

다중 작업 학습의 단계적 특징을 활용한 한국어 속성 기반 감성 분석에서의 대상 추출

박호민^o, 김재훈
한국해양대학교, 컴퓨터정보공학과
homin2006@hanmail.net, jhoon@kmou.ac.kr

Target extraction in Korean aspect-based sentiment analysis using stepwise feature of multi-task learning model

Ho-Min Park^o, Jae-Hoon Kim
Korea Maritime & Ocean University, Department of Computer Engineering

요 약

속성기반 감성 분석은 텍스트 내에 존재하는 속성에 대해 세분화된 감성 분석을 수행하는 과제를 말한다. 세분화된 감성 분석을 정확하게 수행하기 위해서는 텍스트에 존재하는 감성 표현과 그것이 수식하는 대상에 대한 정보가 반드시 필요하다. 그리고 순서대로 두 가지 정보는 이후 정보를 텍스트에서 추출하기 위해 중요한 단서가 된다. 따라서 본 논문에서는 KorBERT와 Bi-LSTM을 이용한 단계적 특징을 활용한 다중 작업 학습 모델을 사용하여 한국어 감성 분석 말뭉치의 감성 표현과 대상을 추출하는 작업을 수행하였다. 제안한 모델을 한국어 감성 분석 말뭉치로 학습 및 평가한 결과, 감성 표현 추출 작업의 출력을 추가적인 특성으로 전달하여 대상 추출 작업의 성능을 향상시킬 수 있음을 보였다.

주제어: 대상 추출, 속성기반 감성 분석, 다중 작업 학습

1. 서론

속성기반 감성 분석(aspect-based sentiment analysis)은 텍스트 내에 존재하는 속성에 대해 세분화된 감성 분석(fine-grained sentiment analysis)을 수행하는 과제를 말한다[1]. 특히 하나의 문장 또는 문서에 여러 개의 속성이 존재할 경우, 존재하는 모든 속성에 대한 감성적인 극성, 수식하는 대상, 발언한 보유자를 분석해야 한다. 예를 들어 “떡볶이는 맛있었는데 중얼원이 불친절했다.” 라는 식당 리뷰 문장에서는 두 가지 속성에 대한 속성기반 감성 분석이 가능하다. 그것은 음식 속성에 대해서는 보유자인 리뷰어가 대상인 떡볶이에 대해서는 긍정적이라는 것과 서비스 속성에서는 보유자인 리뷰어가 대상인 중얼원에 대해서는 부정적이라는 것이다.

위와 같이 하나의 문장 또는 문서에는 여러 개의 속성이 존재할 수 있다. 따라서 내부에 존재하는 감성 표현은 두 가지의 목적에 사용될 수 있다. 그것은 특정 속성에 대한 감성적 극성 분류 작업 또는 해당 감성적 표현이 수식하고 있는 대상(target) 탐지 작업의 단서가 되는 것이다. [2, 3]은 감성 사진을 사용하여 텍스트 내에 존재하는 감성 표현을 찾아낸다. 찾아낸 감성 표현을 기반으로 텍스트 내에서 자주 동시 등장하는 명사 연어를 추출하였다. [4]는 문장의 의존 구문 분석을 통해 감성 표현과 그것이 수식하는 대상을 분석했으며, [5]는 기계 번역 모델을 응용하여 명사와 형용사 사이의 관계를 모델링하여 대상을 추출했다. [6]은 감성 표현, 대상, 보유자(holder)가 단계적인 상응 관계에 있다는 가정을 바

탕으로 다중 작업 학습 모델에 적용하여 최종적으로 보유자 추출 작업의 성능을 향상시켰다.

본 논문은 텍스트 내에 존재하는 감성 표현 정보가 그것이 수식하는 대상을 식별해낼 수 있는 단서가 된다는 점을 응용하여 이러한 결과가 한국어 속성기반 감성 분석에서도 발생하는지 확인한다. 이를 위해 감성 표현 추출, 대상 추출이라는 두 가지 작업을 동시에 수행되도록 감성 표현 추출 작업의 출력을 대상 추출 작업의 입력으로 사용한다.

