통계 정보를 이용한 한국어 자동 띄어쓰기 시스템의 성능 개선

최성자° 강미영 권혁철
부산대학교 컴퓨터공학과
{heya5°, kmyoung, hckwon}@pusan.ac.kr

Improving Korean Word-Spacing System Using Stochastic Information

Sung-Ja Choi° Mi-Young Kang Hyuk-Chul Kwon
Korean Language Processing Lab, School of Electrical & Computer Engineering
Pusan National University

요 약

본 논문은 대용량 말뭉치로부터 어절 unigram과 음절 bigram 통계 정보를 추출하여 구축한 한국어 자동 띄어쓰기 시스템의 성능을 개선하는 방법을 제안한다. 어절 통계를 주로 이용하는 기법으로 한국어 문서를 처리할 때, 한국어의 교착어적인 특성으로 인해 자료부족 문제가 발생한다. 이를 극복하기 위해서 본 논문은 음절 bigram간 띄어쓸 확률 정보를 이용함으로써, 어절로 인식 가능한 추가의 후보 어절을 추정하는 방법을 제안한다. 이와 같이 개선된 시스템의 성능을 다양한 실험 데이터를 사용하여 평가한 결과, 평균 93.76%의 어절 단위 정확도를 얻었다.

1. 서론

한국어 자동 띄어쓰기란, 띄어쓰기가 올바르게 되어있지 않은 문장을 올바르게 띄어쓴 문장으로 자동 수정하는 것이다.

이러한 자동 띄어쓰기는 정보 검색 시 띄어쓰기가 무시된 한국어 문장의 처리, 문자 인식 시 후처리 작업 중 경계 복원, 맞춤법 검사, 문자-음성 변환 등에 필요하다.

본 논문은 통계 기반 자동 띄어쓰기 시스템의 성능을 개선하는 방법을 제안한다. 시스템의 성능을 개선을 개선하기 위해서는 한국어의 어미 변형 특성 때문에 발생하는 어절의 자료부족 문제를 극복하기 위한 방법이 필요하다.

본 논문의 구성은 다음과 같다. 2장에서 통계기반 한국어 자동 띄어쓰기와 관련된 이전의 연구들을 살펴보고, 3장에서 본 연구의 이전 연구인 자동 띄어쓰기 시스템에 관해 간략히 설명한 후, 통계 정보를 활용한 성능 개선 방법을 제안한다. 4장에서 성능 실험 결과를 보이고, 5장에서 결론을 맺는다.

2. 관련 연구

자동 띄어쓰기 방법은 크게 통계 기반 방법[1,2,3]과 규칙 기반 방법[4]으로 나눌 수 있다. 통계 기반 방법은 띄어쓰기 시스템 구축이 쉬우며, 계산적으로도 부담이 적다. 그러나 통계 정보를 추출한 말뭉치의 특성에 따라 유사 분야의 문장들에 대해서는 우수한 성능을 보이지만 다른 분야에 대해서는 정확도가 떨어질 수 있다. 규칙 기반(지식 기반) 방법은 문서의 특성과 상관없이 비교적 일관된 성능을 보이며 중의성 해결에 우수한 성능을 보인다. 그러나 형태소 분석기에 의존적이며 지식베이스를 구축하고 유지 관리하는데 부담이 크다.

