Journal of KIIT. Vol. 20, No. 12, pp. 15-24, Dec. 31, 2022. pISSN 1598-8619, eISSN 2093-7571 
http://dx.doi.org/10.14801/jkiit.2022.20.12.15

# 딥러닝을 활용한 글로벌 주소 데이터 품질 검증과 개선기법

Soeng Saravit*[1], 배진현*[2], 이경희**, 조완섭***

# Global Address Data Quality Verification and Improvement Techniques using Deep Learning

Soeng Saravit*[1], Jin-Hyun Bae*[2], Kyung-Hee Lee** and Wan-Sup Cho***

본 연구는 2022년도 식품의약품안전처의 연구개발비 (21163MFDS517-1)로 수행되었으며 이에 감사드립니다.


## 요 약

본 연구에서는 수입식품 안전관리에 필요한 주소 데이터의 품질 검증과 개선 문제를 다룬다. 수입식품 안전 업무에서는 식품안전 이슈가 발생했을 때 국민의 건강과 안전을 지키기 위해 신속한 안전조치가 필요하며, 주소 정합성 확보가 필요하다. 수입업자가 등록한 해외식품제조업소 주소가 정확한지 검증하고 해당 국가의 행정구역을 식별하며, 표준화된 형식으로 주소를 변환하는 기법을 제안한다. 행정구역 데이터셋을 활용하여 딥러닝 주소 학습 & 분류모형을 생성한 다음 입력한 주소의 행정구역 식별 정확도를 산출하고 표준화된 주소를 반환하여 주소의 품질을 개선한다. 제안하는 모형의 정확도는 90% 이상이며, 중국의 해외제조업소 주소 24,401개 중 1500개를 임의 선택한 다음 검증한 결과 91.4%의 정확도를 보였다.

## Abstract

This study deals with the quality verification and improvement of address data required for imported food safety management. In the imported food safety business prompt safety measures are required to protect the health and safety of the people when food safety issues arise and it is necessary to secure address consistency. We propose a technique for verifying that the address of an overseas food manufacturing business registered by an importer is correct identifying the administrative division of the country and converting the address into a standardized format. After creating a deep learning address learning & classification model using the administrative district dataset, it calculates the administrative district identification accuracy of the entered address and returns the standardized address to improve the quality of the address. The accuracy of the proposed model is over 90%, and as a result of verifying after randomly selecting 1500 addresses out of 24,401 addresses of overseas manufacturing companies in China, the accuracy was 91.4%.



Keywords

address verification, deep learning, address data quality, LSTM, classification model



* 충북대학교 빅데이터학과
- ORCID[1]: https://orcid.org/0000-0001-5704-8820
- ORCID[2]: https://orcid.org/0000-0002-8670-3705
** 충북대학교 경영정보학과 초빙교수
- ORCID: https://orcid.org/0000-0001-5110-123X
*** 충북대학교 경영정보학과 교수(교신저자)
- ORCID: https://orcid.org/0000-0002-4395-1979


· Received: Oct. 13, 2022, Revised: Dec. 08, 2022, Accepted: Dec. 11, 2022
· Corresponding Author: Wan-Sup Cho
Dept. of Management Information System, Korea
Tel.: +82-43-261-2355, Email: wscho@cbnu.ac.kr

## Ⅰ. 서 론

글로벌 무역이나 인터넷 등의 발전으로 국가 간 교류가 활발한 현대사회에서 검증된 주소의 사용은 효율적인 국가 간 물류, 안전한 전자상거래 및 배송, 위치정보서비스 제공 등에서 엄청난 영향력을 발휘한다[1]. 품질이 높은 주소데이터를 사용하면 조직의 업무 처리에 드는 시간과 비용을 절약하고, 고객 만족도를 높이며, 제품 또는 서비스가 고객에게 정확하게 전달되도록 지원하므로 업무 프로세스가 개선된다[2]. 특히, FTA 등으로 국가 간 교류가 활발한 현대사회에서 다양한 국가의 주소를 검증할 필요성은 더욱 높아지고 있다. 그러나 각 국가는 주소체계가 국가의 행정체계에 따른 주소체계DB를 사용하므로 같은 검증기법을 사용할 수 없고, 주소 검증의 목적에 따라 검증알고리즘을 달리 적용해야 한다.

주소 표준화의 이점에 대해 [1]은 세 가지로 설명하고 있다. 첫 번째 경제적인 이점으로 주소 데이터의 상호 운용성을 가능케 하여 주소 데이터의 교환을 쉽게 한다는 점이다. 두 번째 사회적인 이점으로 일부 국가(영국 등)에서 지자체별로 다른 주소체계를 가지고 있으며, 이런 경우 주소 표준화는 사회적인 혼란을 줄이는데 크게 이바지한다. 세 번째 국가 거버넌스적 이점으로 주소가 선거, 인구조사 등의 공공행정 업무의 수행 과정에서 중요한 역할을 한다는 것이다.

