농식품 가격변동 요인분석을 위한 개체명 인식

박찬○, 이경순
전북대학교
{snrnsk5660, selfsolee}@jbnu.ac.kr

Named Entity Recognition for Analyzing Factors of Agrifood Price Fluctuation

Chan Park○, Kung-Soon Lee
Jeonbuk National University

요 약

농식품 가격을 안정적으로 제공하기 위해 농식품 가격 변동에 대한 요인 분석이 필요하다. 본 연구는 농식품 가격 변동의 요인 분석을 위해 인과관계 템플릿을 정의하고, 요약을 위한 개체명 인식 방법을 적용한다. 농식품 일일동향 데이터에 대한 평가에서 딥러닝 기반 BiLSTM-CRF 실험 결과 F1-점수 0.93으로 베이스라인 Bi-LSTM 실험 결과 0.75에 비해 높은 성능을 보였다.

주제어: 개체명 인식, 가격변동, 요인분석, 인과관계 템플릿, BiLSTM-CRF

1. 서론

농식품 계약 재배 담당 기관에서는 농식품 관련 정책, 가격변동, 재배면적, 생산량 등 재배 품목에 대한 정보를 통해 농민들과 계약재배를 한다. 예측에 실패하여 공급이 증가하면 가격 폭락을 대비하여 폐기하게 되고 공급이 감소하면 수급 불안 요인이 생기게 된다. 따라서 농식품 가격 변동 요인을 분석하고 연구할 필요가 있다.

실제로 2020년 3월 마늘 농사가 풍년이 들어 마늘 가격이 내려갈 것을 예상하여 공급량을 조절하기 위해 농민들이 애써 재배한 마늘밭을 갈아엎었다. 이는 농촌에서 매년 되풀이되는 모습으로 이와 같은 일이 발생하는 이유는 계약 재배 담당 기관에서 농식품 가격을 안정적으로 제공하기 위해 공급을 조절하다 실패하여 발생, 농민 개인이 정보가 부족하여 공급조절에 실패하여 발생한다. 이로 인해 2013년부터 2017년까지 산지 폐기로 땅에 묻힌 채소류 규모는 37만 톤으로, 폐기 비용만 약 450억 원에 가깝다. <그림1>에서 2019년에 풍년으로 가격이 폭락하였는데도 2020년에도 농민들이 마늘을 많이 수확하여 수확된 작물을 폐기하고 나서야 가격이 안정화 된 사례로 들 수 있다.

이와 같은 농산품 가격 요인을 분석하고 대비한다면 이익이 된다. 요인분석 하는 방법으로 인과관계 템플릿 구축을 제안한다. 그리고 농식품 가격변동요인 개체명 인식을 통해 데이터를 요약할 수 있는 기반을 제공한다.

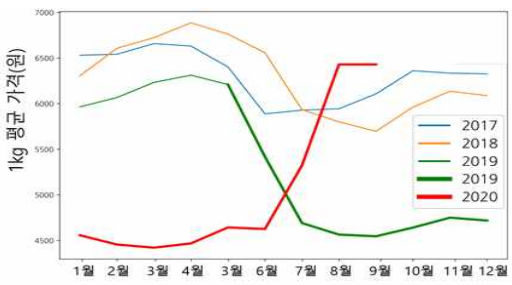

그림 1. 깐마늘(국산)_서울 년도별 평균가격

기존 연구에서는 농식품 가격변동을 여러 요인을 통해 실제 가격을 예측하였다. 과일 도매가격과 날씨 요인에 대한 상관관계 연구[1], LSTM네트워크를 활용한 농산물 가격예측모델[2], 인공신경망의 은닉층 최적화를 통한 농산물 가격예측 모델[3], 인공지능을 이용한 과일 가격 예측 모델 연구[4] 등의 연구가 있다. 본 연구에서 나온 템플릿의 구성과 비교하면 실제 수치로 표현하기 애매한 소비심리와 같은 요인에 대해서는 텍스트로서 요약하여 정보를 전달하는 것도 도움이 되는 정보이다.

본 연구에서는 현재 우수한 성능을 보이는 인공신경망 딥러닝 기술인 BiLSTM-CRF[5]을 이용하고 BIO 표기법으로 개체명 인식기를 모델링하여 태그를 분류한다. 또한 가격변동 요인 인과관계의 정의가 없어 직접 데이터를 분석하여 템플릿을 구축한다.

2절에서 농식품 가격변동요인 인과관계에 대한 템플릿을 정의하고 3절에서는 개체명 태그를 정의하고 4절에서는 학습 및 결과에 대해서 기술한다.