본 논문의 구성은 다음과 같다. 2장에서 한국어 감성 분석 말뭉치(Korean Sentiment Analysis Corpus, KOSAC)[7]와 다중 작업 학습 모델에 대하여 소개한다. 3장에서는 본 논문에서 제안하는 단계적 특징을 활용한 다중 작업 학습 모델에 대하여 기술하고 4장에서는 실험을 통해 감성 표현 추출, 대상 추출 작업의 성능을 평가 및 분석한다. 마지막으로 5장에서는 실험에 대한 결론을 맺고 향후 연구에 대하여 기술한다.

2. 관련 연구

이 장에서는 다중 작업 학습과 실험 및 평가 데이터로 사용될 KOSAC 두 가지를 설명한다.

2.1 다중 작업 학습

다중 작업 학습은 하나의 신경망이 여러 작업을 동시에 수행할 수 있도록 학습하는 접근법으로, 여러 문제들이 하나의 저레벨(low-level) 자질을 공유할 때 효과적

인 방법이다[8]. 다중 작업 학습은 강매개변수공유(hard parameter sharing) 방식[9]과 약매개변수공유(soft parameter sharing) 방식[10]이 존재한다. 강매개변수공유는 신경망 모델의 은닉 계층을 모든 작업들이 공유하고 각 작업에 최적화된 출력 계층을 갖는 구조이다. 이 방법은 다중 작업 학습의 가장 기본적인 구조로 모델이 각 작업마다 적절한 출력을 결정해야 하기 때문에 특정 작업에 과적합되는 것을 방지할 수 있다[11]. 약매개변수공유는 각 작업에 사용되는 별도의 신경망 모델을 사용하며, 각 모델의 매개변수 간의 차이를 정규화하는 방법으로 다중 작업 학습을 수행한다.

다중 작업 학습은 서로 다른 작업을 학습하는 과정에서 데이터에 존재하는 잡음(noise)을 무시 또는 최소화하여 일반화할 수 있다. 이를 통해 동시에 수행하는 여러 작업에 대해 일반화된 잠재적 표현을 학습할 수 있다[8].

본 논문에서는 감성 표현, 대상 추출 작업을 동시에 수행하기 위해 강매개변수공유 방법의 다중 작업 학습을 사용한다. 동일한 텍스트를 이용함과 동시에 감성 표현 추출 작업이 대상 추출 작업의 단서가 되기 때문이다. 공유층의 출력과 감성 표현 추출 작업 특화층의 출력을 입력받은 대상 추출 작업 특화층은 순차적인 학습을 진행함과 동시에 공통된 자질 속에서 작업에 특화된 자질을 학습하여 올바른 결과를 추출할 수 있다.

2.2 KOSAC

KOSAC은 세종 구문분석 말뭉치로부터 선별한 332개 신문 기사, 7,744개의 문장을 주석 대상으로 삼아 만들어진 한국어 감성 분석 말뭉치이다. 일반적으로 텍스트의 감성적 극성만을 분류하는 감성 분석이 아닌, 세분화된 감성 분석을 목적으로 만들어진 말뭉치이다. 따라서 상세하고 풍부한 주석을 위해 한국어 감정 주석 언어(Korean Sentiment Markup Language, KSMML)을 자체적으로 제작하여 서술자의 주관성 여부와 주관적인 표현들을 상세하게 주석하였다.

본 논문에서는 KOSAC의 주석 대상인 각 문장에 존재하는 감성 표현과 그것이 수식하는 대상을 학습 및 실험 데이터로 사용한다. [6]의 실험 내용과 달리, 한국어로 구성된 KOSAC에서는 보유자가 태깅되어 있지 않으므로 최종적인 목표를 보유자 추출이 아닌 대상 추출로 설정하였다. 그러므로 [2-6]의 내용 및 결론과 동일하게 한국어에서도 단계적인 특성이 대상 추출 작업에 효과적이지 확인한다.

3. 단계적 특징을 활용한 다중 작업 학습 모델

본 논문에서는 KOSAC을 이용한 감성 표현, 대상 추출을 동시에 수행하는 다중 작업 학습 모델을 제안한다. 이 장에서는 해당 모델의 구조와 다중 작업 학습 과정에 대해 설명한다. 그림 1은 제안하는 다중 작업 학습 모델

의 전체적인 구조를 보인다. 기본적인 강매개변수공유 모델(hard parameter sharing model) 형태를 따르고 있으나 감성 표현 추출과 대상 추출을 수행하는 작업 특화층의 출력이 순차적으로 다음 작업 특화층의 입력으로 연결된 것이 특징이다.

