음절 단위 및 자모 단위의 Byte Pair Encoding 비교 연구

이찬희○, 이동엽, 허윤아, 양기수, 임희석*
고려대학교 정보대학 컴퓨터학과
{chanhee0222, judelee93, yj72722, willow4, limhseok}@korea.ac.kr

Comparing Byte Pair Encoding Methods for Korean

Chanhee Lee○, Dongyub Lee, YunA Hur, Kisu Yang, Heuiseok Lim*
Department of Computer Science and Engineering, College of Informatics, Korea University

요 약

한국어는 교착어적 특성이 강한 언어로, 교착어적 특성이 없는 영어 등의 언어와 달리 형태소의 수에 따라 조합 가능한 어절의 수가 매우 많으므로 어절 단위의 처리가 매우 어렵다. 따라서 어절을 더 작은 단위로 분해하는 전처리 단계가 요구되는데, 형태소 분석이 이를 위해 주로 사용되었다. 하지만 지도학습 방법을 이용한 형태소 분석 시스템은 다량의 학습 데이터가 요구되고, 비지도학습 방법을 이용한 형태소 분석은 성능에 큰 하락을 보인다.

Byte Pair Encoding은 데이터를 압축하는 알고리즘으로, 이를 자연어처리 분야에 응용하면 비지도학습 방법으로 어절을 더 작은 단위로 분해할 수 있다. 본 연구에서는 한국어에 Byte Pair Encoding을 적용하는 두 가지 방법인 음절 단위 처리와 자모 단위 처리의 성능 및 특성을 정량적, 정성적으로 분석하는 방법을 제안하였다. 또한, 이 방법을 세종 말뭉치에 적용하여 각각의 알고리즘을 이용한 어절 분해를 실험하고, 그 결과를 어절 분해 정확도, 편향, 편차를 바탕으로 비교, 분석하였다.

주제어: 형태소 분석, 자연어처리, Byte Pair Encoding

1. 서론

자연어처리 연구가 가장 활발히 이루어지는 언어인 영어의 경우, 교착어적 특성이 존재하지 않기 때문에 띄어쓰기로 구분되는 어절 단위 처리가 주를 이룬다. 반면 교착어적 특성이 강한 한국어의 경우에는 조합 가능한 어절의 수가 형태소와 비례하여 기하급수적으로 증가하기 때문에 어절 단위 처리 시 어휘집의 크기가 매우 커지게 된다. 많은 자연어처리 시스템들은 메모리 사용량이 어휘집의 크기에 비례하여 증가하므로 시간적/물리적 제약 사항으로 인해 어휘집의 크기를 제한시킨다. 제한된 어휘집에 포함되지 못한 어휘는 모두 동일한 특수 어휘(OOV, Out-of-Vocabulary)로 치환되기 때문에 이는 시스템의 성능 하락으로 이어진다.

이러한 어절 단위 처리의 어려움으로 인해 교착어의 자연어처리에는 일반적으로 형태소 분석 단계가 추가된다. 이를 통해 각 어절은 형태소 단위로 분해가 되고, 형태소 단위 처리 방법을 사용함으로써 어휘집의 크기를 현실적인 수준으로 축소시킬 수 있다.

하지만 형태소 분석 단계가 추가되면서 새로이 발생하는 문제들도 존재하는데, 그중 하나가 학습 말뭉치가 필요하다는 점이다. 비지도학습 방법으로 형태소 분석을 하는 방법들도 존재하지만[1,2,3], 이들은 지도학습 방법에 비해 성능이 낮으며, 따라서 우수한 성능의 형태소 분석 모델을 확보하기 위해서는 양질의 말뭉치가 필수적으로 요구된다.

비지도학습 방법들 중 어휘집의 크기를 감소시키는 알고리즘으로 Byte Pair Encoding(BPE)[4]이 있다. BPE은 단순하면서도 효과적인 데이터 압축 알고리즘으로, 가장 등장 빈도가 높은 바이트 쌍을 하나의 바이트로 표현하는 과정을 반복함으로써 압축을 실현한다. 이를 자연어처리 분야에 응용하여 어휘를 더 작은 단위로 분할하는 방법이 [5]에서 제안되었으며, 기계 번역 분야에 적용되었다.