국제표준화기구인 ISO/TC211은 2015년 「ISO 19160-1: 개념모델」, 2017년 「ISO 19160-4: 우편주소」, 2020년 「ISO 19160-3: 주소 데이터 품질」에 관한 국제표준을 제정하였다. 또한, 디지털 전환과 IoT 시대를 맞아 사물 주소의 표준화 등으로 영역을 확대해 나가고 있다[11]. 그중 2020년 2월 제정된 표 1의「ISO 19160-3: 주소 데이터 품질」은 주소 정확도 검증의 기준이 될 수 있다[11].

이 표준에서는 완전성(Completeness), 논리적 일관성(Logical consistency), 위치 정확성(Positional accuracy), 시간적 품질(Temporal quality), 주제 정확성(Thematic accuracy), 유용성 요소(Usability element)로 구분하고 있다[3].

본 연구는 수입식품 안전을 확인하고 관리하기 위한 분야에서 해외 식품제조업체들의 주소를 검증하고 관리하는데 유용한 주소검증기법을 제안한다.

표 1. 주소검증의 세부 항목(ISO 19610-3 : 주소데이터 품질)
Table 1. Details of address verification(ISO 19610-3 : address data quality)

| Scope | Definition | Quality details |
|---|---|---|
| Completeness | Measuring errors in relationships between entities, properties, and classes at the address class(address, address component, addressable object, reference object) level | excess, omission |
| Logical consistency | Measures adherence to the logical rules of data structures, properties, and relationships | Conceptual coherence, Domain coherence, Formal coherence, Topological coherence |
| Positional accuracy | Measure the positional accuracy of features within a spatial reference system | Absolute Accuracy, Relative Accuracy, Grid Data Position Accuracy |
| Temporal quality | Temporal relation of Address data and quality of emporal atrributes | Accuracy of time measurement, Temporal consistency, Temporal validity |
| Thematic accuracy | The accuracy of these relationships, including the accuracy of quantitative attributes, and the accuracy of non-quantitative attributes and address data classification | Classification Accuracy, Non-Quantitative Attribute Accuracy, Quantitative Attribute Accuracy |
| Usability element | Measure the suitability of an address dataset for a specific application or evaluate quality based on user requirements that cannot be addressed by other data quality factors | |

2021년 6월 현재 기준 대한민국으로 수출하는 해외 식품제조업체의 수가 190여 개 국가, 9만여 개이며, 국가별 제조업소들의 주소를 검증해야 업무의 효율적인 처리가 가능하다. 특히, 국가마다 행정구역 체계와 주소체계가 다르므로 주어진 주소로부터 해당 국가의 세부 행정구역을 인식하는 데 문제가 생길 수 있다.

예를 들어, 행정구역 체계가 4단계 이상으로 복잡하게 구성되는 국가도 있고, 대만이나 싱가포르와 같이 영토의 규모가 작아서 2단계 이내로 구성되는 예도 있다.

이러한 이유로 주소를 검증하고 행정구역 수준을 식별하는 모듈을 국가별로(혹은 유사한 국가들의 그룹별로) 다르게 개발하여 정확도를 높일 필요가 있다. 본 연구에서는 국가별 행정구역 데이터셋을 만들어 모형 개발할 때 학습용 데이터셋으로 활용한다.

주소를 활용하는 분야에 따라 주소 데이터 품질에 세부 항목에서 중요도가 달라질 수 있다. 예를 들어, 식품안전 분야에서는 해외 식품제조업체의 주소 완전성(Completeness)도 중요하지만, 더욱 시급한 것은 식품 위해가 발생했을 때 그 제조업소 주소의 행정구역 단위를 확인하여 통관검사 강화 또는 검사방법 보완 등 알맞은 수입식품 안전 관련 행정업무를 신속히 적용할 필요가 있다.

본 논문에서는 주소 그 자체의 정확성보다 주어진 주소를 행정구역 레벨로 분류하는 방법(행정구역 레벨분류 기법)을 제안한다. 물론 이 과정에서 행정구역 데이터셋을 활용하여 주어진 주소를 표준화하고, 품질을 높이게 된다.

제안된 기법은 구글 지오코딩(Google geocoding) 서비스와 딥러닝 모델을 연계하여 높은 정확도로 주소를 행정구역 레벨로 분류하고, 품질을 개선한다. 지오코딩 서비스는 주어진 주소 문자열에 대하여 행정구역 수준별로 (국가, 광역시도, 시군구 등) 주소구성 요소들을 분리하고, 해당 주소에 대한 위도 경도 정보를 제공한다.