2. 농식품 가격변동 요인 인과관계 템플릿 구축

농식품 가격변동 요인에 대해서 템플릿을 구축함으로써 요인에 대한 전체적인 흐름을 볼 수 있다. 이를 통해 개체명 인식 정의에 참고 사항이 될 수 있다. 농식품 가격변동 요인에 대해서 결과는 오름세, 내림세로 구분하고 가격 변동의 요인을 원인, 결과로 나누었다.

<그림2>의 인과관계 템플릿을 시각화한 것으로 '출하 감소로 인해 반입이 감소하여 비축해놓은 재고 물량이 적어져 오름세' 문장처럼 해석된다. 그리고 두 가지 특징이 있다. 1)결과는 다른 결과의 원인이 될 수 있다.

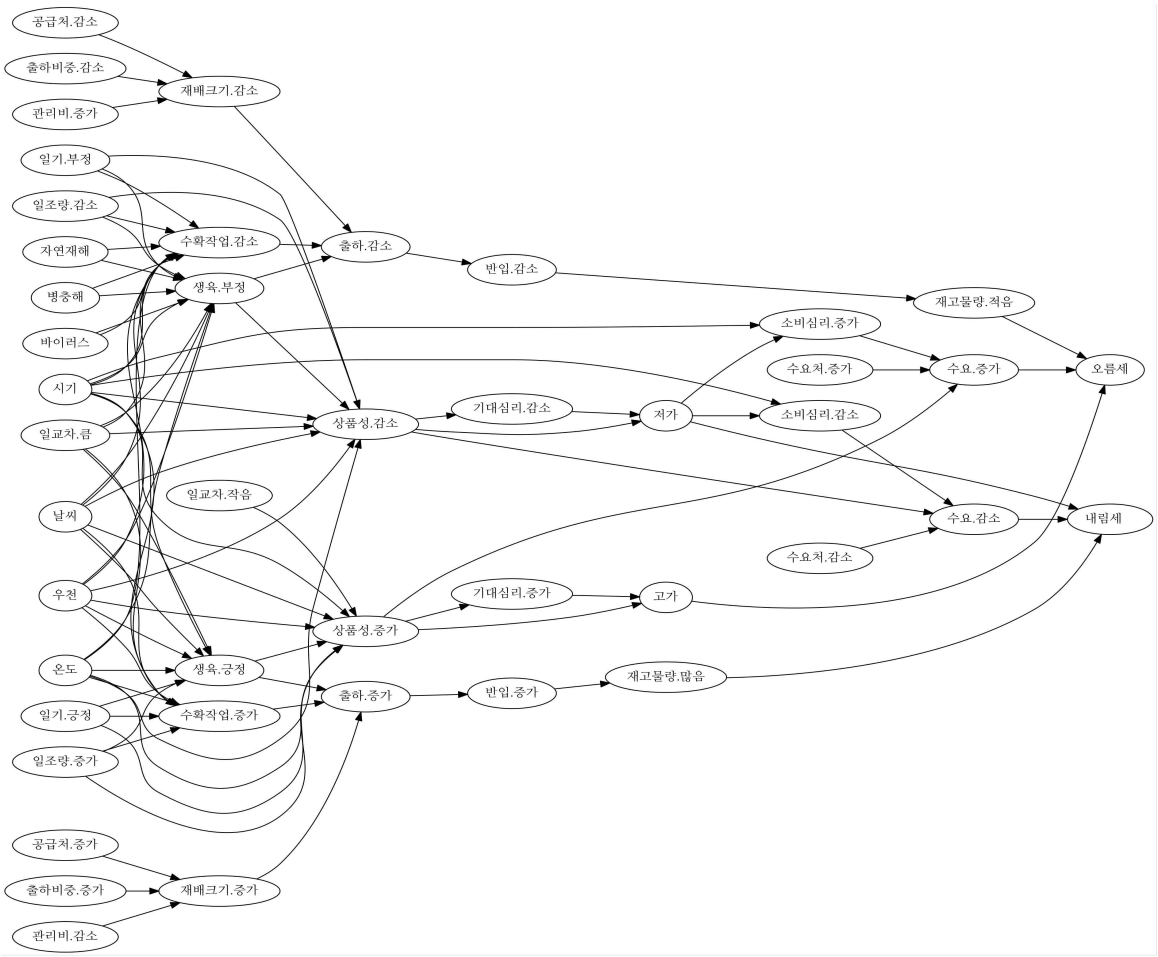

그림 2. 가격 변동에 영향을 미치는 인과관계 템플릿 그래프

예를 들어 반입이 감소하는 것은 재고물량이 감소하는 것의 원인이면서 출하가 감소하는 것의 결과이다. 2) 뛰어넘을 수는 있으나 원인의 순서는 바뀌지 않는다. 예를 들어 수확작업이 감소하면 반입이 감소할 수 있다. 반대로 반입이 감소한다고 수확작업이 감소하지 않는다. 현재 구축된 템플릿에 원인, 결과 쌍은 120쌍이 있다.

