Rough Set을 이용한 형태소 품사 태깅 코퍼스 오류 정량화

박태호○, 박다솔, 신창욱, 박성재, 차정원
창원대학교 컴퓨터공학과
{taehope, dasol_p, papower1, tjdwo1289, jcha}@changwon.ac.kr

Annotated Corpus Error Detection Using Rough Set

Tae-Ho Park○, Da-Sol Park, Chang-Uk Shin, Seong-Jae Park, Jeong-Won Cha
Changwon National University Dept. Computer Engineering

요 약

자연어처리에서 언어정보 부착 말뭉치는 학습과 평가에 매우 중요하다. 언어정보 부착 말뭉치의 일관성과 무결성이 보장되지 않는다면 학습의 효율을 저하시키고 평가의 정당성은 보장되지 못할 것이다. 본 논문에서는 언어정보가 부착된 코퍼스의 일관성을 검사하여 코퍼스의 오류 정도를 정량화시킬 수 있는 알고리즘을 제안한다. 언어정보가 부착된 코퍼스에서 자질을 생성하고, rough set을 이용하여 일관성 검사를 한다. 일관성에 위배된 부분은 오류가 발생한 부분이거나 상황에 따라서 다르게 분석되어야 하는 부분이다. 다양한 품사부착 코퍼스에 제안 알고리즘을 적용하여 실제 오류와 유사한 오류를 찾을 수 있었다. 제안 알고리즘을 이용하면 언어 정보를 부착한 말뭉치의 오류 정도를 정량화할 수 있다.

1. 서 론

기계학습을 이용한 문제해결이 점점 더 많은 분야로 확대되고 있다. 기계학습에서 예측 성능은 학습 데이터의 오류에 영향을 받는다. 따라서 학습 데이터의 오류를 줄이기 위한 방법에 대해서 많은 연구들이 있었다[1-9].

자연어처리에서도 통계 정보에 기반하는 방법과 기계학습을 이용하는 방법이 주류를 이루고 있다. 이 두 방법에서는 학습을 위한 말뭉치가 매우 중요한 역할을 담당한다. 학습 말뭉치는 작성하는데 시간과 비용이 많이 요구된다. 이러한 이유로 인해 지도학습(supervised learning)을 대체하는 비지도학습(unsupervised learning)이나 반지도학습(semi-supervised learning)에 대한 연구가 많이 진행되었다.

비지도학습과 반지도학습의 성공적인 결과에도 불구하고 학습을 위한 정보부착 말뭉치의 중요성은 줄어들지 않고 있다.

대량의 말뭉치를 제작하려고 하면 다수의 사람들이 작업을 하기 때문에 일관성 있는 말뭉치를 제작하기가 매우 어렵다. 세종 말뭉치에 대한 많은 오류 수정 논문들이 이를 증명한다[1, 2].

학습 데이터의 오류 수정에 관한 논문들은 다음과 같다. [3]은 학습 데이터의 오류를 다음과 같이 크게 두 가지로 분류하였다. 1) 속성 오류(attribute noise), 2) 범주 오류(class noise)이다. 속성 오류는 속성값을 입력하는 중에 발생하는 오류이다. 여기에는 속성이 없거나 중복된 값이 있는 경우가 포함된다. 범주 오류는 다시 a)일관성 오류(contradictionary)와 b)분류 오류(misclassification)로 나누어진다. 일관성 오류는 같은 데이터가 다른 범주로 분류된 경우이다. 분류 오류는 같은 데이터가 일관되게 잘못된 범주로 할당된 경우이다. 예를 들어 '나는'에서 '나'는 대명사인데 말뭉치 전체에서 명사로 할당된 경우이다. 자연어처리를 위한 말뭉치의 경우는 범주 오류에 해당된다.

[4, 5]는 대용량, 분산 데이터에 촛점을 맞추었다. 그들은 학습을 수행하는 데이터를 분류 알고리즘이 처리 가능한 보다 작은 크기로 분할(partition)하였다. 분할된 데이터 집합 각각에 대해서 분류기를 생성하였다. 전 데이터 집합에서 오류로 판명되면 빈도수를 증가시키게 되고 높은 확률을 가지는 개별 항목은 최종 오류가 된다. 저자는 오류 항목을 발견하고 삭제하기 위해 최대치(majority)와 일치(non-objection) 전략을 사용하였다.