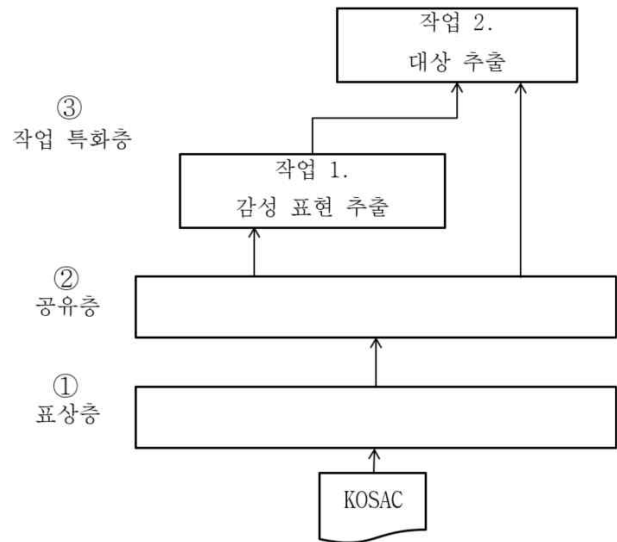


그림 1. 제안하는 다중 작업 학습 모델의 전체 구조

제안하는 다중 작업 학습 모델은 크게 세 개의 계층으로 구성된다. 첫 번째 계층은 표상층(embedding layer)으로 KOSAC 내의 문서를 KorBERT¹⁾로 통과시켜 표상화된 벡터로 변환한다. 두 번째 계층은 공유층으로 표상층의 출력을 입력으로 한다. 공유층은 하나의 Bi-LSTM[12] 구조로 구성되어 있다. 세 번째 계층은 감성 표현 추출, 대상 추출 작업에 특화된 두 개의 작업 특화층들이다. 각 작업 특화층은 하나의 LSTM[13]과 하나의 softmax 계층으로 이루어져 있다. 실제적으로 추출이 수행되는 계층이다.

4. 실험 및 평가

이 장에서는 본 논문에서 제안하는 대상 추출 방법론에 대한 성능 평가 및 평가를 진행한 실험 환경에 대하여 설명한다.

4.1 실험 환경

학습 및 실험에 사용한 데이터셋은 KOSAC을 채택하였다. 7,744개의 문장을 보유하고 있으며 그 중에서 2,654개의 문장이 주관적인 문장이고 나머지 5,090개의 문장이 객관적인 문장으로 구성되어 있다. 객관적인 문장이 더 많은 이유는 신문 기사를 사용했기 때문으로 보인다. 감성 표현과 대상 정보 쌍은 17,582개가 존재하며 한 문장 당 평균적으로 2.3개가 존재한다.

1) https://aiopen.etri.re.kr/service_dataset.php

KOSAC 데이터는 이미 형태소 분석까지 마쳐있는 상태이므로 3장에서 언급한대로 KorBERT-morpheme 모델을 이용하여 표상 벡터를 생성하였다. 학습 및 실험 데이터의 비율은 7:3으로 나누어 학습 데이터 12,307개, 실험 데이터 5,275개로 사용하였다.

4.2 단계적 특성 사용 여부에 대한 실험

본 논문에서는 3장에서 설명한대로 감성 표현 추출 작업 특화층을 제외한 대상 추출 작업 특화층에서는 이전 작업 특화층의 출력을 추가적인 입력으로 사용한다. 학습 데이터에 내재된 추가적인 자질을 공급받음으로써 해당 작업의 성능을 향상시키기 위함이다. 따라서 제안하는 작업 특화층으로 구성된 다중 작업 학습 모델과 공유층의 출력만을 입력으로 사용하는 작업 특화층으로 구성된 모델의 추출 정확도를 비교하여 해당 방법의 효용성을 확인하였다. 표 1은 두 작업의 추가적인 자질 사용 유무에 따른 정확도를 보인다.

