키워드 추출용 구묶음 데이터 구축 및 개선 방법 연구

이민호[O1], 최맹식[1], 김정아[1], 이충희[1], 김보희[1], 오효정[2], 이연수[1]

[1] 엔씨소프트 NLP Center
[2] 전북대학교
[1][minolee, mschoi, kjeongah, forever73, bohui09, yeonsoo] @ncsoft.com

[2]ohj@jbnu.ac.kr

Study on Making Chunking Dataset for Keyword Extraction and its Improvement Methods

Minho Lee[O1], Maengsik Cho[1]i, Jeongah Kim[1], Chunghee Lee[1], Bohui Kim[1], Hyo-Jung Oh[2], Yeonsoo Lee[1]
[1] NCSOFT NLP Center
[2] Jeonbuk National University

요 약

구묶음은 문장을 겹치지 않는 문장 구성 성분으로 나누는 과정으로, 구묶음 방법에 따라 구문분석, 관계 추출 등 다양한 하위 태스크에 사용할 수 있다. 본 논문에서는 문장의 키워드를 추출하기 위한 구묶음 방식을 제안하고, 키워드 단위 구묶음 데이터를 구축하기 위한 가이드라인을 제작하였다. 해당 가이드라인을 적용하여 구축한 데이터와 BERT 기반의 모델을 이용하여 학습 및 평가를 통해 구축된 데이터의 품질을 측정하여 78점의 F1점수를 얻었다. 이후 패턴 통일, 형태소 표시 여부 등 다양한 개선 방법의 적용 및 재실험을 통해 가이드라인의 개선 방향을 제시한다.

주제어: 구묶음, 키워드 추출, 가이드라인 제작, 데이터 구축

1. 서론

자연어처리에서 구묶음(chunking)이란 문장을 명사, 동사 등의 역할을 하는 문장 구성 성분으로 겹치지 않게 나누는 과정을 말한다.[1] 각각의 문장 구성 성분은 하나 이상의 형태소로 이루어져 있으며, 이를 구(chunk)라고 한다. 구묶음은 구문분석이나 의미역 파싱, 개체명 인식 등 자연어처리의 응용 분야에서 처리해야 할 의미 단위를 파악하기 위해 사용된다.

구묶음의 범위는 여러 가지로 정의할 수 있다. [2]에서는 구묶음을 응용분야에 따라 단위를 다르게 해야 한다고 하였다. [3]에서는 '구'를 『같은 역할을 수행하는 형태소들을 묶어 하나의 의미를 가진 부분적인 구문 요소』 라고 정의하였다. 이와 같이 [4][5] 등 많은 연구에서는 구묶음을 형태소 분석과 구문분석의 중간 과정으로 인식하였다. 또 다른 방법으로는 특정 품사만을 대상으로 하는 구묶음이 있다. [6]에서는 명사구만을 대상으로 구묶음을 수행하는 방법을 제시하였다. 위의 방법들은 구문분석에는 도움이 되지만, 검색, 관계 추출, 상호 참조 등 키워드가 중요한 분야에서는 사용하기 어렵다는 문제가 있다. 본 논문에서는 추출된 구가 곧바로 키워드 기반 응용 분야에 사용될 수 있도록 범위를 잡는 키워드 단위 구묶음을 제시한다.

표 1에는 [3], [6]에서 정의한 방식의 구묶음과, 본 논문에서 제시한 키워드 단위 구묶음의 예시가 있다. 첫 번째 예시와 같이, [3]방식의 구묶음은 형태소로 나눈 뒤 여러 어절에 걸쳐 있는 문장 성분을 묶는 것에 중점을 둔다. 반면, 두 번째 예시와 같이 키워드 단위 구묶음은 문장의 주어, 서술어, 목적어 등을 나누는데 중점을 둔다. 이 방법은 문장의 주요 문장 구성 성분이 잘 드러나 있기 때문에, 관계 추출, 상호 참조 등 키워드를 사용해야 하는 태스크에 도움이 될 것으로 생각할 수 있다.

표 1 기존의 구묶음과 키워드 단위 구묶음의 예시

[3][5]	<사내><의> <존재><를> <전혀> <모르><는 듯이> <시치미><를> <떼><고 있><다><.>
[6]	<사내의 존재>를 전혀 모르는 듯이 <시치미>를 떼고 있다.
키워드 단위	<사내의 존재를> <전혀 모르는 듯이> <시치미를 떼고 있다.>

본 논문에서는 키워드 단위 구묶음 데이터를 만들기 위한 가이드라인을 제작하였다. 해당 가이드라인에 따라 데이터를 구축한 뒤, 기계 학습 방식을 이용하여 데이터의 품질을 측정하였다.

