개체명 부착 말뭉치에서 자동 오류 수정

김중한O, 최윤수, 박태호, 차정원
창원대학교
evilangel1@changwon.ac.kr, cyunsu77@changwon.ac.kr, taehope@changwon.ac.kr, chajeongwon@gmail.com

Automatic Named-Entity Corpus of Error Correction

Jung-Han KimO, Yun-Su Choi, Tae-Ho Park, Jeong-Won Cha
Changwon National University

요 약

자연어처리에서 학습 말뭉치는 매우 중요한 부분이다. 본 논문에서는 RDR(Ripple-Down Rules)을 이용하여 개체명 태그의 오류를 수정하는 방법을 제안한다. 규칙 생성에 자질 정보를 사용하기 위해 RDR을 확장하여 템플릿을 사용 가능하게 확장하였다. 학습과 테스트에 블로그 문서를 수집하여 제작한 말뭉치를 사용하며, 소량의 문서를 학습하고 대량의 문서를 평가함으로 일반적인 상황에서도 오류 수정의 효과와 효율을 검증하였다.

1. 서 론

학습 말뭉치는 기계학습에 있어 매우 중요한 부분으로 인식되고 있다. 이는 기계학습의 예측 성능이 학습 데이터에 따라 큰 차이를 보이기 때문이다. 따라서 학습 데이터의 오류를 줄이기 위한 방법에 대한 연구들이 이루어지고 있다[1,2,3,4].

자연어처리 분야에서 학습 말뭉치의 부족으로 인해 발생하는 자료희귀문제(Data Sparseness Problem)를 회피하기 위해 대량의 말뭉치가 필요하게 된다. 하지만 대량의 말뭉치를 제작하는 것은 많은 시간과 비용이 요구된다. 이러한 이유로 학습에 사용되는 학습 말뭉치에 태그가 부착된 말뭉치만 사용하는 지도 학습(Supervised Learning)을 대체하기 위해 비지도 학습(UnSupervised Learning)이나 반지도 학습(Semi-Supervised Learning)에 대한 연구도 진행되었다. 그러나 이러한 연구에도 불구하고 학습 말뭉치의 중요성은 줄어들지 않고 있다[5].

개체명 말뭉치를 제작하기 위해서는 다수의 사람이 직접 수작업으로 제작하거나 개체명 부착 도구를 사용하는 방법으로 말뭉치 제작에 소모되는 비용을 줄인다. 하지만 말뭉치를 제작하는 사람의 숙련도와 개인의 편차로 인해 오류가 포함된다. 이러한 오류는 동일한 개체명에 대한 일관성을 유지하기 힘들게 한다. 따라서 학습 단계에서 문제를 발생시켜 정확도를 떨어뜨리는 요인 된다.

본 논문에서는 개체명 말뭉치 제작 시 잘못 부착된 개체명 오류를 자동 오류 수정 시스템을 통해 수정한다. 그래서 말뭉치 제작에 소모되는 비용을 줄이고 정확한 말뭉치 구축을 목표로 한다.

2. 제안방법

본 연구에서는 여러 연구자들이 손으로 직접 제작한 개체명 말뭉치와 해당 말뭉치를 전문가가 수정한 개체명 말뭉치를 사용한다. 이 두 개의 말뭉치를 RDR 학습에 사용하여 오류 개체명을 자동으로 수정하는 방법을 제시한다.

2.1. 학습방법

[6]에서는 Brill이 제안한 Brill's Tagger[7]의 규칙 생성 알고리즘에 SCRDR(Single Classification Ripple Down Rules)을 추가하여 학습을 진행하였다. 수정 규칙 생성을 위해 초기 말뭉치와 정답 말뭉치를 비교하여 같은 형태소에 다른 품사가 부착된 경우, 이를 학습하여 틀린 품사를 정답으로 수정할 수 있는 규칙을 생성한다. 이렇게 생성된 오류 수정 규칙을 사용하여 오류 수정을 진행할 수 있다. 하지만 이 연구는 형태소 품사를 수정하는 방법이다.

