# 한국어 텍스트 처리를 위한 줄 경계 띄어쓰기 복원

# Automatic Word-Segmentation at Line-Breaks for Korean Text Processing

정영미, 이재윤 (연세대학교 문헌정보학과)

Young-Mee Chung, Jae-Yun Lee
Department of Library and Information Science, Yonsei University

한국어 텍스트의 줄 경계에서의 띄어쓰기 복원을 위해 음절쌍 통계를 이용한 복원 기법을 설계하고 신문기사를 대상으로 통계 정보원과 음절쌍 위치에 따른 가중치를 달리하는 실험을 수행하였다. 실험 결과 처리 대상 기사를 포함하는 1개월 분 기사를 통계 정보원으로 하고 가중치는 균등하게 할 때 가장 높은 성공률을 얻었다. 이 결과는 디지털 원문을 텍스트 방식으로 소급하여 구축하는 경우에 적용될 수 있을 것이다.

## 1. 줄 경계의 띄어쓰기 복원 문제

정보처리나 정보서비스를 위한 디지털 본문 자료를 구축하는 방법으로 원문을 다운로드 받거나 스캐닝한 문서에서 OCR을 이용해서 원문 텍스트를 추출하는 방법이 이용되고 있다. 이때 영어의 경우와 달리 한국어 텍스트는 글자 단위로 줄바꿈을 하므로 줄 경계에서 어절이 분리되는 문제가 발생한다.

예를 들어 아래와 같이 한글 40자폭으로 맞춰져진 입력 텍스트의 경우 첫 번째 줄처럼 줄 경계가 "직원이/회사에"라는 두 어절의 사이이기 때문에 어절이 잘리지 않는 경우도 있지만, 두 번째 줄처럼 "관계/에"라는 어절이 줄 경계 때문에 "관계"와 "에"로 잘릴 수가 있다.

○…김현철씨 측근인 … 알려진 우보전자 직원이 회사에 사흘연속 지각한 … 납치되 사장님과의 관계에 대해 추궁당했다』고 … 일부 언론에 보도되는 등 소동을 빚은 … 자작극으로 밝혀졌다.

만약 줄 경계를 무조건 붙이는 방식을 택할 경우, 이 예에서는 "직원이회사에"와 같은 잘못 붙인 오류가 발생하며, 무조건 띄어쓰는 방식을 택할 경우 "관계 에"와 같은 잘못 띈 오류가 발생한다. 따라서 줄 경계를 원래대로 하나의 문장으로 복원할 때, 어절이 잘렸는지 아닌지를 판단하여 복원할 필요가 있다.

이와 같은 문제를 해결하기 위해서는 한국어 텍스트 자동 띄어쓰기에 적용하는 방식을 이용할 수 있다. 한국어 자동 띄어쓰기 처리 방식으로는 휴리스틱과 형태소분석기를 이용하는 규칙기반 방식(김계성, 이현주, 이상조 1997)과, 음절통계를 이용하는 통계기반 방식(심광섭 1996; 신중호, 박혁로 1997)이 있다.

규칙 기반 방식은 초등학교 교과서를 대상으로 최고 95.2%의 성공률을 보인 것으로 나타났으나, 미등록어와 신규 복합명사가 빈번하게 나타나는 텍스트에 대해서는 높은 정확률을 얻기가 어렵다. 이에 반해 통계기반 방식은 74~85% 정도로 다소 낮은 성공률을 보이지만, 통계 데이터를 획득한 말뭉치에 그대로 적용할 경우에는 93.6%의 성공률을 보여서 통계 데이터를 얻는 말뭉치의 영향이 큰 것으로 나타났다(심광섭 1996).

줄 경계 띄어쓰기 복원 문제는 본문 전체에 대한 띄어쓰기와는 달리 줄 경계 이외의 본문에서 띄어쓰기 판단에 필요한 음절통계 정보를 획득할 수 있으므로 통계적인 방식을 적용할 여지가 높다. 따라서 이 연구에서는 다운로드 받은 한국어 신문기사를 대상으로 음절통계에 기반한 줄 경계

띄어쓰기 복원 방법을 개발하고 실험을 통해 적용 가능성을 확인해보았다.