본 논문의 기여점은 다음과 같이 요약할 수 있다.

- 키워드 추출의 관점에서 구묶음 데이터 구축을 위한 가이드라인을 제시한다.

- 가이드라인을 사용하여 데이터를 구축하고, 데이터 품질 실험을 통해 가이드라인의 유효성을 보인다.

- 데이터 개선 실험을 통해 가이드라인의 개선 방향을 제시한다.

표 2 데이터 통계

분류	문서 수	문장 수	구 수
경제	150	*1612*	*13994*
국제	143	1754	16053
IT	141	2135	19032
생활	147	2516	20742
사설	150	**3134**	**22349**
정치	142	1919	18594
사회	145	2056	18454
스포츠	144	2034	15990
총합	**1162**	**17160**	**145208**

2. 관련 연구

구묶음은 자연어 처리에서 오래 전부터 연구되어 왔던 분야로, [7]에서 구문분석(dependency parsing)의 전처리 과정으로 제시되었다. [8]에서는 CoNLL-2000 Chunking dataset이라는 통일된 데이터를 만들었으며, 영어권에서는 이 데이터를 사용하여 다양한 구묶음 시스템을 개발 및 평가하였다. [9]에서는 PENN treebank의 구문분석 데이터에서 구묶음으로 전환하는 규칙을 학습하였고, [10]에서는 SVM을 사용하여 구묶음을 수행하는 등 기계학습을 사용하지 않은 통계적 모델이 연구가 많이 되었다.

하지만 최근에는 기계 학습의 발전으로 구묶음을 거치지 않고 종단 간(end-to-end) 학습을 통한 구문분석을 하는 흐름이 만들어져, 기계 학습을 사용한 구묶음은 잘 연구되지 않았다. [11]에서는 앞서 언급한 구묶음 데이터를 다루는 실험을 하였지만, 구묶음 자체에 집중하기보다는 넓은 범위의 시퀀스 레이블링 문제를 다루어 다양한 출력 방식을 실험하였다.

한국어에서는 구묶음을 위한 데이터셋을 따로 만들지 않고, 형태소 말뭉치나 구문분석 말뭉치를 사용하여 규칙 기반으로 처리하였다. [4]에서는 형태소 말뭉치에서 규칙을 학습하고, 구문분석 말뭉치를 사용하여 평가 데이터를 구축하였다.

[12]에서 세종계획 말뭉치를 제작한 이후, 많은 구묶음 논문이 세종계획 말뭉치에 포함된 구문분석 말뭉치를 규칙 기반으로 전처리하여 구묶음 데이터를 만드는 방식으로 연구를 진행하였다. [6]에서는 사전 정보를 사용하여 명사열을 중심으로 구묶음을 수행하였다. [5]에서는 Bi-LSTM과 조건부 무작위장(CRF, Conditional Random Field)를 사용한 신경망 기반의 모델로 좋은 성능을 보고하였다. 하지만 각각의 논문에서 데이터를 각자의 방식으로 전처리를 하기 때문에, 구묶음을 위한 통일된 데이터셋은 존재하지 않는다. 본 연구에서는 문장의 주요 키워드 추출에 중점을 두고 수식 구조와 최대 범위 구묶음을 기반으로 한 가이드라인을 구축하였다.

3. 가이드라인 및 데이터 구축

본 논문에서는 두 단계를 통해 구묶음 데이터를 제작하였다. 먼저 구문분석과 통계적 전처리 과정을 통해 기본적인 구묶음을 수행한 뒤, 가이드라인에 맞춰 수동으로 수정하였다.

3.1 데이터 수집 및 전처리

구문분석 데이터로 많이 쓰이는 세종 말뭉치[1]([12])는 잡지와 책, 방송 녹음 전사 자료 등을 사용하여 구축되었다. 최근 국립국어원에서 공개한 모두의 말뭉치[2] 역시 비슷하게 신문 기사와 녹음 자료를 사용하여 구축되었다. 본 논문에서도 이와 같이 신문 기사를 7개 주제를 선정하여 150 개 내외의 문서를 모은 뒤, 전처리 과정과 주석 과정을 통해 데이터를 구축하였다.