[6]에서는 최대 정보량 기준을 사용하였다. [7]은 포화 필터(saturation filter)라고 불리는 방법을 사용하였다. [4, 8]은 C4.5를 사용하여 잠재적으로 오류가 될 항목을 구별하였다. [9]는 인공 신경망을 사용하였다. 이들은 모델에서 벗어난 항목들을 찾아서 제거하는 방법을 사용하였다. [10]은 같은 범주의 중심과 다른 범주와의 거리를 계산하는 식으로 모든 데이터들을 계산하고 주위의 범주를 계산하여 동일 범주에 속하지 않으면 제거한다. 평가는 분류기 등으로 오류를 임의로 생성하여 각 오류율에 따른 성능을 측정하였다.

그렇지만 자연어 처리에서는 데이터에서 오류가 있는 부분을 제거할 수가 없다. 이것을 모두 바르게 수정해야 한다. 또한 자연어 처리에서는 범주를 구성하는 구성원들의 유사도를 범주의 중심에서부터의 거리로 계산할 수가 없다.

본 연구에서는 rough set[11, 12]을 이용하여 언어정보 부착 말뭉치의 무결점 정도를 수치화할 수 있는 방법을 제시하였다.

2. 제안방법

본 연구에서는 다수의 연구자들이 손으로 작성한 품사 부착 말뭉치를 대상으로 실험하였다. 다수의 사람들에 의해서 생성된 말뭉치는 다양한 이유로 일관성에 문제가 발생한다. 이것은 지침이 부족해서 발생할 수도 있고 숙련도의 차이에 의해서 발생할 수도 있다. 언어정보 부착 말뭉치에서 본 연구에서는 일관성 오류(contradictionary)에 집중한다. 언어정보 부착 말뭉치에서 일관성 오류가 분류 오류보다 상대적으로 많고, 분류 오류는 말뭉치 내에서 오류와 정답을 비교할 수 없기 때문에 검출하기 어렵다. <표 1>은 초기 말뭉치에서 측정한 오류 수와 오류율이다. 이것은 전문가가 초기 말뭉치의 오류를 모두 수정한 후 비교한 것이다.

표 1. 실험 문서별 오류 분석 및 정답률

말뭉치	정답 어절 수	오류 어절 수	오류율(%)
1	13,093	260	1.99%
2	80,323	2,681	3.34%
3	6,003	156	2.60%

3. 실험 및 토의

3.1. 실험 방법

본 연구에서는 일부 오류가 남아있는 형태소 품사 정보 부착 말뭉치를 사용하였다. 말뭉치에 존재하는 오류를 분석하기 위해서 동일한 어휘의 어절과 동일한 자질을 지닐 때, 서로 다른 언어정보가 부착되어 있다면 오류로 간주하고 이를 분석하는 방법을 사용하였다. 다음의 <표 2>는 오류를 분석하기 위해 생성한 자질이다. 오류 분석에 사용된 자질은 각 어절의 첫 단어와 마지막 단어의 어휘와 형태소 품사 태그 정보를 사용하였다. 또한 -1 ~ +1 위치에서 동일한 자질을 관측하여 사용하였다.

표 2. 오류 분석을 위한 커널

	Windows		Kernel
Morp Feature	Morpheme	-1	First POS morpheme in eojeol, First POS tag in eojeol, Last POS morpheme in eojeol, Last POS tag in eojeol
		0	First POS morpheme in eojeol, First POS tag in eojeol, Last POS morpheme in eojeol, Last POS tag in eojeol
		+1	First POS morpheme in eojeol, First POS tag in eojeol, Last POS morpheme in eojeol, Last POS tag in eojeol

3.2. 실험 결과

본 논문에서 제안하는 알고리즘을 통해 각 말뭉치에서 다음 <표 3>과 같이 초벌 말뭉치의 오류를 예측하였다. 실제 말뭉치에 존재하는 오류와 비교하였을 때, 1번 말뭉치는 0.88%, 2번 말뭉치는 0.33%, 3번 말뭉치는 1.37% 차이가 발생하였다. 따라서 정답 말뭉치를 사용한 오류율과 거의 차이가 나지 않는 것을 알 수 있다.