## 2. 통계기반 줄 경계 띄어쓰기 복원

### 2.1. 통계기반 띄어쓰기 복원 기법 모형

처리 대상 본문으로부터 연속된 각 음절 쌍에 대해서 다음과 같은 위치별 빈도를 추출한다.

| | |
|---|---|
| $f_{str}$ | 음절 쌍의 두 음절이 붙어서 나타난 빈도 |
| $f_{head}$ | 음절 쌍의 앞이 구분기호인 빈도 |
| $f_{tail}$ | 음절 쌍의 뒤가 구분기호인 빈도 |
| $f_{mid}$ | 음절 쌍의 앞, 뒤가 구분기호가 아닌 빈도 |
| $f_{word}$ | 음절 쌍의 앞, 뒤가 구분기호인 빈도 |
| $f_{seg}$ | 음절 쌍이 두 어절의 경계에 떨어져서 나타난 빈도 |

예를 들어 음절 a와 b가 한 줄에서 다음과 같이 출현한 경우 음절쌍 ab의 각 출현빈도는 아래와 같다.

"ab** *ab *** ab** ab **ab****a b *** ab"
↑ ↑ ↑ ↑ ↑ ↑ ↑
① ② ③ ④ ⑤ ⑥ ⑦

*는 a나 b가 아닌 기타 음절. 빈칸은 빈칸 그대로.

- $f_{str}$ = 6 (①②③④⑤ ⑦)
- $f_{head}$ = 3 ( ③④ ⑦)
- $f_{tail}$ = 2 ( ② ④ )
- $f_{mid}$ = 1 ( ⑤ )
- $f_{word}$ = 1 ( ④ )
- $f_{seg}$ = 1 ( ⑥ )

여기서 $f_{str} >= f_{head} + f_{tail} + f_{mid} - f_{word}$ 인 관계가 성립하는데, 이는 음절 쌍이 줄의 맨 앞이나 맨 뒤에 출현해서 $f_{head}$나 $f_{tail}$을 판단하기 어려운 경우에는 $f_{str}$ 만 증가시켰기 때문이다. 이렇게 한 이유는 음절 쌍의 앞과 뒤에 대한 정보를 최대한 얻기 위해서이다. 만약 $f_{head}$를 어절 앞이면서 음절 쌍 뒤에 구분기호가 아닌 음절이 있는 경우로 센다면 위의 예에서 ⑦번과 같은 경우는 아무런 판단도 할 수가 없게 된다.

이와 같이 얻어진 통계를 이용해서 실제 줄 경계의 띄어쓰기를 복원하는 방법은 다음과 같다. 먼저 신문기사 본문 중에서 임의의 두 줄을 단순하게 표현하면 다음과 같다.

"**********ab"
"cd**********"

*는 공백이나 임의의 음절.

위와 같은 경우 줄 경계에 걸친 음절 b와 음절 c 사이를 붙일 확률 $p(abcd)$는 다음과 같이 세 확률의 합으로 구하여 값이 0.5 이상이면 ab와 cd 사이를 붙이도록 한다.

$$p(abcd) = \alpha\, p_{left}(ab) + \beta\, p_{con}(bc) + \gamma\, p_{right}(cd)$$

여기서 $\alpha\ \beta\ \gamma$는 가중치이고 각 항은 다음을 의미한다.

| | |
|---|---|
| $p_{left}(ab)$ | 음절쌍 ab 뒤에 빈 칸이 아닌 음절이 있을 확률 = 1 - (음절쌍 ab뒤에 빈 칸이 아닌 음절이 없을 확률) |
| $p_{con}(bc)$ | 음절쌍 bc 사이에 빈 칸이 없을 확률 |
| $p_{right}(cd)$ | 음절쌍 cd 앞에 빈 칸이 아닌 음절이 있을 확률 = 1 - (음절쌍 cd앞에 빈 칸이 아닌 음절이 없을 확률) |

실제 확률값은 다음 공식으로 각각 구한다.

$$p_{left}(ab) = \frac{f_{head}(ab) + f_{mid}(ab) - f_{word}(ab)}{f_{head}(ab) + f_{mid}(ab) + f_{tail}(ab) - f_{word}(ab)} = \frac{\text{뒤에 음절이 있는 경우의 빈도}}{\text{어절에 나타난 총 빈도(위치 확실한 경우)}}$$

$$p_{con}(bc) = \frac{f_{str}(bc)}{f_{str}(bc) + f_{seg}(bc)} = \frac{\text{어절에 나타난 총 빈도}}{\text{어절에 나타난 총 빈도} + \text{어절 경계에 나타난 총 빈도}}$$

$$p_{right}(cd) = \frac{f_{tail}(cd) + f_{mid}(cd) - f_{word}(cd)}{f_{head}(cd) + f_{mid}(cd) + f_{tail}(cd) - f_{word}(cd)} = \frac{\text{앞에 음절이 있는 경우의 빈도}}{\text{어절에 나타난 총 빈도(위치 확실한 경우)}}$$