전처리 과정은 은닉 마르코프 모델(Hidden Markov Model) 기반의 형태소 분석기와 [13]에서 제작한 조건부 무작위장 기반의 구문분석 모델을 사용하여 의존구조 트리를 만든 뒤, 규칙과 통계 정보를 사용하여 기초적인 구묶음을 수행하는 것으로 이루어졌다. 통계 정보를 사용한 이유는 구문분석 정보만을 사용하여 전처리를 할 경우 구문분석기의 오류가 있을 수 있고, 어휘 수준에서 자주 등장하는 관용적 표현 등은 의존구조와 상관없이 등장하는 경우가 많기 때문이다.

3.2 가이드라인

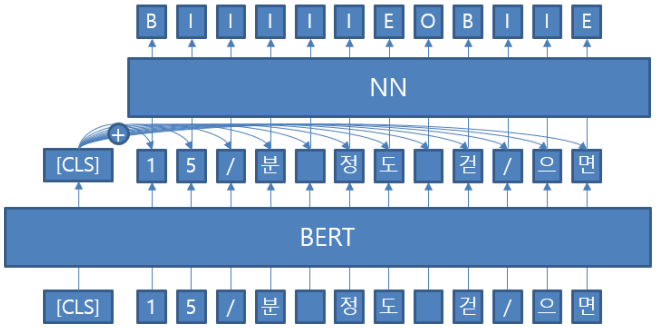
본 논문에서는 키워드 추출을 위해 다음과 같은 원칙을 정하고 데이터를 구축하였다.

1. 구의 단위는 구문분석을 통해 문장 구조를 반영하여 진행하며, 수정 방향은 '최대 범위 묶음'으로 한다.
2. 체언류, 명사구, 동사/형용사 파생 접사가 붙은 체언 어근, 개체명, 기타 명사처럼 쓰이는 외국어, 외래어, 숫자, 기호 등 문장을 이루는 전체 요소들을 대상으로 한다
   2.1 수식하는 말은 수식 받는 말과 되도록 하나로 묶는 것을 원칙으로 한다.
3. 하나의 개체, 행위, 상태 등으로 직관적으로 인식되는 구를 하나로 묶는다.
4. 코퍼스 구축은 전처리 결과를 기준으로 수정 작업을 진행하는 것으로 한다.

문장의 키워드 추출을 위해서는 중심이 되는 명사구와 명사구를 수식하는 수식구를 함께 묶어야 키워드의 정확한 의미를 알 수 있다. 하지만 이렇게 되는 경우 하나의 구가 매우 길어지는 경우가 많아, 일정 기준을 두고 구

[1]https://ithub.korean.go.kr/user/guide/corpus/guide1.do

[2]https://corpus.korean.go.kr/main.do



15/분 정도 걷/으면

15분 정도 걸으면

그림 1 BERT 기반 구묶음 모델

를 나누기로 하였다. 예를 들어, “미국 기업과 거래를 금지당한 중국의 통신장비 기업 ZTE는”과 같은 문구를 볼 때, 수식 관계를 합친다면 전체 범위를 하나의 구로 묶어야 한다. 그러나 이 방식은 하나의 구에 지나치게 많은 정보가 담기게 된다고 판단하였다. 원문의 내용, 작업자 간의 의견 조율 등을 종합적으로 고려해 보았을 때, 복잡한 조건을 기반으로 구를 나누는 것은 작업에 어려움이 있다고 생각하였다. 그 결과 가장 직관적인 기준인 어절 수를 제한하는 방식으로 작업을 진행하기로 하였다.

주석 과정에서는 전처리 과정을 거친 구묶음 데이터를 기반으로 진행되었다. 작업자들은 형태소 정보와 전처리를 통해 구 경계가 표시된 문장을 위의 가이드라인에 맞추어 수정하는 작업을 진행하였다. 이 과정을 통해 얻은 데이터의 통계는 표 2에 나타나 있다.

4. 데이터 품질 측정 실험

본 논문에서는 3.1절에서 제작한 가이드라인과 이를 바탕으로 3.2절에서 만든 데이터를 사용하여 가이드라인과 데이터가 얼마나 품질이 좋은지를 측정하고자 하였다. 이를 위해 [14]에서 발표한 BERT를 사용하여 시퀀스 레이블링(Sequence Labeling) 모델을 만든 뒤, 학습/검증/평가 데이터를 나눠서 학습을 진행하였다.