자연어처리 분야에서 BPE을 적용하여 어휘집의 크기를 감소시키는 방법은 다음과 같다. 우선 부호들의 집합을 문자들의 집합으로 초기화한다. 이후 말뭉치에서 가장 등장 빈도가 높은 부호의 쌍을 묶어 새로운 부호로 부호집에 추가한다. 동시에 해당 부호 쌍을 하나의 부호로 묶는 규칙을 변환 규칙들의 집합에 추가한다. 이러한 과정은 정의된 결합 규칙의 수가 미리 정의한 개수 k가 될 때까지 반복된다.

형태소 분석을 포함한 모든 어절 분해 방법은 각 어절을 하나 이상의 단위로 분해하기 때문에 분해 전과 비교하여 문장의 길이가 길어지게 된다. BPE을 이용한 분해의 경우, 결합 규칙의 수가 증가하면 분해된 요소의 평균적인 길이도 증가하게 된다. 따라서 모든 어절을 표현하는 데 필요한 어휘집의 크기는 증가하는 반면 분해 전과 비교하여 증가하는 문장의 길이는 평균적으로 감소한다. 문장의 길이는 많은 자연어처리 시스템의 계산량 및 속도와 비례 관계에 있으므로 적은 것이 더 효율적이지만, 어휘집의 크기 또한 적은 것이 유리하므로 둘 사이의 균형을 찾는 것이 중요하다. BPE을 사용하는 경우 형태소 분석과 다르게 k의 값을 조절하는 방법으로 필요에 맞게 어휘집의 크기를 결정할 수 있다는 장점이 있다.

한국어에 BPE을 적용하는 데는 두 가지 방법이 있는데, 첫 번째는 음절을 최소 단위로 보는 것이고, 두 번째는 자소를 최소 단위로 사용하는 것이다. 본 연구에서

* Corresponding author

는 두 가지 방법을 정량적, 정석적으로 비교 및 분석하였다. 분석 결과, 어절의 분해 성능은 음절 단위 방법이 미세하게 우수하지만, 어휘집 축소 효율은 자소 단위 방법이 더 우수한 것으로 나타났다.

2. 실험 방법

BPE을 적용하는 두 가지 방법의 정량적 비교를 위해서는 어절의 분해 방법에 대한 정답 값이 필요한데, 한국어에서 이에 가장 널리 이용되는 방법인 형태소 분석 결과를 정답 값으로 사용하였다. 여기서 형태소 분석으로 인해 분해된 개수와 비교 대상 알고리즘으로 분해된 개수가 일치하면 정답으로, 일치하지 않으면 오답으로 정의하고 분해의 정확도를 계산하였다. 이에 더해 비교 대상 알고리즘의 분해 개수가 형태소 분석을 사용했을 때와 비교하여 편향(bias)의 정도와 편차(deviation)의 정도를 통해 비교 대상 알고리즘의 분해 특성을 분석하였다. 편향과 편차는 아래 식에 따라 계산된다.

$$Bias = \frac{\sum_{i=1}^{N}(C_g^i - C_m^i)}{N} \quad (1)$$

$$Deviation = \frac{\sum_{i=1}^{N}|C_g^i - C_m^i|}{N} \quad (2)$$

여기서 C_g^i는 i번째 어절의 분해 개수 정답 값, C_m^i은 i번째 어절의 비교 대상 알고리즘 분해 개수이며, N은 비교에 사용된 어절의 수이다. 또한, 알고리즘별로 최종적으로 생성되는 어휘집의 크기를 비교함으로써 어휘집 축소 효율을 분석하였다.

정성적인 분석으로는 말뭉치에 등장하는 어절들 중 일부를 대상으로 동일한 어절에 대해 형태소 분석, 음절 단위 BPE, 자모 단위 BPE 알고리즘의 분해 결과물을 나열하는 방법을 사용하였다.