단, 위 공식에서 분모가 0이면 음절 쌍이 출현하지 않은 경우이므로 그 확률값을 경험적으로 0.5로 정하였다. 또한 ab나 cd가 음절쌍이 되지 못하고 빈 칸과 같은 구분기호를 포함하여 한 음절일 때에는 음절쌍 통계가 없으므로, a가 구분기호인 경우에는 경험적인 분석에 의해 $p_{left}(ab) = 0.7$로, d가 구분기호인 경우에는 $p_{right}(cd) = 0.7$로 정하였다. 그리고 각 항의 반영비율을 나타내는 가중치 $\alpha$, $\beta$, $\gamma$는 $\alpha+\beta+\gamma = 1$ 이 성립되며, 실험을 통해 최적의 경우를 파악하였다.

### 2.2. 통계 기반 띄어쓰기 복원 사례

실제로 앞의 방법을 이용하여 줄 경계의 띄어쓰기 판단을 하는 사례를 살펴보기로 한다. 기사 본문의 어떤 두 줄이 다음과 같은 경우, 이 기사가 속한 종합면 1996년 4월 기사 집합에서 얻은 음절쌍의 위치별 빈도는 <표 1>과 같다.

의 축소 수사 여부와 … 3김정치청산, 역사바로세우기 등을 놓고 … 공약도 쏟아졌다. 대드

<표 1> 음절쌍의 위치별 빈도

| 음절쌍 | $f_{str}$ | $f_{head}$ | $f_{tail}$ | $f_{mid}$ | $f_{word}$ | $f_{seg}$ |
|---|---|---|---|---|---|---|
| "사바" | 1 | 0 | 0 | 0 | 0 | 3 |
| "바로" | 27 | 21 | 19 | 0 | 0 | 1 |
| "로세" | 4 | 0 | 0 | 3 | 0 | 4 |

이때, 음절쌍 "사바" 뒤에 빈칸이 아닌 음절이 있을 확률 $p_{left}$(사바)와, 음절쌍 "바로" 사이를 붙일 확률 $p_{con}$(바로), 음절쌍 "로세" 앞에 빈 칸이 아닌 음절이 있을 확률 $p_{right}$(로세)는 각각 다음과 같이 계산된다.

$$p_{left}(\text{사바}) = \frac{f_{head}(\text{사바}) + f_{mid}(\text{사바}) - f_{word}(\text{사바})}{f_{head}(\text{사바}) + f_{mid}(\text{사바}) + f_{tail}(\text{사바}) - f_{word}(\text{사바})} = \frac{0+0-0}{0+0+0-0} \Rightarrow 0.500$$

$$p_{con}(\text{바로}) = \frac{f_{str}(\text{바로})}{f_{str}(\text{바로}) + f_{seg}(\text{바로})} = \frac{27}{27+1} = 0.964$$

$$p_{right}(\text{로세}) = \frac{f_{tail}(\text{로세}) + f_{mid}(\text{로세}) - f_{word}(\text{로세})}{f_{head}(\text{로세}) + f_{mid}(\text{로세}) + f_{tail}(\text{로세}) - f_{word}(\text{로세})} = \frac{0+3-0}{0+3+0-0} = 1.000$$

가중치를 각각 $\alpha$=0.3, $\beta$=0.4, $\gamma$=0.3으로 정했을 때, "사바"와 "로세" 사이를 붙일 확률은 다음과 같이 구해진다.

$$p(\text{사바로세}) = 0.3\times0.500 + 0.4\times0.964 + 0.3\times1.000 = 0.836$$

결국 확률값이 0.5 이상이므로 "사바"와 "로세" 사이는 빈 칸이 없는 형태로 붙여서 "역사바로세우기"라는 한 어절로 복원된다.