본 연구의 자동 띄어쓰기 시스템은 통계 정보를 기반으로 한다. 기존의 통계 기반 자동 띄어쓰기에 관한 연구들에 주로 사용된 통계 정보는 음절 N-gram이다. 강승식(2001)[2]은 음절 bigram 통계 정보를 이용하여 임의의 두 음절 사이에 공백이 삽입될 확률을 이용해 자동 띄어쓰기를 한다. 이 방법의 어절 단위 정확도[1]는 71.22%이다.[2] 이도길 외(2003)[3]는 음절 trigram 통계 정보를 이용한다. 이 연구는 다른 통계 기반 자동 띄어쓰기 방법에서 이전의 띄어쓰기 상태를 고려하지 않기 때문에 발생하는 문제점을 극복하기 위하여, 자동 띄어쓰기를 품사 부착과 같은 분류 문제로 간주하고 은닉 마르코프 모델을 확장한 띄어쓰기 모델을 제안한다. 이 모델은 이전의 2개 음절을 고려할 때 가장 좋은 성능을 나타낸다. 이 방법의 어절 단위 정확도는 93.06%이다. 부산대학에서 개발하고 있는 자동 띄어쓰기 시스템은 어절 통계 정보를 중심으로, 음절 bigram을 이용해 보완하는 방법을 취한다[5,6]. 통계 정보만을 이용했을 때, 이 시스템의 어절 단위 정확도는 90.5%이다.

사용하는 통계 정보별로 각각의 자동 띄어쓰기에 필요한 기억 공간을 비교해보면 <표 1>과 같다.[3]

<표 1> 사용하는 통계 정보별 필요한 기억 공간

	필요한 기억 공간
[illegible]	4.1MB
[illegible]	4.1MB + 25.1MB = 29.2MB
[illegible]	63.7MB

<표 1>에서 가장 작은 기억 공간을 차지하는 음절

1) 실험 성능은 시스템이 반환한 전체 어절 수 중에 시스템이 바르게 반환한 어절 수의 비율, P_{word}(Word Precision)과, 실험 데이터의 전체 정답 어절 수에 대한 시스템이 바르게 반환한 어절 수의 비율, R_{word}(Word Recall)로 나타낸다. '어절 단위 정확도'는 P_{word}을 말한다.

2) 강승식(2001)의 논문은 어절단위 정확도를 구해 놓은 것이 없으므로, 제시한 어절 단위 정확도는 이도길 외(2003)에서 동일한 조건으로 모델을 구현하여 실험한 결과이다.

3) 통계 정보를 추출한 말뭉치의 구성은 <표 2>와 같다.

bigram만을 사용하는 방법이다. 그러나 이 방법은 성능이 높지 못하다. 어절과 음절 bigram 통계 정보를 이용한 경우가 음절 trigram 통계 정보를 이용할 때보다 더 적은 기억 공간을 차지한다. 그러나 성능이 다소 낮다. 본 연구에서 이를 개선할 수 있는 방안을 모색한다.

3. 자동 띄어쓰기 성능 개선 방안

3.1. 통계 정보를 이용한 자동 띄어쓰기 방안

본 연구는 자동 띄어쓰기를 위해 어절 unigram와 음절 bigram을 사용한다. 따라서, 어절 unigram과 음절 bigram을 다음 말뭉치로부터 추출하였다.

<표 2> 통계 정보를 추출한 말뭉치의 구성

	[illegible]	[illegible]
(A) A신문의 2년분 [illegible]	1,564,070	18,961,771
(B) B신문의 1년분 [illegible]	650,539	9,432,258
(C) TV 뉴스 방송 [illegible]	409,962	5,249,855
총 합	1,950,068	33,643,884

추출한 통계를 이용하여 임의의 음절 연속이 독립된 어절로 존재할 확률 $P(W_i)$는 전체 말뭉치에서 나타난 어절의 수에 대한 특정 어절 W_i가 말뭉치에서 나타난 횟수로 구한다.

다음으로, 음절 쌍(음절 bigram)을 띄어쓴 경우와 붙여쓴 경우의 통계 정보를 추출하여, 음절 쌍(x, y) 사이에 공백이 올 확률$(P_{innerS}(x,y))$을 구하였다.

본 연구의 자동 띄어쓰기 방법은 어절 확률$(P(W_i))$을 이용하여 어절 후보가 되는 음절의 연속을 찾고, 어절 간 음절 bigram 통계 정보를 이용해 값을 보완한다.