4.1 모델

본 논문에서는 시퀀스 레이블링 방법으로 구묶음 모델을 제작하였다. 시퀀스 레이블링은 순차적 데이터에 Begin, Inside, Outside, End, Single을 의미하는 BIOES 태그를 통해 범위를 표현하는 방식이다. 다양한 신경망 모델 중 Attention을 기반으로 한 BERT 모델을 선택하여 실험을 진행하였다. 그림 1에는 모델의 구성과 예시 입력이 나타나 있다. "15분 정도 걸으면" 이라는 문구가 있을 때, 어절 내의 형태소를 "/"로 구분한 형태가 BERT의 입력으로 들어간다. 이후 BERT 층을 통과하면 각각의 토큰에 대한 임베딩과 문장 전체에 대한 내용을 담고 있는 CLS 임베딩을 얻을 수 있게 된다. CLS 임베딩의 값을 다

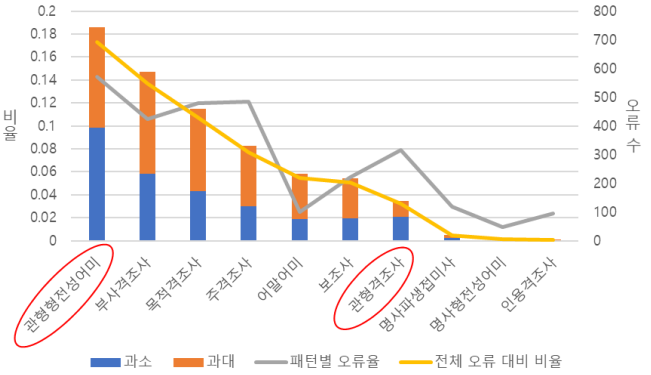
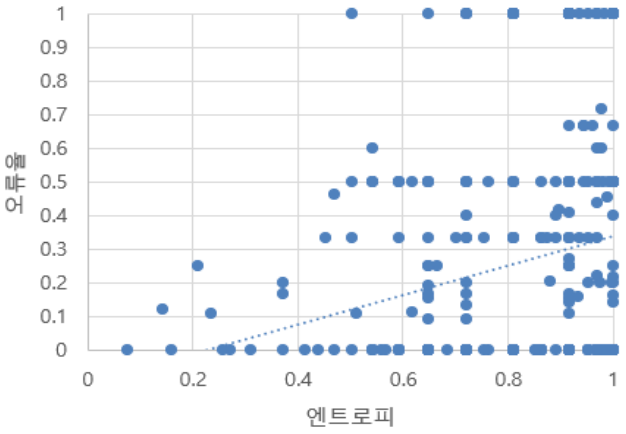


그림 2 형태소 태그별 오류율 그래프



← 일관적 일관적이지 않음 →

그림 3 패턴의 일관성과 오류율의 관계를 나타낸 그래프. 하나의 점은 하나의 패턴을 의미한다.

른 모든 토큰의 임베딩 값에 각각 더한 뒤, 단일 층의 신경망을 사용하여 각각의 태그에 대한 확률값을 계산하였다. 위의 예시에서는 <15분 정도> <걸으면> 이라는 구를 추출할 수 있다.

한국어 BERT 사전 학습 모델은 구글에서 만든 다국어(multilingual), ETRI에서 만든 KorBERT 등이 있는데, 이들은 WordPiece 단위로 만든 통계 기반의 토큰화만 지원하거나, 형태소 태그를 부착한 형태만을 지원한다. 본 논문에서는 뉴스 기사를 사용하여 음절 기반, 형태소 기반의 BERT 모델을 사전 학습하여 실험을 진행해 보았다. 음절 기반의 모델이 성능이 더 좋아 음절 기반 모델을 기준으로 결과 성능을 기록하였다.

또한, 시퀀스 레이블링 모델에서는 조건부 무작위장을 마지막 층에 적용하여 성능이 향상되었다는 연구가 많지만, 본 실험에 적용해 본 결과 속도와 성능이 모두 저하되어 사용하지 않았다.

4.2 실험 평가 방법

본 논문에서는 모델의 성능을 평가하기 위해 CoNLL-2000[8], CoNLL-2003[15] 등의 시퀀스 레이블링을 평가하는 방법과 같은 기준을 적용하였다. 시스템이 예측한 구와 정답의 범위가 일치할 경우를 "같은 구"로 정의한다.