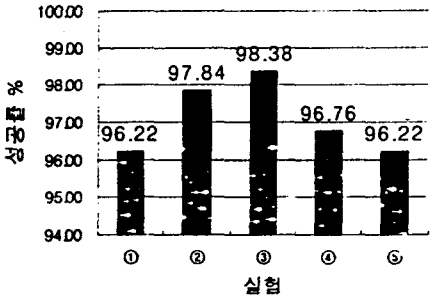
### 2.3. 통계정보 추출원의 최적 규모 판단 실험

앞에서 제시한 모형을 적용해서 실제 줄 경계의 띄어쓰기를 복원할 때 음절쌍 통계정보를 추출하는 분석 대상 텍스트 집단의 규모에 따라서 띄어쓰기 처리의 성공률이 달라질 수 있다. 통계정보 추출원의 규모가 너무 작으면 적용할 통계정보가 부족할 것이고, 반대로 너무 커도 잡음정보가 많아져서 성공률이 떨어질 수 있다.

통계정보 추출원의 규모에 따른 성공률의 차이를 평가하기 위해 1996년 4월 종합면 기사 중에서 임의로 기사번호 1~5, 41~45, 81~85, 121~125의 40건을 추출하여 처리대상 실험 집합 JH9604를 구성하였다. JH9604에는 띄어쓰기 여부를 판단해야 할 줄바꿈이 185건 있다. 음절쌍 통계정보를 추출하는 정보원으로 다음의 5가지 경우를 이용하여 실험하고 각각의 성공률을 평가해 보았다.

① 처리 대상 기사 집합 (=JH9604)
② 처리 대상 기사가 포함된 12일 분량의 정치면 신문기사 집합
③ 처리 대상 기사가 포함된 그 달의 정치면 신문기사 집합
④ 처리 대상 기사가 포함된 10개월 간의 정치면 신문기사 집합
⑤ 처리 대상 기사가 포함된 10개월 간의 정치, 사회, 경제면 모두의 신문기사 집합

실험 결과는 아래 <그림 1>과 같다.




<그림 1> 통계 정보원별 복원 실험 결과

실험 결과를 보면 처리 대상 기사가 포함된 한 달 분량의 같은 면 기사 집합을 통계정보 추출원으로 사용한 ③의 경우에서 가장 좋은 성공률을 얻었고, 이보다 집합이 작아지거나 커지면 오류도 많아지는 것으로 나타났다.

한편, 통계정보가 희박한 ①의 경우와 반대로 잡음정보가 많은 ⑤의 경우에 오류 건수는 같은 7건이었지만, 그중 1건(127번)을 제외하고는 서로 다른 오류를 보이는 것으로 나타났다.

<표 2> 처리 대상 185건 중 실패한 경우

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ① | 3 | | | 70 | 91 | | | 127 | | 134 | 158 | | | 183 |
| ② | | | 58 | | | | | 127 | | | | 176 | 177 | |
| ③ | | 6 | | | | 115 | | | | | | 176 | 177 | |
| ④ | | | | | 114 | 115 | 125 | 127 | 133 | | | 176 | | |
| ⑤ | | | 58 | | 114 | 115 | 125 | 127 | 133 | | | 176 | | |

### 2.4. 최적 가중치 판단 실험

줄 경계를 불일 확률을 판단하는 공식에서 윗줄 끝의 음절쌍에 대한 가중치 $\alpha$, 줄 경계에 나뉜 음절쌍에 대한 가중치 $\beta$, 아래줄 맨 앞 음절쌍에 대한 가중치 $\gamma$는 각각 띄어쓰기를 판단할 때 앞 부분 음절쌍, 가운데 음절쌍, 뒷 부분 음절쌍에 대한 통계정보를 어느 정도 반영할지를 결정하는 것이다. 이 가중치가 달라지면 성공률이 변할 수 있다고 보고, 어느 경우가 최적인지를 실험을 통해 알아보았다.

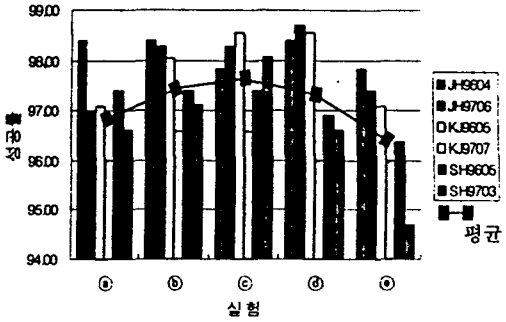
실험을 위해서 정치, 사회, 경제 면의 3개 면 기사 중에서 1996년 5개월 분중에서 1개월 분, 1997년 5개월 분 중에서 1개월 분의 기사를 임의로 선정하여 면종별, 시기별로 고른 6개 집합을 구성하였다. 이 6개 집합에 대해서 각각 기사번호 1~5, 41~45, 81~85, 121~125인 20개 기사씩을 추출하여 실험집단 6개를 구성하였다.

