다단계 기계학습 기법을 이용한 구묶음 성능향상

전길호○, 서형원, 최명길,
남유림, 김재훈
한국해양대학교, NLP연구실
asone7784@naver.com, wonn24@gmail.com, cmg5478@naver.com
zin1984@nate.com, jhoon@hhu.ac.kr

Performance Improvement of Chunking Using Cascaded Machine Learning Methods

Kil-Ho Jeon○, Hyeong-Won Seo, Myung-Gil Choi
Yoo-Rim Nam, Jae-Hoon Kim
Korea Maritime University

요 약

기계학습은 학습말뭉치로부터 문제를 해결하기 위한 규칙을 학습하여 모델을 생성한다. 생성된 모델의 성능을 높이기 위해서는 문제에 적합한 자질들을 많이 이용해야 하지만 많은 자질들을 사용하면 모델의 생성시간은 느려지는 것이 사실이다. 이 문제를 해결하기 위해 본 논문에서는 다단계 기법을 적용한 기계학습으로 구묶음 시스템을 제작하여 학습모델의 생성시간을 단축하고 성능을 높이는 기법을 제안한다. 많은 종류의 자질들을 두 단계로 분리하여 학습하는 기법으로 1단계에서 구의 경계를 인식하고 2단계에서 구의 태그를 결정한다. 1단계의 학습자질은 어휘 정보, 품사 정보, 띄어쓰기 정보, 중심어 정보를 사용하였으며, 2단계 학습자질은 어휘 정보와 품사 정보 외에 1단계 결과에서 추출한 구의 시작 품사 정보와 끝 품사 정보, 구 정보, 구 품사 정보를 자질로 사용하였다. 평가를 위해서 본 논문에서는 ETRI 구문구조 말뭉치를 사용하였다.

주제어: 기계학습, 학습모델, 구묶음, 자질

1. 서론

컴퓨터의 하드웨어 기술의 발전으로 기계학습을 이용해서 다양한 문제를 효과적으로 해결하고 있다. 그러나 학습 자료나 그 자질의 수가 많을 경우 학습시간이 지나치게 오래 걸린다. 본 연구팀에서는 구묶음 문제를 기계학습을 이용하고자 한다. 학습 말뭉치로는 ETRI 구문구조 말뭉치[1]를 이용하고 학습자질로 어휘 정보, 품사 정보, 띄어쓰기 정보, 구 정보를 사용하여 모델을 학습했으나(CRF++[1]) 학습시간이 너무 오래 걸려 좋은 결과를 얻을 수 없었다. 본 논문에서는 이를 해결하기 위해 다단계 기계학습 기법을 사용한다. 본 논문에서 구묶음 문제를 구 경계 인식과 구 품사 결정 단계로 나누었으며 이를 다단계 기계학습 기법이라고 한다.

본 논문은 2장에서 관련연구로 기계학습을 이용한 구묶음 시스템들을 소개한다. 3장과 4장에서 제안된 다단계 기계학습 기법과 실험 결과를 기술한다. 끝으로 5장에서 결론을 맺고 향후 과제에 대해 논의한다.

2. 관련연구

구묶음은 문장에서 겹치지 않는 덩어리로서 구문트리의 하위트리를 형성하는 것이라고 정의하였다[2]. 즉, 구묶음은 여러 단어를 하나의 구문 단위로 묶어서 구문분석에서 하나의 단위로 간주하므로 구문분석의 복잡도를 크게 개선할 수 있다. 기계학습을 통한 구묶음 시스템을 제작하는 기법은 그 종류가 다양하며 크게 구 경계에 괄호를 삽입하는 기법[3], 구의 정보를 포함하는 추가 품사를 이용하는 기법[4] 등이 있다.

구 경계에 괄호를 삽입하는 기법은 아래와 같이 구의 시작과 끝을 괄호로 표시하는 기법이다.

생각하/eyVBB +
[르/eyEEI + 수/eyNDF + 있/eyVAB] + 는/eyEEI +
처지/eyNNF + 이/eyPOC + 라면/eyEEG [2]

이 기법은 부가적으로 표시되는 정보를 최소화하면서 기반 명사구(base noun phrase) 정보를 표시할 수 있다는 장점이 있지만 말뭉치를 추가적으로 가공해야 한다는 단점이 있다.