형태소의 오류가 나타나는 유형과 개체명 오류가 나타나는 오류의 유형은 다르다. 그 이유로 형태소 수정은 모든 형태소가 반드시 하나의 품사를 부착하고 있지만 개체명 수정에서는 단어가 개체명 태그를 반드시 가진다고 볼 수 없다. 본 연구에서는 개체명 오류의 유형을 분석하고 <표 1>과 같이 4 가지로 분류하였다.

표 1. 개체명 말뭉치에서 나타나는 오류 유형

유형	설명
1	동일한 단어에 서로 다른 개체명
2	초기 말뭉치에만 존재하는 개체명
3	정답 말뭉치에만 존재하는 개체명
4	서로 다른 경계를 가지는 개체명

오류 유형 1은 개체명을 인식했으나 잘못된 개체명 태그를 부여한 경우이고, 유형 2는 개체명으로 인식한 단어가 개체명이 아닌 경우이다. 유형 1, 유형 2는 개체명 인식에 대한 오류보다는 분류에 대한 오류라고 볼 수 있다. 왜냐하면 'Null'도 하나의 개체명 태그라고 생각할 때, 'Null'이 아닌 다른 태그를 부착했다고 생각할 수 있기 때문이다. 이와 다르게 유형 3은 개체명인 단어를 인

식하지 못한 경우이다. 또한 유형 4는 개체명의 일부분만을 인식해 잘못된 개체명이다. 그래서 유형 3과 유형 4는 개체명 인식의 문제로 볼 수 있다.

본 연구에서 제안하는 규칙을 기반으로 하는 모델에서는 개체명 인식보다는 개체명 분류의 문제에서 보다 나은 성능을 보인다. 따라서 유형 1, 유형 2가 오류 수정의 대상이 된다. 하지만 유형 3에 대한 오류를 개체명 사전을 사용한 태깅을 통해 유형 1 또는 유형 2의 문제로 변경할 수 있다. 이를 위해 개체명 사전과 확장 개체명 사전을 사용해 초기 말뭉치에 추가적인 개체명 태그를 부착하였다. 그러나 유형 4는 다른 유형의 오류에 비해 나타나는 빈도가 미미하고 개체명 사전으로 처리하기에는 한계성이 있으므로 제외한다.

2.2. 자질 선택

RDR 규칙 생성을 위해 말뭉치를 자질 집합으로 변환하는 자질 파일(Feature file)을 생성한다. 자질 정보는 <표 2>와 같이 사용한다.

표 2. 자질 파일(Feature file) 생성에 사용하는 자질

순번	자질	설명
1	형태소/품사	
2	조사/품사	형태소가 개체명 일 때 동일한 어절에 있는 조사
3	이전 어절의 형태소/품사	
4	다음 어절의 형태소/품사	
5	지배소/품사	형태소의 지배소 형태소와 품사
6	개체명	형태소가 개체명의 일부분 일 때 개체명 단어

3. 실험 및 토의

개체명은 ETRI에서 정의한 15개의 카테고리에 대한 개체명 태그를 사용한다. 규칙 학습을 위해 규칙 선택 임계치(threshold)는 1을 사용한다. 임계치를 1로 선택한 이유는 개체명이 나타나는 빈도가 높지 않으므로 보다 많은 규칙을 생성하기 위함이다. 기본 실험을 진행하여 최적의 템플릿을 구하고 학습을 진행하였다. 실험은 소량의 문서로 학습하여 대량의 평가 말뭉치를 분석한다.

3.1. 실험 환경

본 논문에서는 실험을 위해 두 개의 블로그 문서를 사용하였다. 학습을 위해서 총 3,500문장의 블로그 문서를 사용하고 평가를 위해서 총 5,438문장의 블로그 문서 사용하였다. 학습과 평가에 사용한 3,500문장과 5,438문장의 블로그 문서에 대한 정답 개체명 수와 유형별 오류 개체명 수와 오류율은 <표 3>과 같다. 또한 오류율은 (식 1)과 같이 계산하였다.

$$\text{오류율} = \left(1 - \frac{\text{정답 개체명 수}}{\text{총 개체명 수} + \text{오류 유형 3 개체명수}}\right) \times 100 \qquad \text{(식1)}$$

초기 말뭉치의 정답 개체명 수 6,222개와 총 개체명 수 6,538개, 오류 유형 3의 개체명 수 161을 (식1)을 사용해 계산하면 오류율은 7.12%가 된다.