3. 농식품 인과관계 분석을 위한 개체명 인식

형태소 분석기는 심층언어 모델 HanBert[1] 기반 형태소 분석기를 사용하였다. 조사에 대한 정보는 요약을 하기 위한 실마리 단어(clue-word) 등을 사용 할 수 있게 한다. 말뭉치로부터 개체명 인식하기 위해 보편적인 방법인 BIO 표기법을 사용하여 형태소 분석 결과에 적용하여 데이터셋을 만든다. BIO 표기법이란 B은 Begin의 약자로 개체명의 시작 부분, I는 Inside의 약자로 개체명 내부 부분을 의미하며, O는 Outside의 약자로 개체명이 아닌 부분을 의미합니다. <표 1>에서 예를 사용해보면 (수요,##량)은 (B-소비,I-소비)로 나타낼 수 있고 (~~로)와 같은 개체명이 아닌 부분은 (O)로 표기한다.

개체명 인식에 많이 쓰이는 딥러닝 모델 양방향 장단기 메모리(Bidirectional Long Short-Term Memory models)와 조건부 무작위장(Conditional random field) 기반 개체명 인식 비교 실험한다.

대분류	중분류(소분류)
공급	반입, 출하, 공급처(사람, 장소), 출하비중, 수확작업, 재배크기(출하지역, 작목전환)
수요	소비, 수요처(사람, 장소), 소비심리
시기	일기, 흐림, 맑음, 다습, 건조, 일교차, 일조량, 우천, 온도(이름, 고온, 저온)
시세	가격, 고가, 저가, 오름세, 보합세, 내림세
상품	이름, 종류, 상품성
기대심리	가격, 상품
자연재해	(중.소분류 없음)
병충해	(중.소분류 없음)
바이러스	(중.소분류 없음)

표 1. 농식품 개체명 태그 정의 일부

Bi-LSTM 모델은 노드가 방향을 가지고 있어 문자들이 순차적으로 등장하는 데이터 처리에 적합한 모델로 RNN(Recurrent Neural Networks)의 일종인 LSTM을 사용한다. RNN은 문자열이 길어지면 성능이 급격하게 저하되어 LSTM을 사용한다. 단방향 LSTM에서는 이전 시점의 단어에 대해서 다음 단어를 예측하는 것인데 추가로 다음 시점의 단어도 참고하여 성능을 높일 수 있는 Bi-LSTM 모델을 사용한다.

BiLSTM-CRF 모델은 Bi-LSTM 모델에 CRF 층을 추가하여 Bi-LSTM의 제약사항[2]을 학습하기 위해 추가했다. 제약사항으로 1) 문장의 첫 번째 단어에서는 I가 나오지 않는다. 2) O-I패턴은 나오지 않는다. 3) B-I-I 패턴에서 개체명은 일관성을 유지한다. Bi-LSTM 에서는 각 레이블을 따로 보고 각 점수가 높은 개체명을 예측했지만 CRF는 여러 레이블을 고려(의존성)하여 개체명을 예측한다.

4. 실험 및 평가

4.1 실험 데이터 구축 및 실험환경

농수산식품기업지원센터(KAMIS)의 “동향/전망 - 일일동향” 카테고리의 문서에 대해 2018년~2019년, 문서 데이터 개수 기반 상위 7개 품목(풋고추, 호박, 배추, 오이, 시금치, 파, 미나리), 각 품목 100개씩 총 700개 문서, 체계적 표본추출 방법으로 데이터 수에 균등하게 추출한다.

기간	2018.1.1.~2019.12.31. (2년)
농산품 품목	데이터 개수 기반 상위 7개 품목 (풋고추, 호박, 배추, 오이, 시금치, 파, 미나리)
데이터	700개(품목당 100개)

표 2. 농수산식품 실험 데이터

작은 크기의 학습데이터에 대한 실험이라 교차검증(5-fold cross validation)으로 학습한다. 데이터를 5등분으로 나누어 훈련 집합(train set, 560개, 80%), 테스트 집합(test set, 140개, 20%)으로 교차검증을 통해 오차율의 평균으로 한다.

각 문서에는 출하지역, 현재 가격변동, 가격변화, 미래 가격변동 예상에 대한 정보 중에서 현재 가격변동이 어떻게 이루어졌는지에 대한 정보가 필요하므로 가격변화, 미래 가격변동 예상에 대한 텍스트는 제외하였다.