한편, 구의 정보를 포함하는 추가 태그를 이용한 기법은 다양한 형태의 구 경계 태그[5-7]를 이용하며 일반적으로 B(Begin), I(In), O(Out) 태그가 널리 사용된다. B(Begin)는 구의 시작으로 표시하고 I(In)과 O(Out)는 각각 구의 내부에 포함된 경우와 그렇지 않은 경우를 표시한다.

[3]에서는 명사구 인식을 위해 B, I, O 정보 외에 띄어쓰기 정보를 자질로 사용했다. 띄어쓰기 정보는 다음 형태소와 같은 어절에 속하는 경우에 '+'로, 어절의 끝인

1) http://crfpp.sourceforge.net

2) 태그들의 자세한 설명은 [1]을 참고하시오.

경우는 '.'로, 하나의 형태소로 이루어진 어절인 경우에는 '*'로 표시하였다. 뿐만 아니라 한국어의 특성을 고려하여 어절의 중심어 정보도 자질로 사용하였다. 기능어의 중심어와 내용어의 중심어는 각각 fh, ch로 표시하였고, 중심어가 아닌 형태소는 fx, cx로 표시하였다. 또한 기능어와 중심어가 동일한 형태소는 위치에 따라 cf, fc로 나누어 표시하였다.

[8]에서는 구묶음 기반 의존명사 처리 기법을 제안하고 있다. 의존명사란 자립성이 없는 특수한 명사를 일컬으며 그 앞에 어떤 한정 성분이 나타나지 않으면 홀로 쓰일 수 없는 비자립적인 명사를 의미한다. 여기서는 의존명사를 크게 단위 의존명사와 비단위 의존명사구 구분하여 처리기법을 제안하고 있는데 단위 의존명사와 비단위 의존명사는 각각 6개의 규칙을 고안하여 구묶음 처리를 하고 있으며, 충돌오류에 대해 3가지 규칙을 이용하여 오류를 해결하고 있다.

3. 다단계(Cascade) 기계학습 기법

일반적으로 구묶음 문제를 기계학습으로 해결하기 위해서는 구 경계 태그(BIO tag)와 품사 태그를 결합하여 하나의 태그를 사용한다. 결합된 태그는 품사 태그의 수에 약 3배가 되며 이는 기계학습의 속도를 떨어뜨리는 커다란 원인이 된다. 또한 태그의 수가 많기 때문에 성능을 저하시키는 원인이 되기도 한다. 본 연구에서는 이를 해소하기 위해 구 경계를 먼저 인식하고(1단계), 인식된 구에 대한 품사를 결정한다(2단계). 이와 같은 방법을 본 논문에서는 다단계 기계학습이라고 한다. 이 경우 태그의 수가 늘어나지 않으며 이로 인해 학습 속도를 크게 개선할 수 있을 것이다.

3.1 기계학습을 통한 구묶음

본 논문에서는 기계학습을 통한 구묶음 모델을 생성하기 위해 ETRI 구문구조 말뭉치[1]를 사용한다. 이 말뭉치에는 한 문장에 대해 품사과 구묶음 그리고 구문분석 정보가 부착되어 있다. 본 논문에서는 품사와 구묶음 정보를 이용해서 학습에 필요한 자질을 추출한다.

가. 1단계 학습자질 추출

<표 1> 구 경계 인식(1단계)을 위한 학습자질의 예

어휘	품사	공백	중심어	태그
편안하	eyVAB	−	ch	O
게	eyEEJ	+	fh	O
생각하	eyVBB	−	ch	O
ㄹ	eyEEI	+	fh	B
수	eyNDF	+	cf	I
있	eyVAB	−	ch	I
는	eyEEI	+	fh	O
처지	eyNNF	−	ch	O
이	eyVFD	−	fc	O
라면	eyEEG	+	fh	O

1단계 학습자질은 어휘 정보, 품사 정보, 공백 정보, 중심어 정보이며 그 예는 <표 1>이다. 어휘 정보와 품사 정보의 자질값은 각각 형태소와 그 품사이다. 공백 정보의 자질값은 형태소 다음에 띄어쓰기 유무에 따라 `+'와 `−'이다. 중심어 정보의 자질값은 [3]에서 사용한 중심어 정보는 그대로 사용한다.

나. 2단계 학습자질 추출

구 품사 결정 단계(2단계)에서 기본 단위도 구 경계 인식 단계와 마찬가지로 형태소이며 구의 품사를 형태소에 부여하는 것은 문제가 있다. 본 논문에서 이 문제를 구의 시작 형태소에 구의 정보를 부여하는 방법으로 해결하였다(<표 2 참조>). 2단계 학습자질은 어휘 정보, 품사 정보, 구의 시작과 끝의 형태소 품사 정보, 그리고 1단계에서 인식된 구 경계 정보이다.