표 3. 블로그 문서의 개체명 수와 오류율

		학습문서		평가문서	
		초기	정답	초기	정답
문장 수		3,500	3,500	5,438	5,438
총 개체명 수		6,538	6,531	10,447	10,396
정답 개체명 수		6,222	6,531	9,951	10,396
오류 개체명 수 (비율)	유형1	100 (21%)		128 (18%)	
	유형2	168 (35%)		276 (38%)	
	유형3	161 (34%)		225 (31%)	
	유형4	48 (10%)		92 (13%)	
오류율 (단위 : %)		7.12		6.76	

3.2. 사전을 이용한 개체명 부착

오류 유형3은 RDR을 이용하여 생성한 규칙으로는 수정이 용이하지 않다. 이를 보완하기 위해 개체명 사전을 이용하여 개체명을 부착하는 전처리 과정을 거친다. 개체명 사전은 평가 문서의 초기 말뭉치에 부착되어 있는 개체명을 사용하여 제작된다. 사전은 해당 개체명의 태그와 개체명이 포함된 어절의 바로 앞뒤 어절 정보를 함께 사용한다. 확장 사전은 해당 개체명의 태그와 개체명의 뒤에 오는 괄호 기호 정보를 함께 사용한다. 개체명 부착방법은 우선 사전에 있는 개체명과 일치하는 형태소를 탐색한다. 사전은 그 형태소가 속한 어절의 앞뒤 어절이 사전에 있는 개체명의 앞뒤 어절 정보와 일치하는 경우에만 개체명 태그를 부착한다. 확장 사전은 그 형태소의 뒤에 괄호 기호가 나타나면 괄호 기호로 싸인 단어에 개체명을 부착한다. 사전을 사용해 개체명을 부착하는 방법은 개체명 미인식에 대한 오류를 일부 수정할 수 있다. 하지만 단어 정보만을 이용하기 때문에 인식 오류를 증가시킨다. <표 4>는 사전을 이용한 개체명 부착 방법에 대한 성능을 분석한 결과이다.

3.3. 실험 결과

실험은 최소한의 학습 데이터로 더 큰 데이터를 평가하기 위해 작은 양의 문장을 학습하고 많은 양의 문장을 평가하였다. 하지만 작은 양의 문장에 나타나는 개체명은 빈도가 낮아 임계치(threshold)를 낮게 조정하여 학습하였다. 임계치를 높게 설정할 경우 생성되는 규칙이 적어 수정되

는 개체명이 거의 나타나지 않았다. 또한 낮은 임계치는 오류만 학습하는 기존의 RDR 시스템[6]에서는 오히려 오류를 발생시키는 규칙도 생성하는 문제점이 나타났다. 이를 해결하기 위해서 오류와 정답을 모두 학습해 규칙을 생성하게 하였다.

표 4. 개체명 사전을 이용한 개체명 부착 성능

		초기	사전 적용 후	확장 사전 적용 후
총 개체명 수 (증감)		10,447	10,614 (+167)	10,652 (+205)
정답 개체명 수 (증감)		9,951	10,002 (+51)	10,005 (+54)
오류 개체명 수 (증감)	유형1	128	128 (+0)	136 (+8)
	유형2	276	390 (+114)	416 (+140)
	유형3	225	172 (-53)	159 (-66)
	유형4	92	94 (+2)	95 (+3)
오류율 (단위 : %)		6.76	7.27	7.46

사전과 확장 사전은 평가 문서의 초기 말뭉치를 사용하여 사전을 생성한다. 그 결과를 RDR 학습과정에서 생성되는 규칙을 사용해 초기 말뭉치의 오류를 수정하고 평가 정답 말뭉치와 비교하여 성능을 측정하였다. <표 5>는 평가 문서의 성능 분석표이다.