실험을 위한 말뭉치는 세종 말뭉치 중 현대 문어 형태 분석 말뭉치를 사용하였다. 또한, 해당 말뭉치에서 등장하는 형태소 분석 결과를 정답 분해 결과로 사용하였다. 본 연구에서는 정확한 형태소 분석보다는 분해 개수를 목적으로 하므로, 형태소 분석의 모호성으로 인해 여러 분해 개수가 존재하는 어절의 경우 성능 측정에서 제외하였다. 그 결과 총 1,492,079개의 고유한 어절이 실험에 사용되었으며, 이는 전체 말뭉치에 등장하는 고유한 어절의 98.91%에 해당하는 것으로 나타났다.

3. 실험 결과

3.1 어절 분해 정확도

앞서 기술된 대로, 어절 분해 정확도는 형태소 분석을 이용하여 분해된 개수와 비교 대상 알고리즘을 이용하여 분해된 개수의 일치 여부에 따라 정확도를 계산하여 측정하였다. 여기서 비교 대상이 되는 BPE 알고리즘의 경우, 결합 규칙의 수 k가 유일한 하이퍼 파라미터이다. 따라서 k의 값을 1,000부터 100,000까지 1,000단위로 변화시켜가며 성능 변화를 측정하였다. 실험 결과는 그림 1과 표 2에 정리되어 있다.

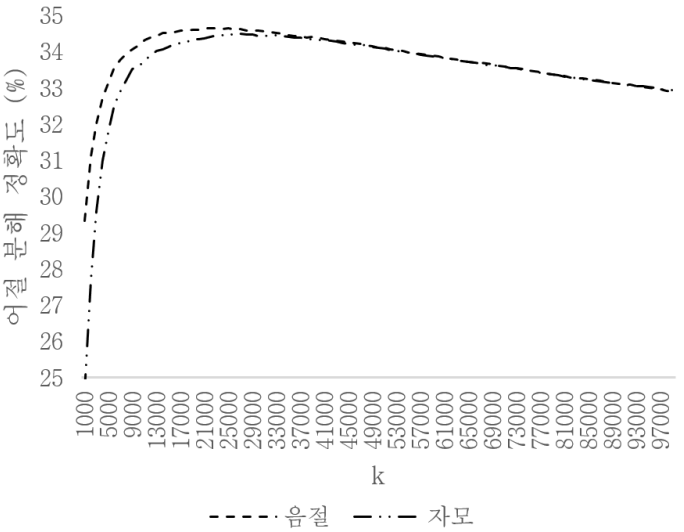

그림 1. k값의 변화에 따른 음절 단위 BPE과 자모 단위 BPE의 어절 분해 정확도.

음절 단위 BPE의 경우, 24,000개의 결합 규칙을 생성하였을 때 가장 높은 34.65%의 성능을 보였다. 반면 자모 단위 BPE 알고리즘을 사용했을 때는 28,000개의 결합 규칙을 생성하였을 때 가장 높은 34.49%의 성능을 보여 음절 단위 BPE 알고리즘이 미세하게 우수한 것으로 나타났다. 두 알고리즘 성능의 절대 차이는 0.16%이며, 자모 단위 대비 음절 단위 방법의 상대 오차 감소량(Relative Error Reduction)은 0.24%이다. 또한, 자모 단위 BPE 알고리즘이 최대 성능을 내기 위해 더 많은 결합 규칙을 필요로 하는 것으로 나타났는데, 이는 한 음절이 2개 이상의 자모로 구성되기 때문에 자모 단위 처리의 경우 상대적으로 더 많은 횟수의 결합이 필요한 것이 그 원인으로 추정된다.

3.2 편향 및 편차

다음으로, 식(1) 및 식(2)에 따라 편향과 편차를 계산하여 두 알고리즘의 특성을 분석하였다. 앞선 실험에서와 마찬가지로 k의 값은 1,000부터 100,000까지 변화시켜가며 편향과 편차를 계산하였다. 본 실험의 결과는 그림 2, 3과 표 2에 나타나 있다.