표 3 BERT 모델의 구묶음 성능

	정밀도	재현율	F1
Baseline	49.47	58.30	53.52
BERT-형태소	59.05	60.66	59.84
BERT-음절	**78.07**	**79.51**	**78.79**

표 4 일관적이지 않은 패턴의 예시

1) <비둘기들이> <아예> <서식을 하는 **경우가 많다**>
2) <전문가의 상담이 필요한 **경우가**> <**많다**>

표 5 일관적이지 않은 패턴 일부의 통계 정보

패턴	분절	결합	엔트로피
것으로 보인다.	38	71	0.932
것으로 알려졌다.	34	51	0.971
도움이 된다.	11	16	0.975

같은 구의 수를 시스템이 예측한 총 구의 수로 나눈 것을 정밀도(precision)으로, 같은 구의 수를 정답 데이터 셋에 있는 구의 수로 나눈 것을 재현율(recall)으로 계산한다. 이렇게 구한 정밀도와 재현율의 조화평균으로 계산되는 F1 점수를 최종 성능으로 측정하였다. [8][15]에서는 각각의 덩이마다 역할 태그가 부여되어 있어 역할 태그까지 맞추는 것을 정답의 기준으로 삼았지만, 3장에서 구축한 데이터에는 NP, VP 등 구가 어떤 역할을 하는지는 표시되어 있지 않다. 따라서 구의 역할 태그를 제외하고, 추출된 구의 일치 여부만을 평가 대상으로 하였다.

4.3 실험 결과 및 분석

표 3에는 3장에서 만든 데이터를 기반으로 4.1절에서 제작한 모델을 학습, 평가한 결과가 나타나 있다. Baseline은 3.2절에서 언급한 전처리 과정만을 수행한 데이터와 주석한 데이터를 비교한 점수이다. 실험 결과 BERT모델의 성능은 79점으로, Baseline보다는 많이 좋아졌지만 기존의 구묶음 모델과 비교했을 때 낮은 성능이 나오는 것을 확인하였다.

그림 2는 어절의 끝 형태소의 종류별로 오류율을 계산한 그래프이다. 여기서 오류율은 분절 오류의 합을 정답 구 수로 나눈 값으로 계산하였다. 그림에서 관형형 전성어미와 관형격 조사의 패턴 내 오류율(녹색 실선)이 비교적 높은 축에 속하는 것을 알 수 있다. 하지만 가이드라인상 "수식하는 말은 수식받는 말과 하나로 묶는 것이 원칙" 이라는 조항을 적용하여 본다면, 관형형 전성어미와 관형격 조사는 뒤의 체언을 수식하는 기능으로 쓰이기 때문에 반드시 뒤따라오는 구와 하나로 묶여야 할 것으로 보인다. 그럼에도 불구하고 오류가 생기고, 심지어 오류율이 비교적 높게 나타났다는 것은 여기에 일관적이지 않은 패턴이 있는 것으로 판단이 가능하다. 여기서 패턴이란 연속된 두 어절을 말하고, 일관적인 패턴이란 특정 패턴이 등장한 모든 문장에서 모두 분절되거나 모두 결합되어 있는 경우를 의미한다. 예를 들어, 표 4의 문장에서 "경우가 많다"와 같은 패턴의 경우 문장 1에서는 분절하지 않고, 문장 2에서는 분절한다. 이 경우 일관적이지 않은 패턴이 된다. 이는 3.2절에서 언급한 어절 수 제한 때문인 것으로 유추해 볼 수 있다.

어절 수 제한 조항은 관형형 조사만이 아니라 모든 패턴에 영향을 미칠 수 있다고 생각하여, 구축한 데이터 전체를 대상으로 패턴의 일관성과 오류율의 상관관계를 구해 보고자 하였다. 그림 3은 패턴별로 일관성의 지표인 엔트로피(Entropy)와 오류율을 계산하여 좌표평면상에 흩뿌린 그래프이다. 엔트로피란 정보이론에서 정보의 표현량을 정의한 지표로, 특정 데이터가 일관적일수록 낮은 값을 가진다. 엔트로피는 다음과 같은 식으로 계산된다.

$$H(X) = -\sum_{x \in X} p_x \log(p_x)$$

위 식을 사용하여 문장 1, 2에서 등장한 "경우가 많다"라는 문구의 엔트로피를 계산하면 1이 나온다.

그림 3에서, 일관적이지 않은 패턴일수록 오류율이 높은 것을 알 수 있다. 표 5에는 자주 등장한 일관적이지 않은 패턴과, 그에 대한 분절, 미분절 수, 계산된 엔트로피의 예시를 정리해 두었다. 이에 대한 개선 실험은 5장에서 자세히 분석한다

5. 품질 개선 실험

4.2절에서 제시한 실험 결과는 [5] 등 최근의 구묶음 시스템의 성능에 비해 성능이 낮다. 이는 모델보다는 데이터의 문제일 것이라고 생각하여, 다양한 방식으로 구축된 데이터를 바꿔 동일 환경에서 학습과 평가를 반복해 보았다.