그러나 지오코딩에서 받은 주소구성 요소들이 실제 해당 국가의 행정구역DB와는 일치하지 않을 수도 있으며, 주소 문자열에 오타나 약어가 포함된 경우(주소를 영어로 번역하거나 혹은 지명에 대한 약어를 사용하는 경우) 지오코딩 처리가 불가할 수도 있다. 본 연구에서는 이와 같은 지오코딩의 한계를 극복하기 위해 수작업으로 각국의 '행정구역 데이터셋'을 수작업으로 구축하고, 이를 학습데이터로 활용하여 딥러닝 모델을 개발한다. 행정구역 데이터셋은(국가, level1, level1_name, level2, level2_name, level3, level3_name)와 같은 구조로 구성되며, 3장에서 상세한 데이터 구조를 소개한다.

생성된 딥러닝 분류 모델은 주소 문자열에 오타나 약어 등이 포함되어 있는 경우에도 입력한 주소에 해당하는 행정구역 수준을 정확하게 식별하는 등 기존 지오코딩의 한계를 극복하는데 유용하다. 제안된 기법의 정확도를 평가하기 위해 중국, 일본, 캄보디아를 대상으로 각각 모형을 개발하고, 평가하였다. 중국과 일본의 경우 90% 이상의 정확도를 보여주고 있으며, 캄보디아의 경우 78% 정도의 정확도를 보여주고 있다.

논문의 구조는 다음과 같다. 제 2장에서는 주소검증에 관한 관련연구를 소개한다. 제 3장에서는 주소검증기법을 소개한다. 제4장에서는 연구결과를 실제 중국회사 주소 데이터셋에 적용한 사례를 소개하고, 현업에서 사용할 수 있도록 개발한 주소검증 시스템을 소개한다.

## II. 관련 연구

현대사회에서 주소는 거주지 개념을 넘어 물류, 우편, 전자상거래, 위치기반산업 등 산업 전반과 연결되는 기본 요소로서의 역할과 그 활용 범위가 점차 확대됨에 따라 국제 사회는 산업 전반에 걸친 유통체계 비용 절감 등을 위해 주소를 국제표준으로 제정하고 있다. 최근 들어 표준 제정의 범위를 주소의 품질·교환 및 지도 등으로 확대 중에 있으며, IoT 시대를 맞아 사물주소표준화 등으로 확장해 나가고 있다.

그림 1은 주소검증의 사례를 보여주는 그림이다. 일반적인 주소검증은 주소 문자열을 입력받아 파싱(Parse), 형식화(Format), 표준화(Standardize), 지오코딩(Geocoding), 검증(Verify) 과정을 거쳐 올바른 주소 문자열로 변환해서 출력한다(그림 1).

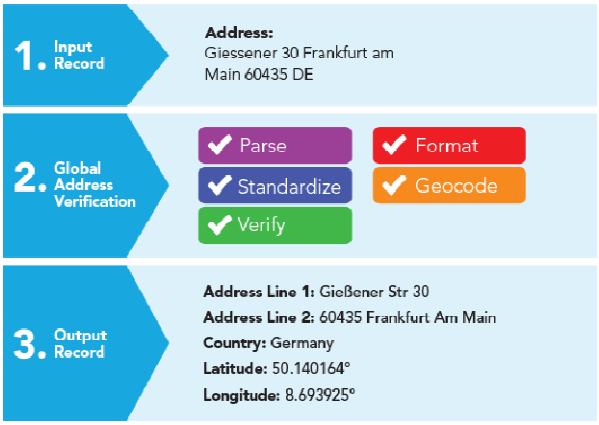


그림 1. 주소검증 서비스 예시
Fig. 1. Example of address verification service[4]

[5]에서는 호주 국가 우편 주소 지침과 호주 주소 데이터베이스(G-NAF, Geocoded National Address File)를 기반으로 세부 주소 태그를 구축하고, 확률적 은닉 마르코프 모델(HMM)을 사용하여 주소 정리 및 표준화를 위한 자동화된 접근 방식을 제시했다(그림 2).


그림 2. 세분화된 호주 주소 예시
Fig. 2. Example of granular australian address

[6]의 연구에서는 비표준화된 주소 문제를 해결하고 모든 개체명 인식 NER(Named Entity Recognition) 문제에 적용할 수 있는 우편 주소 구문분석을 위해 다양한 딥러닝 기법을 사용하여 다국적 주소를 구문 분석하기 위한 최첨단 라이브러리인 DeepParse를 구축했다.

구축된 라이브러리는 비표준화된 주소 데이터에 대해 향상된 일반화 기법, 클래스 혼합 문제와 개체명 인식 문제의 해결 방안을 제시하였다.

[7] 연구는 지오태깅이 되지 않은 텍스트 데이터에서 행정구역이나 기관명, 도서관, 영화관 등이 들어가는 확장된 개념의 장소 정보를 탐지하는 방법을 제안하였다. 뉴스, 기사, 블로그, 소셜미디어 등에서 추출되는 비정형 텍스트 데이터를 가지고 라벨링, 단어 임베딩, 어텐션 기반의 딥러닝 모델을 사용해서 이진 분류기를 만들고 장소 정보의 포함 여부를 예측하였다.