편향의 경우 알고리즘의 분해 개수가 정답 값에 비해 평균적으로 어느 방향으로 얼마나 치우쳐 있는지를 나타낸다. 따라서 형태소 분석과 동일하게 어절 분해가 가능하다면 편향의 값은 0이 되며, 이 값에 가까울수록 알고리즘의 분해 성능이 우수하다고 해석할 수 있다. 음절 단위 BPE과 자모 단위 BPE 모두 최고 성능이 나타난 k값과 유사한 26,000과 29,000에서 0에 가장 가까운 편향

표 2. k값의 변화에 따른 음절 단위 BPE과 자모 단위 BPE의 어절 분해 정확도, 편향, 편차. 지면의 한계로 인해 두 알고리즘의 최대 성능이 나타나는 21,000~30,000, 45,000~50,000구간의 수치만 정리하였다.

k		21000	22000	23000	24000	25000	26000	27000	28000	29000	30000	…	45000	46000	47000	48000	49000	50000
분해 정확도 (%)	음절	34.60	34.64	34.64	**34.65**	34.62	34.64	34.61	34.62	34.58	34.59	…	34.29	34.25	34.23	34.21	34.16	34.14
	자모	34.34	34.36	34.40	34.45	34.45	**34.49**	34.46	34.49	34.47	34.47	…	34.25	34.21	34.20	34.19	34.17	34.13
편향	음절	-0.07	-0.05	-0.04	-0.02	-0.01	**0.01**	0.02	0.03	0.04	0.05	…	0.19	0.20	0.20	0.21	0.22	0.22
	자모	-0.12	-0.10	-0.09	-0.07	-0.05	-0.04	-0.02	-0.01	**0.00**	0.02	…	0.16	0.17	0.18	0.19	0.19	0.20
편차	음절	0.96	0.96	0.96	0.96	0.95	0.95	0.95	0.95	0.95	0.95	…	0.95	**0.95**	0.95	0.95	0.95	0.95
	자모	0.98	0.97	0.97	0.97	0.97	0.96	0.96	0.96	0.96	0.96	…	0.95	0.95	**0.95**	0.95	0.95	0.95

값이 관찰되었다. 이는 편향을 바탕으로 알고리즘의 대략적인 분해 정확도를 예측할 수 있음을 의미한다.

편차는 각 알고리즘의 분해 결과가 형태소 분석에 따른 분해 개수로부터 얼마나 떨어져 있는지를 나타낸다. 따라서 형태소 분석 방법의 경우는 편향과 마찬가지로 그 값이 0이 되지만, 편향과는 다르게 그 값이 작을수록 성능이 우수함을 의미한다. 음절 단위 BPE과 자모 단위 BPE은 각각 46,000과 47,000에서 가장 작은 편차인 0.95를 보였다. 어절 분해 정확도와 편향과는 다르게, 두 알고리즘 모두 최소 편차가 나타나는 분해 규칙의 수와 각 시점의 편차에 있어서 큰 차이가 없는 것으로 나타났다. 이로부터 두 알고리즘의 어절 분해 정확도는 다르지만, 실제 분해되는 개수와 정답 값의 차이는 유사한 수준이라는 점을 확인할 수 있다.

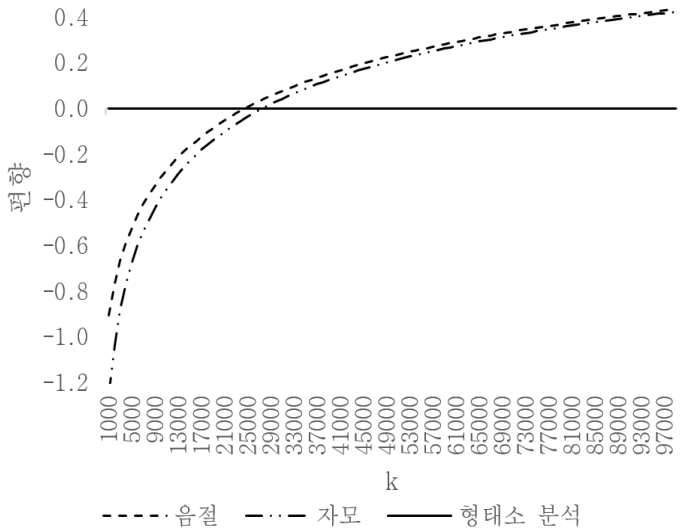

그림 2. k값의 변화에 따른 음절 단위 BPE과 자모 단위 BPE의 편향.