각 실험집합에 대해 적용하기 위해 사용하는 통계정보는 앞의 실험에서 최적의 경우로 증명된 같은 달, 같은 면, 한 달 분 기사집합에서 추출하였다. 실험은 $\beta$값을 0.5에서 0.2까지 5단계로 줄이면서 수행하였고, $\alpha$와 $\gamma$값은 1에서 $\beta$를 뺀 나머지를 2등분한 값으로 정하였다. 실험결과 복원 성공률은 <표 3>과 같고 이를 <그림 2>에 도표로 나타내었다.

개별 복원 실험 중에서 최고값은 JH9706을 대상으로 $\alpha$=0.35, $\beta$=0.30, $\gamma$=0.35일 경우의 98.71% 였지만, $\alpha$=0.33, $\beta$=0.34, $\gamma$=0.33로 균등하게 가중치를 부여하였을 때 평균 성공률이 97.67%로 가장 높고 실험 집단 사이의 표준 편차도 0.71로 가장 낮은 것으로 나타났다.

<표 3> 가중치별 띄어쓰기 복원 실험 결과

| 정보원 | 처리건수 | 복원 성공률 (%) | | | | | |
|---|---|---|---|---|---|---|---|
| | | $\alpha,\beta,\gamma$의 값 | | | | | |
| | | 0.25<br>0.50<br>0.25 | 0.30<br>0.40<br>0.30 | 0.33<br>0.34<br>0.33 | 0.35<br>0.30<br>0.35 | 0.40<br>0.20<br>0.40 | 평균 |
| | | ⓐ | ⓑ | ⓒ | ⓓ | ⓔ | |
| JH9604 | 185 | 98.38 | 98.38 | 97.84 | 98.38 | 97.84 | 98.16 |
| JH9706 | 232 | 96.98 | 98.28 | 98.28 | 98.71 | 97.41 | 97.93 |
| KJ9605 | 205 | 97.07 | 98.05 | 98.54 | 98.54 | 97.07 | 97.85 |
| KJ9707 | 350 | 96.00 | 96.57 | 96.57 | 96.00 | 96.00 | 96.23 |
| SH9605 | 193 | 97.41 | 97.41 | 97.41 | 96.89 | 96.37 | 97.10 |
| SH9703 | 207 | 96.62 | 97.10 | 98.07 | 96.62 | 94.69 | 96.62 |
| 건당평균 | | 96.94 | 97.52 | 97.67 | 97.38 | 96.50 | 97.20 |
| 표준편차 | | 0.80 | 0.72 | 0.71 | 1.16 | 1.14 | 0.79 |




<그림 2> 가중치별 복원 실험 결과

## 3. 결론

음절쌍 통계를 이용한 한국어 신문기사 텍스트의 줄 경계에서의 띄어쓰기 복원 실험 결과 동일한 면의 1개월 분 기사 집합을 통계 정보원으로 사용하고, 음절쌍의 위치별 가중치를 균등하게 하는 것이 97.67%로 가장 높은 성공률을 보이는 것으로 나타났다.

실험에서 띄어쓰기 복원에 실패한 경우는 통계 데이터가 부족하거나 잡음 데이터가 많은 경우였다. 이는 통계적인 기법을 적용할 때에 피할 수 없는 오류이므로, 성공률을 이보다 높이기 위해서는 규칙 기반 방식을 일부 혼용하는 연구가 필요하다.

## 참고문헌

김계성, 이현주, 이상조. 1997. "음절 정보를 이용한 한국어 띄어쓰기의 구현," 1997년도 한국정보과학회 가을 학술발표논문집, 24(2): 243-246.

신중호, 박혁로. 1997. "음절단위 bigram정보를 이용한 한국어 단어인식모델," 제9회 한글 및 한국어 정보처리 학술대회 발표논문집: 255-260.

심광섭. 1996. "음절간 상호정보를 이용한 한국어 자동 띄어쓰기," 정보과학회논문지(B), 23(9): 991-1000.