<표 2> 구 품사 결정(2단계)을 위한 학습자질의 예

어휘	품사	시작품사	끝 품사	구 정보	BIO	태그
편안하	eyVAB	N	N	N	O	N
게	eyEEJ	N	N	N	O	N
생각하	eyVBB	N	N	N	O	N
ㄹ	eyEEI	eyEEI	eyVAB	ㄹ_수_있	B	eyEVK
수	eyNDF	N	N	N	I	N
있	eyVAB	N	N	N	I	N
는	eyEEI	N	N	N	O	N
처지	eyNNF	N	N	N	O	N
이	eyVFD	N	N	N	O	N
라면	eyEEG	N	N	N	O	N

4. 실험 및 분석

앞에서 언급했듯이 실험에 사용된 말뭉치는 ETRI 구문구조 말뭉치이다. 평가를 위한 학습 말뭉치과 실험 말뭉치로 나뉘었고 학습 말뭉치와 실험 말뭉치는 각각 56,936 문장과 5,756문장으로 구성되었다. 기계학습 방법을 1단계와 2단계 모두 CRF++을 사용하였다. 측도는 정확률(P)과 재현률(R)을 사용하며 아래와 같이 구한다.

$$\text{정확률(P)} = \frac{\text{제시한 구묶음(품사) 정답수}}{\text{출력한 구 전체 수}}$$

$$\text{재현율(R)} = \frac{\text{제시한 구묶음(품사) 정답수}}{\text{말뭉치 구 전체 수}}$$

<표 3>은 제안된 시스템의 성능 평가 결과이다. 구 경계 인식 단계에서는 정확률과 재현율이 각각 84.5%와 89.6%였으며 구 품사 결정 단계에서는 정확률과 재현율이 각각 83.8%과 91.0%였다.

<표 3> 성능 평가

	1단계	2단계
정확률	87.3%	83.8%
재현율	88.9%	91.0%

본 논문에서 제안된 시스템은 초벌 시스템으로 충분한 성능을 보이지 못했으며 여러 가지 면에서 개선의 여지가 있다.

5. 결론

본 논문에서는 다단계 기계학습을 통해 구묶음 시스템을 제안했다. 이 방법은 학습시간을 크게 줄일 수 있었으며 시스템에서 결정해야 하는 태그의 수가 많은 경우에 매우 유용한 방법이다. 성능에서도 개선될 수 있을 것을 예상되지만 동일한 말뭉치로 성능을 비교하지 못하여 단정할 수는 없다. 향후 연구과제를 통해서 두 시스템의 성능도 비교해볼 계획이다.

참고문헌

[1] 김재훈 외, "구문구조 부착 말뭉치 구축", 모비코앤시스메타㈜, 최종보고서, 2005.

[2] Steven Abey, Chunks and dependencies: Bringing processing evidence to bear on syntax. In Jennifer Cole, Georgia Green, and Jerry Morgan, editors, Computational Linguistics and the Foundations of Linguistic Theory, CSLI, pp. 145-164, 1995

[3] 서충원 외, "어절의 중심어 정보를 이용한 한국어 기반 명사구 인식", 제15회 한글 및 한국어 처리 학술대회 발표논문집, pp. 145-151, 2003.

[4] 양재형, "규칙 기반 학습에 의한 한국어의 기반 명사구 인식", 정보과학회논문지 제27권 제10호, pp. 1062-1071, 2000.

[5] Lance A. Ramshaw and Mitchell P. Marcus, Text Chunking Using Transformation-Based Learning, In: "Proceedings of the Third Workshop on Very Large Corpora", Cambridge, MA, USA, 1995.

[6] Erik F, Tjong Kim Sang, "Representing text chunks.", EACL'99, 1999.

[7] Erik F, Tjong Kim Sang, "Memory-Based Shallow Parsing.", Machine Learning Research volume2, 2002.

[8] 박의규 외, "한국어 구문분석을 위한 구묶음 기반 의존명사 처리", 인지과학, 제17권 제2호, pp.119-138, 2006.

[9] 황영숙 외, "자질집합선택 기반의 기계학습을 통한 한국어 기본구 인식의 성능 향상", 정보과학회논문지, 제29권 제9호, pp. 654-668, 2002.