임의의 문장 S가 있다고 할 때, 최우추정에 의해 최댓값을 가지는 S를 찾음으로써 최적의 띄어쓰기 위치를 찾을 수 있다. Viterbi 알고리즘을 사용하여 가능한 모든 문장의 확률 값 S중 가장 큰 값을 가지는 경우를 채택한다.

전처리 단계로 음절 bigram 통계를 이용하여 항상 띄어쓸 위치를 미리 찾아 띄어줌으로써 분석해야할 음절의 연속을 가능한 짧게 만든다. 그리고 문장의 확률 계산은 곱하기 연산이기 때문에 입력 문장을 많이 띄어쓸수록 값이 작아지게 되므로, 띄어쓴 부분이 띄어쓸 확률이 큰 위치일 경우에는 값을 보상해준다. Underflow 문제를 막기 위해 문장의 확률 계산에 log연산을 적용하였으며, 임의의 m값을 곱하여 음절 통계 정보를 이용한 보상 값의 영향력을 더 크게 한다. 어절 unigram와 음절 bigram 통계 정보를 이용하여 띄어쓸 위치를 찾는 모델은 다음과 같다.

$$\arg\max_{S} \sum_{k=1}^{n} \{ \log P(W_k) + m \log \left[\frac{P_{innerS}(LS\ of\ W_k,\ FS\ of\ W_{k+1})}{1 - P_{innerS}(LS\ of\ W_k,\ FS\ of\ W_{k+1})} \right] \}$$

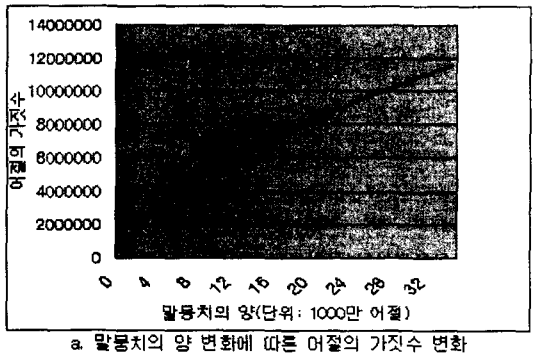
LS: 어절의 마지막 음절
FS: 어절의 첫 음절
P_{innerS}: 두 음절 사이 띄어쓸 확률

3.2. 통계 정보에 기반한 자료부족 문제 해결

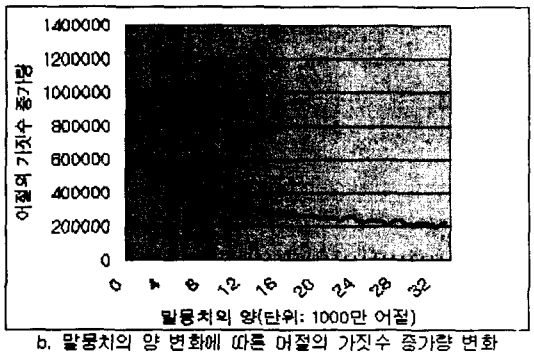
앞장에서 서술한 시스템은 자료부족 문제로 인하여 어절을 인식하는 데 어려움이 있다.

<그림 1>은 기본 어절 인식 중심의 띄어쓰기를 위해 통계 정보를 추출할 적절한 말뭉치의 크기를 알아보기 위해 말뭉치의 양 변화에 따른 어절의 가짓수 변화를 조사한 결과이다.

