H_1 H_2) = P(H_1) * P(H_2 \mid H_1) \qquad (3.1)$$

$$P(H) = \sum_{i=1}^{N} P(H \mid E_i) P(E_i) \qquad (3.2)$$

$$P(H_K \mid E) = \frac{P(E \mid H_K) P(H_K)}{\sum_{i=1}^{N} P(E \mid H_i) P(H_i)} \qquad (3.3)$$

H_k (k=1,2, ,N) : 고려할 수 있는 상호배타적인 가설
E : 주어진 증거

베이스의 식을 이용하면 복잡한 네트워크 구조에서 노드의 확률을 구할 수 있다.



S:주어 O:목적어 N:없음 X:기타

[그림 3.1] 영역별 네트워크 구조

[표 3.1] 앞 영역의 조건에 따른 노드 확률

S	O	N	X	Cnt true	Cnt false	P
F	F	F	F	0	0	0.00
T	F	F	F	1200	9563	0.13
F	T	F	F	99	2971	0.03
T	T	F	F	322	3976	0.08
F	F	T	F	2091	17462	0.12
T	F	T	F	2722	23962	0.11
F	T	T	F	1102	11730	0.09
T	T	T	F	909	8313	0.11
F	F	F	T	1400	10299	0.14
T	F	F	T	1766	14390	0.12
F	T	F	T	710	7992	0.09
T	T	F	T	507	4833	0.10
F	F	T	T	4006	33196	0.12
F	T	T	T	1890	17193	0.11
T	T	T	T	1784	15560	0.11

영역1은 아무런 증거 노드를 가지지 않는다. 영역2부터 영역5까지의 각 노드들은 바로 앞 영역의 4가지 증거 노드를 가지므로 2^4개의 조합으로 표 3.1과 같은 베이지언 확률 테이블을 구성한다.

$$\arg\max_{i,j} \sum_{i=1}^{5} \sum_{j=1}^{5} P(S_i O_j)(i \neq j) \quad \text{(수식3.4)}$$

i,j는 영역을 나타낸다. 한 영역에서 주어, 목적어 모두 나올 수 없으므로 i,j 가 같을 경우는 고려하지 않는다. 예1은 수식3.4를 이용하여 실제 문장의 확률을 구하는 과정을 보여준다.

예1) 컴팩은 IBM PC가 수행한 모든 일을 하...
S1 N2 S3 X4 N5 O5

$$P(S_1,O_5)=P(S_1|O_5)*P(O_5)=0.019 \quad \text{(수식3.5)}$$

$$P(S_2,O_5)=P(S_2|O_5)*P(O_5)=0.016 \quad \text{(수식3.6)}$$

두 확률들은 O_5가 S_1과 S_2를 얼마나 지원하는지 나타낸다. 표3.2, 그림3.2와 표3.3, 그림 3.3는 각각 목적어가 증거 이벤트 일 때의 주어 이벤트 확률, 주어가 증거 이벤트 일 때의 목적어 확률을 계산한 결과를 보여준다.

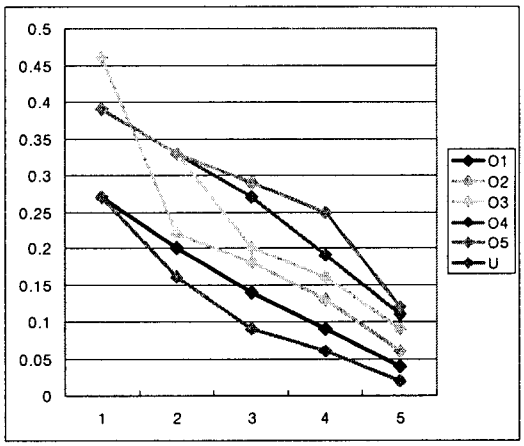
E : 증거 U : 전체

S_i : i번째 영역의 주어

O_i : i번째 영역의 목적어

[표 3.2] 증거 이벤트에 따른 주어 확률

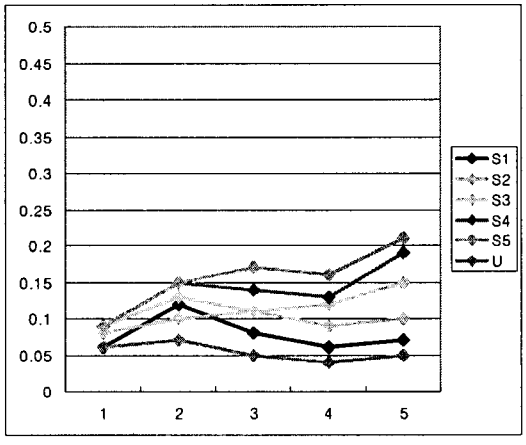
E	P(S_1|E)	P(S_2|E)	P(S_3|E)	P(S_4|E)	P(S_5|E)
U	0.27	0.16	0.09	0.06	0.02
O_1	0.27	0.20	0.14	0.09	0.04
O_2	0.46	0.22	0.18	0.13	0.06
O_3	0.39	0.33	0.20	0.16	0.09
O_4	0.39	0.33	0.27	0.19	0.11
O_5	0.39	0.33	0.27	0.25	0.12