학습 환경은 Ubuntu 16.04.2 버전, tensorflow 1.15.3 버전, python 3.6.9 버전에서 교차검증(5-fold cross validation) 학습하여 성능 평가는 F1-score로 한다.

딥러닝 모형의 성능 최적화를 위해서는 알고리즘에 사용되는 최적의 하이퍼파라미터(hyper-parameter) 값을 찾아 설정하는 것이 모형 설계에서 가장 중요하다. 튜닝 과정은 정해진 방법 없이 반복적인 실험을 통해 찾을 수 있다. 가장 좋은 값을 찾을 수는 없지만 데이터와 모형에 따라 적합한 값을 찾을 수 있다.[6] 하이퍼파라미터 입력값에 대해서 반복적으로 실험을 진행하였다.

최종적으로는 BiLSTM-CRF는 Neuron(100), output_dim(100), loss_fuction(crf.loss in keras), activation(Relu), batch_size(50), Adam(0.001)가 가장 좋은 성능을 가졌다.

4.2 딥러닝 모델 비교 실험

성능 평가 방법에서 F1-점수[7]는 정확률(Precision)과 재현율(Recall)을 이용해서 평가하는 방법으로, 보통 문장에는 값에 O(Outside)의 개수가 많은데 정확도(accuracy)는 O가 맞은 상황도 맞았다고 평가하므로 문제가 되어 F1-점수를 통해 비교한다.

딥러닝 모델	F1
Bi-LSTM	0.75
BiLSTM-CRF	0.93

표 3. 개체명 인식 비교 실험 결과

BiLSTM-CRF 모형은 인식률이 93%로 잘 인식하고 있다고 볼 수 있고 새로운 단어에 대해서는 인식되지 않는 부분이 있다.

예) 대구 지역 대형마트, 할인마트들이 오이 할인 행사를 진행하면서 수요대비 물량이 부족하여 오름세에 거래됨

<표 4>는 개체명 인식 결과를 보여준다. 인과관계 부분은 향후 추출 계획이다.

형태소 해석 결과	개체명	인과관계
대구	지역	지역
대형마트	수요처|장소	이벤트
할인마트	수요처|장소	
오이	상품|이름	
할인행사	시기|기간	
수요	수요|소비	원인
물량	재고물량	
부족	상태|적음	
오름세	시세|오름세	결과
거래	수요|소비	

표 4. 개체명 인식 예

5. 결론

가격 변동 요인 분석을 하기 위해 템플릿을 구축하여 요인 정보를 이해하는 데 도움이 될 것으로 기대한다.

딥러닝 기반 개체명 인식을 통해 F1-점수 0.93의 성능으로 개체명을 잘 인식한다.

향후 과제로 농식품 개체명 인식을 기반으로 태그를 추가하고 텍스트 데이터에 대해 인과관계를 요약한다.

감사의 글

이 논문은 2019년도 정보통신산업진흥원의 지원을 받아 수행된 기초연구입니다.

참고문헌

[1] Chang, J. H., Kim, J. W., Kwak, D. E., & Aziz, N. A Correlation Study Between Fruit Wholesale Price And Weather Factor. In Proceedings of the Korea Information Processing Society Conference. Korea Information Processing Society.(pp. 706-708).(2017).

[2] 신성호, 이미경, & 송사광. LSTM 네트워크를 활용한 농산물 가격 예측 모델. 한국콘텐츠학회논문지, 18(11), 416-429. (2018).

[3] 배경태, & 김창재. 인공신경망의 은닉층 최적화를 통한 농산물 가격예측 모델. 한국정보기술학회논문지, 14(12), 161-169.(2016).

[4] 임진모, 김월용, 변우진, & 신승중. 인공지능을 이용한 과일 가격 예측 모델 연구. The Journal of the Convergence on Culture Technology (JCCT), 4(2), 197-204.(2018).

[5] Huang, Z., Xu, W., & Yu, K. Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991.(2015).

[6] “Sequence tagging with LSTM-CRFs”, depends-on-the-definition, last modified Nov 27. 2017, access Feb. 2019, https://www.depends-on-the-definition.com/sequence-tagging-lstm-crf/

[7] Reimers, N., & Gurevych, I. Optimal hyperparameters for deep lstm-networks for sequence labeling tasks. arXiv preprint arXiv:1707.06799.(2017).

[8] Derczynski, L.. Complementarity, F-score, and NLP Evaluation. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) (pp. 261-266). (2016, May)

1) https://twoblockai.com/2020/01/22/hanbert를-공개합니다/

2) https://wikidocs.net/34156