본 연구에서 통계 정보 추출을 위해 사용한 말뭉치의 크기인 33,643,884어절은 <그림 1-b>에서 어절의 가짓수 증가량 변화곡선이 완만해지는 지점의 양보다 더 많기 때문에, 기본적인 어절 정보를 충분히 포함한 양이라고 볼 수 있다. 또한, <그림 1-a>에서 볼 수 있듯이 말뭉치의 양을 충분히 늘리더라도 어절의 가짓수는 수렴하지 않는다. 따라서, 통계를 통해 인식할 어절의 수를 늘리는 것, 즉, 더 많은 양의 말뭉치에서 어절 통계를 추출하는 것은 시스템의 자료부족문제로 인한 오류의 근본적인 해결책이 되지 못한다.



a. 말뭉치의 양 변화에 따른 어절의 가짓수 변화



b. 말뭉치의 양 변화에 따른 어절의 가짓수 증가량 변화

<그림 1> 말뭉치의 양 변화에 따른 어절의 가짓수 변화

따라서 어절 통계 기반 시스템에서는 띄어쓸 단위로 인식할 수 있는 후보 어절의 보충이 필요하다. 본 연구는 자료부족 문제를 통계 정보를 이용하여 해결하기 위한 후보 어절의 보충 방법을 제안한다. 기본적으로 본 시스템은 말뭉치에서 추출한 통계 정보 중 어절 unigram 통계 정보는 어절을 인식하는 데 사용하고, 음절 bigram 통계 정보는 인식된 어절의 타당성을 검토하여 보상 값을 적용하는 데 사용한다. 이에 더해서, 제안하는 시스템은 어절 unigram과 음절 bigram 통계 정보

를 이용하여 추가 후보 어절을 생성한다.

한국어는 다양한 어미 변형, 확장으로 인하여 자료 부족 문제가 심각하게 나타난다. 따라서, 어절 통계 정보를 중심으로 입력 문장에서 어절로 가능한 음절 연속을 검색한 다음, 그 어절 후보의 마지막 음절과 그 다음에 오는 음절 쌍 사이에 띄어쓰는 통계 정보를 이용해 후보 어절(스무딩 어절)을 추가하여, 임의의 가능한 어절에 어미가 연속하여 붙음으로써 학습 말뭉치에는 나타나지 않은 더욱 긴 어절을 예측한다. 이때, 어절 후보의 마지막 음절과 그 다음에 오는 음절 쌍 사이에 붙여쓸 확률이 임계치[4] 이상이면 붙여쓸 수 있는 것으로 보고 그 다음 음절에 대해서 순차적으로 계산한다. 최장일치법을 이용하여 가장 긴 한 개의 어절을 어절로 인식 가능한 후보로 추가한다.

$$SW_i = W_i \times \prod_{k=1}^{n} [\{1 - P_{innerS}(LS\ of\ W_i, NS_k\ of\ W_i)\} \times weight]$$

SW_i : 스무딩 어절

$NS_k\ of\ W_i$: W_i의 다음 k번째 음절

$weight$: 임의의 가중치[5]

4. 실험 및 평가

시스템의 성능 평가 실험을 위해 내부 데이터(학습한 말뭉치), 21세기 세종계획 말뭉치, ETRI 품사부착 말뭉치, 인터넷상에 오른 신문기사에 대해 독자가 남긴 100자 평에서 각각 2000문장씩 추출하였다.

<표 3> 실험 데이터의 구성

[illegible]	[illegible]	[illegible]	총 [illegible] 수
[illegible]	2,000	25,020	103,196
[illegible]	2,000	13,971	40,353
ETRI [illegible]	2,000	17,191	52,688
100자 [illegible]	2,000	12,504	40,088

위의 실험 데이터를 어절 unigram과 음절 bigram을 이용한 기본적인 자동 띄어쓰기 시스템으로 처리한 결과, 평균 91.68%의 어절 단위 정확도를 얻었다. 통계 정보를 이용하여 어절로 인식 가능한 후보 어절을 추가하였을 때 결과(<표 4>)는 평균 93.76%로, 2.08%의 성능 개선이 이루어짐을 확인하였다. 이 시스템의 외부 데이터에 대한 성능만 비교해 보았을 때 평균 2.75% 향상을 보인다.