[그림 3.2] 증거 이벤트에 따른 주어 확률 그래프

[표 3.3] 증거 이벤트에 따른 목적어 확률

E	P(O_1|E)	P(O_2|E)	P(O_3|E)	P(O_4|E)	P(O_5|E)
U	0.06	0.07	0.05	0.04	0.05
S	10.06	0.12	0.08	0.06	0.07
S_2	0.08	0.10	0.11	0.09	0.10
S_3	0.09	0.13	0.11	0.12	0.15
S_4	0.09	0.15	0.14	0.13	0.19
S_5	0.09	0.15	0.17	0.16	0.21



[그림 3.3] 증거 이벤트에 따른 목적어 확률 그래프

그림 3.2은목적어가 증거가 되었을 때 주어의 영역별

확률이다. 그림 3.3는 반대로 주어가 증거가 되었을 때 목적어의 영역별 확률이다. 그림 3.2에서는 전반부가 높고 뚜렷하게 하강되며, 그림 3.3에서는 후반부 영역의 목적어 확률이 전체적으로 증가하나 그림 3.2의 하강폭에 비해 그 증가 폭은 작으며 도중에 감소하는 경우도 있어 뚜렷한 특성을 반영한다고 보긴 어렵다. 목적어가 주어를 지지하는 정도가 주어가 목적어를 지지하는 정도보다 높고, 신빙성도 높다.

임의의 S와 O를 선택할 경우, 정확한 S, O일 확률은 $\frac{1}{n(S)n(O)}$ 이 된다. 아래 표에 P1이 이런 확률을 뜻한다. P2는 본 논문에서 제시한 베이지언 모델을 이용하여 선택한 결과의 정확성을 나타낸다. 소설, 설명문, 논문 등의 문서에서 주어와 목적어를 선택하는데 모호성을 가지고 있는 문장 135개를 대상으로 다음과 같은 결과를 계산하였다.

n(S) : 주어의 수 n(O) : 목적어의 수
S : 정답을 찾는데 성공 F : 정답을 찾는데 실패

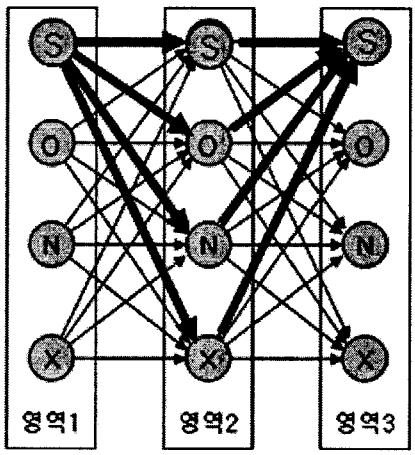
[표 3.3] 주어, 목적어 수에 따른 확률

n(S)	n(O)	S	F	비중	P1	비중*P1	P2	비중*P2	상승률
1	2	12	4	11%	0.50	0.0550	0.75	0.0825	150%
1	3	7	5	8%	0.33	0.0264	0.58	0.0464	175%
1	4	2	4	4%	0.25	0.0100	0.50	0.0200	200%
2	1	5	4	6%	0.50	0.0300	0.55	0.0330	110%
2	2	12	17	21%	0.25	0.0525	0.41	0.0861	164%
2	3	3	7	7%	0.16	0.0112	0.30	0.0210	187%
2	4	4	9	9%	0.12	0.0108	0.30	0.0270	250%
3	1	6	6	8%	0.33	0.0264	0.50	0.0400	151%
3	2	2	4	4%	0.16	0.0064	0.33	0.0132	206%
3	3	1	5	4%	0.11	0.0044	0.20	0.0080	181%
3	4	0	3	2%	0.08	0.0016	0.00	0.0000	0%
4	1	4	3	5%	0.25	0.0125	0.57	0.0285	228%
4	2	1	4	3%	0.12	0.0036	0.25	0.0075	208%
Sum		59	76	100%		0.0175		0.0295	168%

n(S)나 n(O) 중 하나라도 높은 경우는 잘 나타나지 않는 경우이다. 본 실험에서 n(S)=3, n(O)=4인 예제는 135개 예제 중 3번 밖에 없었다. 발생한 회수를 전체로 나눠서 비중을 계산하여 실험이 많이 이루어 진 항목일수록 신빙성을 높이게 했다. 무작위로 선택한 주어, 목적어가 맞을 확률에 비해 본 연구의 베이지언 모델을 적용 했을 경우 전체적으로 168%의 정확도가 상승하였다.