표 1. 두 작업의 단계적 특징 사용 유무 정확도

작업 종류	이전 작업 특화층의 출력	
	미사용	사용
감성 표현 추출	71.32	-
대상 추출	65.59	71.68 (▲6.09)

구조 상 감성 표현 추출 작업에서는 실험이 불가능하지만 대상 추출 작업에서 이전 작업 특화층의 출력을 사용한 경우의 정확도가 상대적으로 높게 측정되었다. 6.09%의 상당한 정확도의 증가를 보였으므로 [2-6]의 결론과 동일하게 한국어에서도 감성 표현 정보가 대상 추출 작업에 있어 중요한 단서임을 확인하였다. 다만 모델의 전체적이고 종합적인 정확도가 [6]에 비해서는 그다지 10% 정도 낮은 수준으로 측정되어 아직 발전 및 추가적인 연구의 여지가 남아있다고 볼 수 있다. 이에 대한 원인으로서는 두 가지가 예상된다. 첫 번째는 신문 기사를 이루는 KOSAC의 범위로 인해, 식당 및 노트북의 리뷰를 기반으로 동일 모델을 사용한 [6]의 결과와 차이가 있을 수 있다는 것이다. 일반적으로 최대한 감성 및 감정적인 정보를 지우고 최대한 중립적으로 내용을 전달하려는 것이 신문 기사 분야이므로 동일 모델이라도 문제의 난이도가 더 높았을 것이란 예상이다. 두 번째로는 [6]의 결론과 같이 상대적으로 높은 문장 당 정보 쌍의 개수가 영향을 끼쳤을 수도 있다. [6]에서 한 문장에 존재하는 감성어, 대상, 보유자 쌍이 많을수록 정확도가 하락하는 결과를 보였다.

5. 결론

본 논문에서는 KOSAC을 이용하여 KorBERT와 Bi-LSTM을

결합한 다중 작업 학습 모델로 대상 추출 작업을 진행하였다. 결과적으로 이전 작업 특화층의 출력을 사용함으로써 확실한 성능의 향상을 보였다. 향후 연구로는 [6]과 같이 보유자 추출 작업을 한국어에서 진행할 수 있도록 실험을 진행할 예정이다.

참고문헌

- [1] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos and S. Manandhar, "SemEval-2014 task 4: aspect based sentiment analysis", *In Proceedings of the 8th International Workshop on Semantic Evaluation*, pp. 27-35, 2014.
- [2] M. Hu and L. Bing, "Mining opinion features in customer reviews", *In Proceedings of the 19th National Conference on Innovative Applications of Artificial Intelligence*, pp. 755-760, 2004.
- [3] A. M. Popescu and E. Oren, "Extracting product features and opinions from reviews", *In Proceedings of Human Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pp.339-346, 2005.
- [4] G. Qiu, B. Liu, J. Bu and C. Chen, "Opinion word expansion and target extraction through double propagation", *Computational linguistics*, vol. 37, no. 1, pp. 9-27, 2011.
- [5] K. Liu, X. Liheng and Z. Jun, "Opinion target extraction using word-based translation model" *In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Linguistics*, pp. 1346-1356, 2012.
- [6] H. M. Park and J. H. Kim, "Stepwise multi-task learning model for holder extraction in aspect-based sentiment analysis", *Applied Sciences*, vol. 12, no. 13, pp. 6777, 2022.
- [7] 김문형, 장하연, 조유미, 신호필, "KOSAC(Korean Sentiment Analysis Corpus): 한국어 감정 및 의견 분석 코퍼스", 2013년 한국컴퓨터종합학술대회 논문집, pp. 650-652, 2013.
- [8] S. Ruber, "An overview of multi-task learning in deep neural networks", *arXiv preprint arXiv:1706.05098*, 2017.
- [9] R. Caruana, "Multitask learning", *Autonomous Agents and Multi-Agent Systems*, vol. 27, no. 1, pp. 95-133, 1998.
- [10] L. Duong, T. Cohn, S. Bird and P. Cook, "Low resource dependency parsing: cross-lingual parameter sharing in a neural network parser", *In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pp. 845-850, 2015.

- [11] J. Baxter, “A Bayesian/information theoretic model of learning to learn via multiple task sampling” , *Machine Learning*, vol. 28, no. 1, pp. 7-39, 1997.
- [12] M. Schuster and K. K. Paliwal, “Bidirectional recurrent neural networks” , *IEEE transactions on Signal Processing*, vol. 45, no. 11, pp. 2673-2681, 1997.
- [13] S. Hochreiter and J. Schmidhuber, “Long short-term memory” , *Neural computation*, vol. 9, no. 8, pp. 1735-1780, 1997.