데이터 개선 방법은 두 가지 방식을 사용하였다. 일관적이지 않은 데이터를 통일하는 방법, 형태소 표시 여부에 변화를 준 뒤, 4장에서 제시한 실험을 반복하였다. 각각의 데이터 개선 실험에 대한 결과는 밑의 절에서 소개한다.

5.1 일관적이지 않은 패턴 통일

4.2절에서 보였듯이, 일관적이지 않은 패턴은 오류율이 높은 경향을 보이고, 이는 성능에 안 좋은 영향을 미친다. 이를 통일하면 성능이 높아질 것으로 기대하여 세 가지의 방법으로 데이터를 통일하여 실험을 하였다. **기본**은 4장에서 만든 모델의 예측 결과이다. **분절**은 패턴의 두 어절 사이를 분절하는 방식이고, **결합**은 반대로 두 어절을 결합하는 방식이다. **다수결**은 패턴별 분절, 결합 수를 구하여 더 많은 쪽으로 데이터를 통일한 방식이다. 이 방법들을 사용하여 학습, 평가를 진행한 결과를 표 6에 정리하였다,

실험 결과, 다수결 방식으로 패턴을 통일했을 때 수정하지 않은 데이터에 비해 성능이 1점가량 오른 것을 알 수 있다.

표 6 패턴 수정 방식별 성능 점수

모델	정밀도	재현율	F1
기본	78.07	79.51	78.79
분절	78.23	79.19	78.71
결합	78.29	79.81	79.04
다수결	**78.76**	**80.47**	**79.61**

표 7 패턴 수정 방식 데이터간 유사도. 유사도는 일치하는 구의 수를 전체 구 수로 나눈 값으로 계산한다.

	기본	분절	결합	다수결
기본	-	99.96	99.93	99.96
분절	-	-	99.89	99.93
결합	-	-	-	99.97
다수결	-	-	-	-

표 8 형태소 표시 여부에 따른 성능 점수

모델	형태소 표시 F1	형태소 미표시 F1
기본	78.79	78.46
분절	78.71	77.75
결합	79.04	78.71
다수결	**79.61**	78.51

여기서 흥미로운 점은 각각의 데이터 간 같은 구 수의 비율이 99.9%를 넘는다는 것이다. 표 7에는 수정한 데이터 간의 데이터 일치율을 측정해 기록하였다. 이를 볼 때 패턴의 통일성은 모델의 성능 점수에 유의미한 개선을 보인다는 것을 알 수 있다.

5.2 형태소 표시 여부

형태소 정보를 표시한 것이 과연 도움이 될지에 대해 평가해 보기 위해, 형태소 표시를 하지 않고 원문을 사용하여 학습과 평가를 다시 진행해 보았다. 즉, 그림 1의 예시에서, BERT의 입력 토큰이 "15/분 정도 걷/으면" 대신 "15분 정도 걸으면"과 같은 형태로 들어가게 된다. 본 논문에서 사용한 음절 단위 BERT에서는 형태소 구분자인 "/"가 UNK 토큰으로 들어가기 때문에 성능에 부정적인 역할을 줄 수 있을 것이라고 판단하였다. 이에 대한 성능을 비교하여 표 8에 정리하였다.

실험 결과, 모든 데이터에서 형태소를 표시한 것이 형태소를 표시하지 않은 것보다 성능이 1점가량 더 좋은 것을 알 수 있다. 이는 형태소 구분자로 인해 생기는 노이즈보다 형태소를 표시한다는 것 자체가 더 큰 이득이 되는 것으로 해석할 수 있다. 또한, 이는 UNK 토큰이 들어가도 학습 과정에 큰 영향을 주지 않는다고도 볼 수 있는데, 이는 BERT의 단어 사전에 한글의 거의 모든 글자가 들어가 있어 구분자만이 UNK 토큰으로 들어가게 되기 때문이다.