4. 빈도 확률 모델

어절을 주어(S), 목적어(O), 없음(N), 기타(X), 네 가지로 범주화 하고, 어절의 길이에 따라 네트워크를 각각 구성한다. 다섯 개 이상의 어절을 가진 문장은 서로 공유한다. 마코프 모델에 따라 각 노드의 전이 확률을 구하고, 어휘적 정보를 저장한다. 다섯 개 보다 많은 어절을 다섯 개 영역에 넣으면 한 영역에 두 개 이상의 어절이 포함된다. 이 때문에 선택의 문제가 생기고 모호성이 발생한다. 여러 가지 경로 중 최대 값을 구하기 위해 비터비 알고리즘을 사용한다.



[그림 4.1] 네트워크에서 입력 문장에 따른 영역의 활성화

계산하고자 하는 문장에 따라 경로가 다양하게 활성화 된다. 그림 4.1처럼 영역1에서 4개 경로 모두 활성화 되고 영역2에서 S_3로 가는 한 개의 경로만 활성화 되었다고 가정한다면, S_2, O_2, N_2, X_2는 S_1에서 전이 확률을 저장하고, 어휘 확률을 전이 확률에 곱한다. 영역2의 노드들은 모두 S_3로의 경로 하나만 가진다. S_3에서는 (S_1,S_2), (S_1,O_2), (S_1,N_2), (S_1,X_2) 경로를 거쳐 온 확률 중 최대값을 택하고 어휘 확률을 곱한다. 이 과정을 마지막 영역까지 수행한다.

공유하지 않는 네트워크1부터 네트워크4에서는 보다 간편한 아래와 같은 알고리듬으로 구한다.

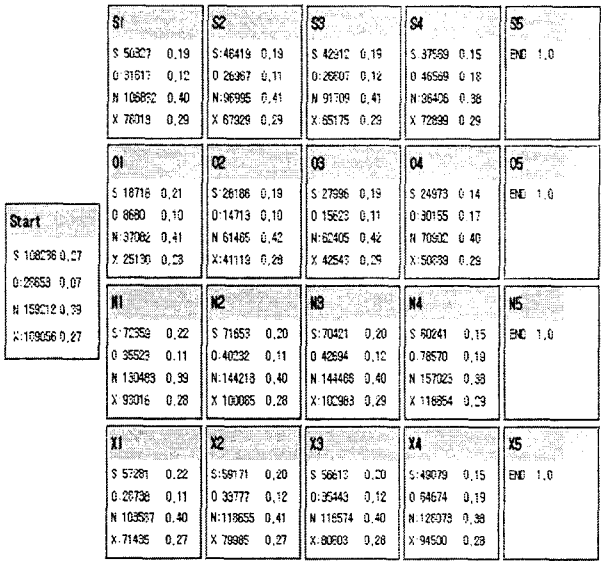
```
P = 1
For( i 1 to 어절 수 ) {
P1i = 이전 노드에서 범주i 노드로의 전이 확률
P2i = 범주I 노드에서 어절i의 어휘 확률
P = P * P1i
P = P * P2i
i영역의 범주i로 이동
}
```

본 논문에서는 "하다" 용언을 포함한 총 25만 문장을

사용하여 앞의 어절 수에 따라 다른 네트워크로 구성하였다. 어절 수가 6개부터 40개까지의 문장은 네트워크5에 같이 구성되기 때문에 네트워크5의 수치는 두드러지게 많다. 표 4.1에서 네트워크를 구성한 문장의 수를, 그림 4.2에서 네트워크5의 전이 빈도와 확률을 확인 할 수 있다.

[표 4.1] 네트워크 구성 문장

네트워크	1	2	3	4	5
문장수	17,931	17,202	17,956	16,168	183,397



[그림 4.2] 노드간 전이 확률

검증을 위한 실험은 소설, 설명문, 논문 등의 문서 중에서 직접 주어와 목적어를 표시한 문장 200개를 대상으로 하였다. 아래 표4.2는 비터비 알고리듬을 이용해 순위화한 경로들 중 영역 안에 정답이 나올 확률이다.