표 3. 실험 문서별 오류 분석 및 정답률

말뭉치	오류율(%)	예측 오류 어절 수	예측 오류율(%)
1	1.19	271	2.07
2	3.34	2,419	3.01
3	2.60	74	1.23

표 4. 일관성 오류 유형

	오류 유형	
1	이/MM	이/NP
2	있/VA+습니다/EF+./SF	있/VV+습니다/EF+./SF
3	지난/MM	지나/VV+ㄴ/ETM
4	점/NNB+이/JKS	점/NNB+이/JKC
5	4/SN+만/NR+원/NNB+대/XSN	4/SN+만/NR+원/NNB+대/NNB

4. 결론

본 연구를 통해 언어정보 말뭉치에 존재하는 일관성과 무결성의 오류 정보를 정량화하였다. 언어정보가 부착된 말뭉치에서 자질을 생성하고, 이를 통해 말뭉치에서 나타나는 오류를 일관성 오류와 분류 오류로 나누어 측정하였다. 실험을 통해 최대 1.37% 이내의 오차가 발생하였다. 따라서 정답 말뭉치를 작성하지 않아도 초벌 태깅된 말뭉치의 오류를 예측할 수 있다. 향후 연구로는 현재 오류를 분석할 수 있는 말뭉치는 형태소 품사 부착 말뭉치에만 적용가능하기 때문에 이를 다양한 언어정보 부착 말뭉치에서 사용할 수 있도록 자질을 생성하는 템플릿과 각 말뭉치에 따라 나타날 수 있는 새로운 오류 유형에 적용 가능하도록 개선할 예정이다. 또한 오류 정량화를 통해 분석된 오류를 자동으로 수정할 수 있는 방안에 대한 연구도 진행예정이다.

사 사

이 논문은 2016년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임.
(No. R0101-16-0054, WiseKB: 빅데이터 이해 기반 자가학습형 지식베이스 및 추론 기술 개발)

참 고 문 헌

[1] 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, PAGE.417-428, 2013.

[2] M. Choi, H. Seo, H. Kwon and J. Kim, “Detecting and correcting errors in Korean POS-tagged corpora.”, Journal of the Korean Society of Marine Engineering, Vol.37, No.2, pp.227-235, 2013.

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

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

[5] Zhu. X., Wu. X. and Chen Q, “Bridging Local and Gobal Data Cleansing: Identifying Class Noise in Large”, Distributed Data Datasets, Data Mining and Knowledge Discovery, pp 275-308, Dec. 2006.

[6] Guyon, Isabelle, Matic. N. and Vapnik. V., “Discovering informative patterns and data cleaning”, Advances in Knpwledge Discovery and Data Mining, AAAI/MIT Press, pp 181-203, 1996.

[7] Gamberger, Dragan, Lavrac. N. and Groselj. C., “Experiments with noise filtering in a medical domain”, Proc. of 16th ICML Conference, pp. 143-151, San Francisco, CA. 1999.

[8] John, G. H., “Robust decision trees: Removing outliers from databases”, Proc. of the First International Conference on Knowledge Discovery and Data Mining, pp.174-179, AAAI Press. 1995.

[9] Zeng, Xinchuan and Martinez. T., “A noise filtering method using neural networks”, SCIMA 2003. IEEE International Workshop on Soft Computing Techniques in Instrumentation, Measurement and Related Applications, pp. 26-31. 17 May 2003.

[10] Edwards, G., and Compton, P., “Peirs: A pathologist maintained expert system for the interpretation of chemical pathology reports”. Pathology, Vol. 25, No. 1 , Pages 27-34, 1993.

[11] Pawlak, Zdzislaw. “Rough sets.“ International Journal of Computer & Information Sciences 11.5 341-356. 1982.

[12] Chouchoulas, Alexios, and Qiang Shen, “A rough set-based approach to text classification.“ New Directions in Rough Sets, Data Mining, and Granular-Soft Computing. Springer Berlin Heidelberg, 118-127. 1999.