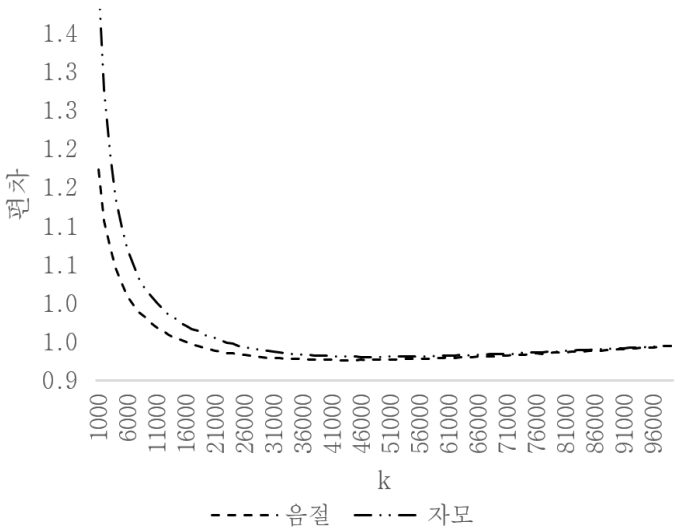

그림 3. k값의 변화에 따른 음절 단위 BPE과 자모 단위 BPE의 편차.

3.3 어휘집 축소 효율

본 장에서는 세종 말뭉치의 모든 어절을 표현하기 위해 각 알고리즘이 몇 개의 어휘(부호)를 필요로 하는지를 앞선 실험들과 마찬가지로 k값을 변화시켜가며 비교하였다. 세종 말뭉치에 등장하는 고유한 어절의 수는 1,508,561개이며, 형태소 분석을 할 경우 고유한 형태소의 수는 216,080개이다. k값의 변화에 따른 음절 단위 BPE과 자모 단위 BPE의 어휘집 크기는 그림 4에 나타나 있다.

동일한 수의 결합 규칙을 사용할 때, 자모 단위 BPE은 음절 단위 BPE에 비해 더 적은 수의 어휘집이 요구되는 것으로 나타났다. 이를 바탕으로 어휘집 크기 축소에는 BPE가 더 효율적인 것으로 확인되었다.

표 3. 원본 어절과 그에 해당하는 형태소 분석, 음절 단위 BPE, 자모 단위 BPE 적용 결과.

어절	형태소 분석	음절 단위 BPE	자모 단위 BPE
건축구조체계와	건축+구조+체계+와	[O] 건축+구조+체계+와	[O] 건축+구조+체계+와
드러낸다고	드러내+ㄴ다고	[O] 드러+낸다고	[O] 드러내+ㄴ다고
노나니	놀+나니	[O] 노+나니	[O] 노+나니
가득채운	가득+채우+ㄴ	[O] 가득+채+운	[X] 가득+채운
만족시키고자	만족+시키+고자	[O] 만족+시키+고자	[X] 만족+시키고자
머물렀다가	머무르+었+다가	[O] 머물+렀+다가	[X] 머물러+ㅆ다가
다녔으므로	다니+었+으므로	[X] 다녔+으므로	[O] 다녀+ㅆ+으므로
문체공격으로	문체+공격+으로	[X] 문+체+공격+으로	[O] 문체+공격+으로
떠올릴지	떠올리+ㄹ지	[X] 떠올+릴+지	[O] 떠올리+ㄹ지
의미하지만	의미+하+지만	[X] 의미+하지만	[X] 의미+하지만
달여주시는	달이+어+주+시+는	[X] 달+여+주+시는	[X] 달+여+주+시는
입출력장치	입출력+장치	[X] 입+출+력+장치	[X] 입+출+력+장치