[표 4.2] 마코프 모델의 영역별 정답확률

순위(%)	정답확률	누적확률
0~10	73.2%	73.2%
10~20	12.1%	85.3%
20~30	6.8%	92.1%
30~40	2.2%	94.3%
40~50	1.0%	95.3%
50~60	2.1%	97.4%
60~70	1.0%	98.4%
70~80	1.6%	100.0%
80~90	0%	100.0%
90~100	0%	100.0%

순위가 낮아질수록 정답 확률이 떨어지는 결과를 보였고 이는 본 논문에서 제시한 비터비 알고리듬을 이용한 순위화의 효과가 있었음을 뜻한다.

5. 결 론

사람의 개입을 필요로 하는 규칙 기반의 방법이 언어처리 분야에서 오랫동안 연구되어 왔지만 훌륭한 성과를 이루지 못했다. 대량의 코퍼스 어휘들을 통해 자동으로 학습하는 보다 구축과 확장이 용이한 시스템 구축하였다. 이 시스템으로 여러 용언을 대상으로 정보 네트워크를 구축하여 복잡한 문장의 구조를 파악하는데 사용 될 수 있다.

마코프 모델을 이용해 어절 길이 별 네트워크를 구성하여, 노드마다 어휘 테이블을 구축한 후, 이 정보를 이용하여 최적 확률을 구하였고, 베이지언 확률로 범주의 패턴에 맞는 확률을 구하였다. 첫 번째 방법은 연속적인 확률을 이용하여 최적 경로를 구하기 위한 방법이고, 두 번째 방법은 경로와는 상관없이 패턴으로만 주어와 목적어를 선택하는 방법이다. 두 가지 방법 모두 임의의 주어, 목적어 선택 방법보다 높은 정확성을 보였다.

어절들을 발생 위치에 따라 5개의 그룹으로 구성된 네트워크는 마코프 모델과 베이지언 모델의 기본 골격으로 이용이 되었다. 그 결과 상대적인 어절의 위치에 대해서 현저하고 지속적인 패턴을 형성함을 알 수 있었다. 용언의 하위 범주들이 다양한 문장의 길이에도 불구하고 상대적 위치에서 구문적 패턴을 구성하며, 또한 예측 능력을 가지고 있음이 증명되었다.

한계점으로는 태깅 되지 않은 자료를 조사 리스트로만 어절을 범주화 시켰기 때 오류가 생길 수 있다. 단순한 어휘적 정보만을 이용한 어절들의 범주화는 정확률이 떨어질 수 밖에 없고 이는 형태소 분석기를 이용하지 않았기 때문에 필연적이다. 베이지언 모델에서는 최대값을 가지게 하는 패턴을 선택하게 되므로 패턴이 같을 때 항상 같은 결과를 도출하게 된다. 추가적인 연구로 마코프 확률모델과 베이지언 모델을 병행하여 보다 상황에 따라 적응 하게 할 수 있다. "하다" 동사들 이외에 다른 다양한 용언에서는 본 실험과는 다른 결과가 나올 것으로 예상되므로 용언마다 따로 구축 해볼 수 있다. 동사와 이에 해당하는 주어 목적어를 찾을 수 있다면 이들의 경계를 찾아내고, 이를 이용하는 선행된 파싱 연구들을 적용 시킬 수 있다[7].

참고 문헌

[1] 서충원, 최기선 "다단계 묶음을 이용한 명사구와 동사구의 의존관계 추출 시스템", 2004

[2] 강승식, "음절 bigram 특성을 이용한 띄어쓰기 오류의 인식", 한글 및 한국어 학회, 2000

[3] 정경미, "규칙을 기반으로 한한국어 텍스트에서의 문장 분할", 포항공과대학교 석사 학위논문, 1996.

[4] 정영미, 이재윤, "한국어 텍스트 처리를 위한 줄 경계 띄어쓰기 복원", 한국어정보관리학회 학술대회 논문집, 1999

[5] 강미영, 권혁철, "효율적인 문서처리를 위한 띄어쓰기 교정 기법 개선", 정보과학회 2003년 춘계학술대회

[6] Young-Sook Hwang, So-Young Park, Hoo-Jung Chung, Yong-Jae Kwak, Hae-Chang Rim, "Shallow Parsing By Weighted Probabilistic Sum", Proc. of the 2001 International Conference on Computer Processing of Oriental Languages

[7] [Abney91]S.P. Abney, "Parsing by chunks", R.C. Berwick, S.P. Abney and C. Tenny, editors, Principle-Based Parsing : Computation and psycholinguistics, Kluwer, Dordrecht, 1991