표 5. RDR 시스템 성능

		초기	RDR 적용		
			사전 미사용	사전	확장 사전
총 개체명 수 (증감)		10,447	10,350 (-97)	10,350 (-97)	10,422 (-25)
정답 개체명 수 (증감)		9,951	9,987 (+36)	10,037 (+86)	10,041 (+90)
오류 개체명 수 (증감)	유형1	128	92 (-36)	93 (-35)	100 (-28)
	유형2	276	59 (-217)	160 (-116)	186 (-90)
	유형3	225	225 (0)	172 (-53)	159 (-66)
	유형4	92	92 (0)	94 (+2)	95 (+3)
오류율 (단위 : %)		6.76	4.49	4.61	5.10

3.4. 토의

개체명 말뭉치를 형태소 단위로 매핑된 자질 파일(Feature file)을 생성하여 학습과 평가 모두에 사용한다. 이는 개체명의 경계 인식에 대한 문제를 제외하기 위한 방법으로 오류 유형 1, 유형 2에 대해서만 생성한 규칙을 적용하기 위함이다. 또한 개체명 사전을 사용하는 방법으로 개체명 인식의 문제점인 오류 유형 3에 대해서 보완하였다. 평가 말뭉치와 학습 말뭉치에서 나타나는 개체명이 서로 다르고, 다른 주변 정보를 많이 포함하고 있었다. 이 때문에 오류 유형 1에 대한 수정이 오류 유형 2에 대한 수정보다 적은 수가 수정되는 결과로 이어졌다. 그리고 개체명 사전을 사용하는 개체명 인식 과정을 통해 오류 유형 3도 감소된 결과를 얻을 수 있었다. 하지만 한 번도 문서 내에 나타나지 않은 개체명에 대해서는 개체명 인식이 불가능했다.

5. 향후 연구 및 결론

본 연구에서는 개체명 태그 부착 말뭉치의 오류를 자동으로 수정하는 방법을 제시하였다. 성능은 3,500문장을 학습하고 5,438문장을 평가했을 때 2.27%의 성능이 향상되었다. 개체명 사전과 확장 개체명 사전을 사용하였을 때 개체명 사전의 목적인 오류 유형 3의 수가 감소되었지만 오류율은 증가하였다. 따라서 RDR 학습 과정에서 유형 1, 유형 2을 정답으로 수정할 수 있는 규칙이 학습되면 개체명 사전이 효용성을 가질 수 있다.

이번 연구에서 개체명 부착 말뭉치를 수정하기 위해 사용된 자질 정보 이외의 추가적인 자질 정보에 대한 연구가 필요하다. 또한 개체명 사전을 사용한 개체명 인식의 정확도를 향상시킬 수 있는 방법에 대한 연구도 진행되어야 한다.

참고문헌

[1] Edwards. G, and Compton. P, “Peirs : A pathologist maintained expert system for the interpretation if chemical pathology reports”, Pathology 25, pp.27-34, 1993.

[2] J. Hong, J. Cha, “Error Correction of Sejong Morphological Annotation Corpora using Part-of-Speech Tagger and Frequency Information”, Journal of KIISE SA, ISSN.1226-2285, VOL.40, NO.7, pp.417-428, 2013.

[3] Wu. X, “Knowledge acquisition from database”, Ablex Publishing Corp USA, 1995.

[4] Zhu. X, Wu. X, Chen Q, “Eliminating Class Noise in Large Datasets”, Proceedings of the 20th ICML International Conference on Machine Learning, pp.920-927, 2003.

[5] Edwards. G, and Compton P, “Experience With Ripple-Down Rules“, AI 2005 SI, Vol.19, Issue.5, pp.356-362, 2006.

[6] Dat Quoc Nguyen, Dai Quoc Nguyen, Dang Duc Pham, Son Bao Pham, “RDRPOSTAgger : A Ripple Down Rules-based Part-Of-Speech Tagger“, EACL, p.17-20, 2014.

[7] Eric Brill, “Transformation-based error-driven learning and natural language processing : a case study in part-of-speech tagging“, Comput. Linguist, pp.543-565. 1995.