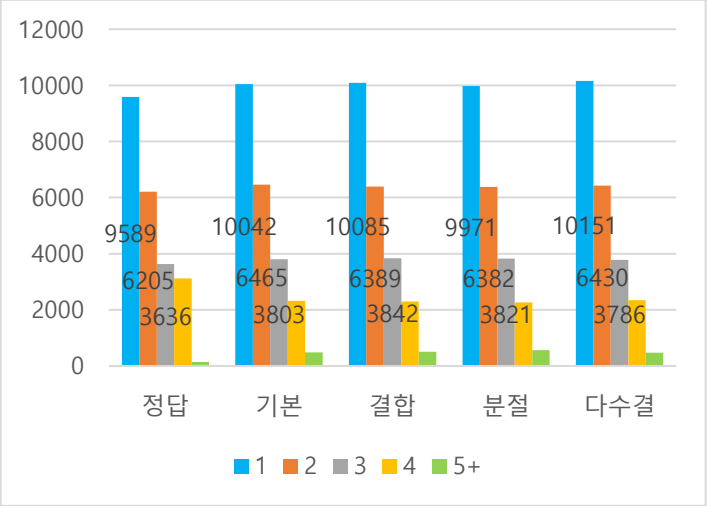


그림 4 정답과 패턴 통일 데이터별 학습 모델의 예측 결과에서 구당 어절 수

5.3 사례 연구

표 9에는 예시 문장에 대한 정답과, 각각의 데이터셋에서 학습한 모델의 구묶음 예측 결과가 나타나 있다. **정답**은 3.2절에서 제작한 데이터이고, 기본, 분절, 결합, 다수결은 5.1절에서 제시한 데이터 수정 방법을 사용하여 학습한 모델의 예측 결과이다. 우선 모든 모델이 수식 구조가 없는 단순한 주어-술어 구조는 잘 나누는 경향을 보인다. 첫 번째 문장의 "행정소송법 개정위원회를 구성해" 와 세 번째 문장의 "과거의 흔적이 가감 없이 드러낸다" 와 같이 수식이 없는 주어-술어가 명확한 패턴이 이에 해당한다. 구축된 데이터도 이러한 패턴에서는 일관적으로 주석되어 있다. 그러나 두 번째 문장의 "켜켜이 쌓인 세월의 흔적"이나 세 번째 문장의 "이 나간 시멘트 벽돌에서"와 같이 수식구가 붙은 경우는 학습된 모델들이 서로 다른 예측을 하고 있다.

또한, 데이터를 수정하더라도 추가 분절이 일어나거나, 분절한 부분을 결합하는 식으로 학습이 진행되지 않고, 위치만 서로 바뀌는 경향을 보였다. 이는 표 9의 2, 3번째 문장과 그림 4의 구당 어절 수의 통계를 통해 확인할 수 있다. 모든 패턴을 분절한 데이터에서 학습한 모델이 1어절로 이루어진 구를 추출한 수보다, 평균적으로 구의 길이가 길 것으로 예상되는 결합 데이터에서 1어절로 이루어진 구를 추출한 수가 오히려 더 많은 것을 알 수 있다. 이를 통해 분절 수와 학습 결과가 비례하지 않는다는 것을 알 수 있는데, 이는 [16]에서 실험적으로 보인 BERT의 특성 때문으로 유추해 볼 수 있다. BERT는 사전 학습 과정에서 주변 토큰은 물론, 의존구조상의 지배소 등 언어학적 관계가 있는 토큰에 높은 가중치가 붙도록 학습이 된다. 이를 세부 조정하는 과정에서 언어학적 요소를 고려하지 않은 기계적 분절 및 결합이 문제가 되었을 수 있다. 오히려 BERT의 사전 학습된 가중치가 약해져서 문법적으로 맞지 않는 구묶음이 등장하게 되는 원인이 될 것으로 추측된다.

표 9 사례 분석

정답	대법원이	지난 2002년	행정소송법 개정위원회를	구성해	처음 논의를 시작한 이후	16년째다.
기본	대법원이	지난 2002년	행정소송법 개정위원회를	구성해	처음 논의를 시작한 이후	16년째다.
분절	대법원이	지난 2002년	행정소송법 개정위원회를	구성해	처음 논의를 시작한 이후	16년째다.
결합	대법원이	지난 2002년	행정소송법 개정위원회를	구성해	처음 논의를 시작한 이후	16년째다.
다수결	대법원이	지난 2002년	행정소송법 개정위원회를	구성해	처음 논의를 시작한 이후	16년째다.

정답	공장들 사이	사이마다	**켜켜이 쌓인**	**세월의 흔적**	위에	…
기본	공장들 사이	사이마다	**켜켜이 쌓인**	**세월의**	**흔적** 위에	…
분절	공장들 사이	사이마다	**켜켜이 쌓인**	**세월의 흔적**	위에	…
결합	공장들 사이	사이마다	**켜켜이 쌓인**	**세월의 흔적**	위에	…
다수결	공장들 사이	사이마다	**켜켜이**	**쌓인 세월의 흔적**	위에	…

정답	벗겨진 페인트칠이나	**이 나간**	**시멘트 벽돌에서**	과거의 흔적이	가감 없이 드러낸다.
기본	벗겨진 페인트칠이나	**이 나간**	**시멘트 벽돌에서**	과거의 흔적이	가감 없이 드러낸다.
분절	벗겨진 페인트칠이나	**이 나간**	**시멘트 벽돌에서**	과거의 흔적이	가감 없이 드러낸다.
결합	벗겨진 페인트칠이나	**이 나간**	**시멘트 벽돌에서**	과거의 흔적이	가감 없이 드러낸다.
다수결	벗겨진 페인트칠이나	**이 나간**	**시멘트 벽돌에서**	과거의 흔적이	가감 없이 드러낸다.