<표 4> 통계 정보를 이용한 어절 후보 추가 전, 후 정확도

[illegible]	[illegible]	[illegible]	[illegible] 후
내부 데이터	[illegible]	98.45	98.49
	[illegible]	98.19	97.85
[illegible]	[illegible]	90.91	93.45
	[illegible]	93.63	95.15
ETRI [illegible]	[illegible]	90.50	93.36
	[illegible]	93.67	95.64
[illegible]	[illegible]	86.88	89.74
	[illegible]	90.89	92.78

5. 결론

본 논문에서 제안하는 자동 띄어쓰기 시스템은 한글 문장의 띄어쓰기 접근법으로 가능한, 어절 인식 접근법[4]과 공백 삽입 접근법[1,2,3] 중, 이전의 통계 기반 방법들의 공백 삽입 접근법과 달리, 어절 통계 정보를 이용한 어절 인식 접근법을 취하고 있다.

또한 이전의 어절 인식 접근법들은 형태소 분석을 통해 어절을 인식하는 반면, 본 시스템은 통계 정보를 이용해 어절을 인식한다. 이 방법은, 구축이 손쉬우며, 실세계의 다사용 어절들을 우선으로 어절을 인식할 수 있는 장점이 있다.

통계 정보를 추출하는 말뭉치의 양이 시스템의 성능 변화에 영향을 크게 미치는 범위는 한정되어 있고, 아무리 큰 말뭉치에서 어절을 추출하여, 인식될 어절 후보를 늘려나가더라도 시스템은 통계 정보만으로는 실세계에서 나타나는 방대한 가짓수의 어절을 처리하기 힘들다. 이와 같은 문제에 대한 해결 방법으로 본 연구의 이전 연구에서 최장의 형태소 분석 결과를 어절의 viable-prefix 값으로 취하고 이를 후보 어절로 추가하는 방법을 제안하였다. 이 방법은 규칙을 이용하여 해결하는 방법으로, 세종 말뭉치에서 추출한 테스트 데이터에 대해서 어절단위 정확도 96.59%의 성능을 보인다[6].

본 논문에서는 통계 기반 어절 인식 기법의 자료 부족 문제의 극복 방법으로 통계 정보만을 이용한 방법을 제안하였다. 어절 unigram과 음절 bigram 통계 정보만을 이용하여 최장 어절을 유추하고, 이 어절을 인식 가능한 후보 어절로 추가한 결과, 시스템의 성능이 향상됨을 확인하였다.

참고 문헌

[1] 심광섭, "음절간 상호 정보를 이용한 한국어 자동 띄어쓰기", 정보과학회 논문지(B), 23권 9호, pp.991-1000, 1995.

[2] 강승식, "음절 bigram를 이용한 띄어쓰기 오류의 자동 교정", 음성과학회 논문지, 제8권 제2호, pp.83-90, 2001.

[3] 이도길, 이상주, 임희석, 임해창, "한글 문장의 자동 띄어쓰기를 위한 두 가지 통계적 모델", 정보과학회 논문지: 소프트웨어 및 응용 제30권 제4호, pp.358-370, 2003.

[4] 강승식, "한글 문장의 자동 띄어쓰기", 제 10회 한글 및 한국어 정보처리 학술발표 논문집, pp.137-142, 1998.

[5] 최성자, 강미영, 허희근, 권혁철, "음절 N-Gram과 어절 통계 정보를 이용한 한국어 띄어쓰기 시스템", 제15회 한글 및 한국어 정보처리 학술발표 논문집, pp.47-53, 2003.

[6] Mi-young Kang, Sung-ja Choi, Ae-sun Yoon, Hyuk-chul Kwon, "Stochastic Word-Spacing System with Dynamic Increase of Word List", Proceedings of the IJC-NLP04, 게재예정.

4) $1 - P_{innerS}(x, y) \geq 0.53$

5) 학습을 통해 얻은 값 $weight$ = 0.00014