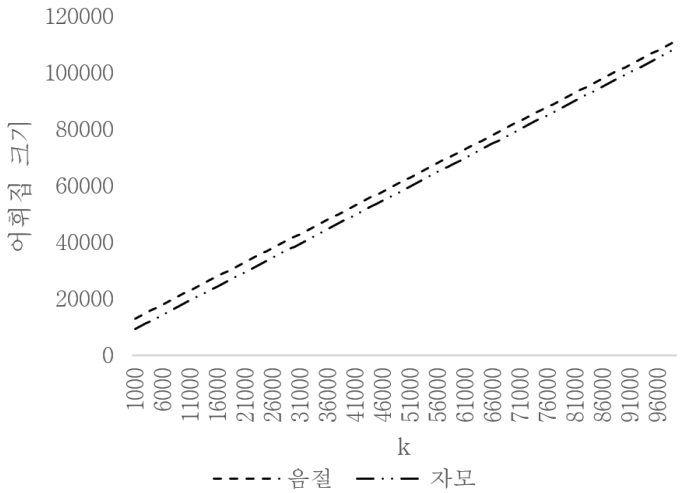

그림 4. k값의 변화에 따른 음절 단위 BPE과 자모 단위 BPE의 어휘집 크기.

3.4 정성적 분석

각 분해 알고리즘의 결과물에 대한 정성적인 분석을 위해 동일한 어절에 대한 형태소 분석, 음절 단위 BPE, 자모 단위 BPE의 적용 결과물을 표 3에 정리하였다. “건축구조체계와”에서의 예와 같이, 독립적인 어절을 구성할 수 있는 형태소들의 결합으로 생성된 어절의 경우 두 BPE 알고리즘 모두 형태소 분석과 동일하게 분해를 함을 확인할 수 있다. 다만, “노나니”, “머물렀다가”, “다녔으므로”, “달여주시는”에서 보이듯 형태소 간의 결합으로 인해 형태소에 변형이 생기는 경우에는 분해 과정에서 변형은 일어나지 않는 BPE 알고리즘의 한계로 인해 형태소 분석과 동일한 결과를 얻는 것은 불가능하다. 한 가지 흥미로운 사실은 “드러낸다고”와 “떠올릴지”에서의 예처럼, 형태소 분석 단계에서 음절 내에 분해의 경계가 존재하는 경우 음절 단위 BPE 알고리즘은 정상적으로 분해를 할 수 없지만, 자모 단위 BPE 알고리즘은 이를 형태소 분석과 동일하게 분해할 능력이 존재한다는 점이다.

4. 결론

본 연구에서는 데이터 압축 방법의 하나인 BPE을 이용하여 비지도학습 방법으로 교착어적 특성이 강한 한국어의 어절을 더 작은 단위로 분해하는 방법을 비교, 분석하였다. BPE의 최소 단위를 음절로 사용하는 것과 자모로 사용하는 것의 차이를 형태소 분석을 기준으로 분해 정확도, 편향, 편차를 이용하여 비교한 결과, 분해의 정확도 및 편향 면에서는 음절 단위의 BPE이 더 우수한 반면, 어휘집 축소의 효율은 자모 단위의 BPE가 더 좋은 성능을 보이는 것으로 확인되었다.

향후 연구로는, 각각의 BPE 적용 단위에 따라 기계 번역과 같은 상위 자연어처리 시스템의 성능이 어떠한 영향을 받는지를 정량적으로 확인할 것이다. 또한, BPE 알고리즘을 개선시켜 어절 분해 정확도를 향상시키는 것도 흥미로운 연구 방향이 될 것으로 기대한다.

Acknowledgement

본 연구는 과학기술정보통신부 및 정보통신기술진흥센터의 대학ICT연구센터지원사업의 연구결과로 수행되었음 (IITP-2018-0-01405)

참고문헌

[1] Poon, H., Cherry, C. and Toutanova, K., Unsupervised morphological segmentation with log-linear models. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (pp. 209-217). Association for Computational Linguistics, 2009.

[2] Bernhard, D., Unsupervised morphological segmentation based on segment predictability and word segments alignment. In Proceedings of 2nd Pascal Challenges Workshop (pp. 19-24), 2006.

[3] Demberg, V., A language-independent unsupervised model for morphological segmentation. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (pp. 920-927), 2007.

[4] Gage, P., A new algorithm for data compression. The C Users Journal, 12(2), pp.23-38, 1994.

[5] Sennrich, R., Haddow, B. and Birch, A., Neural machine translation of rare words with subword units. arXiv preprint arXiv:1508.07909, 2015.