이런 사례들을 볼 때, 좋은 구묶음 데이터를 만들기 위해서는 언어학적 요소를 고려하여 패턴을 수정할 필요가 있다. 데이터를 제작하는 과정에서 사용한 의존구조 트리의 수식 조건을 제약 조건으로 사용하는 등의 방법을 고려할 수 있다.

6. 결론 및 향후 연구

본 논문에서는 키워드 추출이라는 관점에서 구묶음을 하기 위한 가이드라인을 제작하였다. 가이드라인에 맞는 데이터를 제작하고, 이를 기반으로 모델을 학습하여 데이터의 품질을 측정하였다. 실험과 분석을 하여 개선할 부분을 찾고, 추가적인 실험을 통해 가이드라인의 수정 방향을 제시하였다. 구가 지나치게 길어지는 것을 방지하기 위해 어절 수 제한을 두었으나, 이는 일관적이지 않은 패턴이 만들어지는 원인이 되는 것을 알 수 있었다.

의존 구조를 목적으로 하는 구묶음과 달리, 키워드 추출을 위한 구묶음은 특정하지 못하는 일반 명사를 수식 관계를 함께 묶음으로써 단일 대상으로 특정하는 것이라고 할 수 있다. 이런 관점에서 현재 구축한 데이터는 키워드는 잘 구분했으나, 수식 관계를 처리하는 데 있어 아직 추가적인 개선이 필요한 상태이다.

향후 연구로는 5장에서 제시한 방법대로 언어학적 의미가 있는 패턴만을 고려하여 수정하거나, 4장의 모델을 개선하는 등 더 나은 평가 모델을 제작해 볼 예정이다. 또한, 개체명 인식, 관계 추출 등 구묶음을 사용하는 하위 태스크에 적용하여 실제로 어떤 방법이 가장 잘 맞는지를 분석해 볼 예정이다.

참고문헌

[1] Daniel Jurafsky, Speech and Lanugage Processing, 3rd edition, 2019, web draft

[2] 김재훈, 부분 구문분석 방법론, 2000, 정보처리학회지

[3] 남궁영 외 6명, 구문 분석을 위한 한국어 말덩이 정의, 2018, 제30회 한글 및 한국어 정보처리 학술대회

[4] 임지희 외 3명, 자동 구축된 구문패턴사전과 규칙을 이용한 구묶음, 2004, 한국정보과학회 언어공학연구회 학술발표 논문집

[5] 남궁영 외 7명, 한국어 말덩이 정의와 구묶음: 한국어 말덩이 부착 말뭉치와 Bi-LSTM/CRFs 모델을 활용하여, 2020, 정보과학회논문지

[6] 안광모, 서영훈, 명사 의미 부류를 이용한 연속된 명사열의 구묶음, 2010, 한국콘텐츠학회논문지

[7] Steven Abney, Parsing bu Chunks, 1991, "Principle-Based Parsing"

[8] Erik F. Tjong Kim Sang, Sabine Buchholz, Introduction to the CoNLL-2000 Shared Task: Chunking, 2000, CoNLL

[9] Lance A. Ramhaw, Mitchell P. Marcus, Text Chunking using Transformation-Based Learning, 1999, Natural Language Processing Using Very Large Corpora

[10] Taku Kudo, Yuji Matsumoto, Chunking with Support Vector Machines, 2001, NAACL

[11] Feifei Zhai et al., Neural Models for Sequence Chunking, 2017, AAAI

[12] 황용주, 최정도, 21세기 세종 말뭉치 제대로 살펴보기, 2007, 국립국어원 누리집

[13] 최맹식, 정석원, 김학수, CRFs를 이용한 의존구조 분석 및 의존 관계명 부착, 2014, 정보과학회논문지: 소프트웨어 및 응용

[14] Jacob Devlin et al., BERT: Pre-Training of Deep Bidirectional Transformers for Lanugauge Understanding, 2018

[15] Erik F. Tjong Kim Sang, Fien De Meulder, Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition, 2003, CoNLL

[16] Kevin Clark et al., What Does BERT Look At? An Analysis of BERT's Attention, 2019, arXiv preprint arXiv:1906.04341