[8]의 연구에서는 딥러닝 기법을 활용하는 주소 정합성 검증 프로그램을 개발하였으며, 검증 프로세스는 그림 3과 같다. 먼저 국가별 행정구역 데이터의 행정구역명으로 행정구역 레벨을 예측하는 RNN-LSTM(Recurrent Neural Network-Long Short Term Memory) 모델을 생성한다.

이후 실제 주소검증 단계에서는 입력된 해외 제조업소 주소 데이터를 구글 지오코딩 하여 행정구역을 추출하고, 생성된 RNN-LSTM 모델에 적용하여 예측 행정구역의 정확도 및 표준 주소를 도출한다. 기존 연구에서는 RNN-LSTM 모델에 다중 클래스 분류(Multi-Class Classification)를 구현하여 추출된 각 주소구성 요소에 대하여 총 4개의 클래스('Country', 'Level 1', 'Level 2', 'Level 3') 중에 하나로 분류하였다.

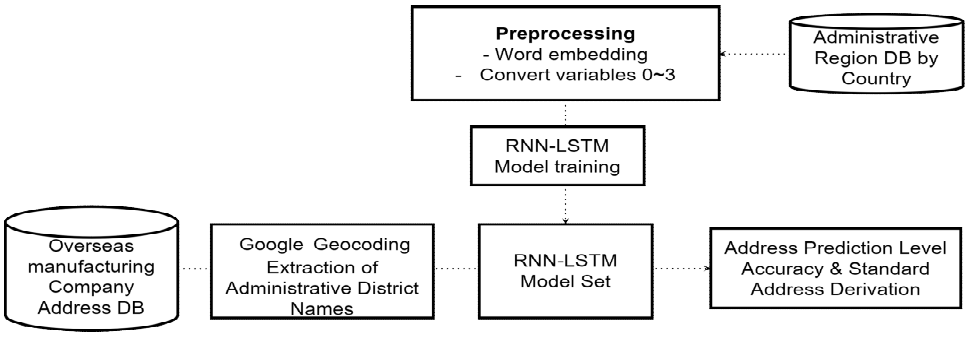


그림 3. 딥러닝 기법을 이용한 주소검증
Fig. 3. Address verification using deep learning technique

한편, 주소 국제 표준화는 2009년부터 ISO TC 211 19160(지리정보분과) Working Group7 (WG7)에서 진행하고 있으며, 개념모델(ISO 19160-1), 주소부여(ISO 19160-2), 품질관리(ISO 19160-3), 국제우편(ISO 19160-4), 지도표기(ISO 19160-5) 등으로 나누어 표준화를 추진하고 있다(그림 4)[9].

## III. 주소검증기법

본 장에서는 주소 문자열을 주소구성 요소들로 분리한 후, 각 구성요소를 적절한 행정구역 레벨로 분류하는 딥러닝 기반 분류모델 구축 방법을 소개하고, 정확도를 검증한다.

### 3.1 딥러닝 기반 국가별 주소 분류모델 생성

#### 3.1.1행정구역 데이터셋 구축

주소를 행정구역 레벨로 분류하기 위해 먼저 각국의 행정구역 데이터셋을 구축하여 학습데이터로 사용한다. 각국의 행정구역 데이터셋은 각국의 주소체계 DB와 Post Code Query 웹사이트 데이터 등 다양한 소스를 활용하여 수작업으로 만들어진다. 표 2는 예시로 만든 한국에 대한 행정구역 데이터셋이다. 행정구역 데이터셋을 전처리하여 주소에 포함된 행정구역 명칭을 알맞은 행정구역 레벨로 분류하기 위한 학습 데이터를 구성한다.

표 2. 국가별 행정구역 주소체계 DB(한국사례)
Table 2. Administrative district address system by country DB(Korea case)

| country | Korea | … | Korea |
|---|---|---|---|
| level_1 | Chungcheongbuk | … | Gyeonggi |
| v1_division | Do | … | Do |
| level_2 | Cheongju | … | Anseong |
| lv2_division | Shi | … | Shi |
| level_3 | Seowon | … | Bogae |
| lv3_division | Gu | … | Myeon |

#### 3.1.2 딥러닝 기반 검증 및 분류모델 구축

그림 5는 딥러닝 기반 주소 검증 및 분류 모델 아키텍처를 보여준다[10]. 분류모델은 학습데이터를 기반으로 RNN(Recurrent Neural Network) 기법을 이용하여 학습한다. 분류모델은 단어 임베딩, LSTM 레이어 및 Dense레이어를 적용한 텐서플로(TensorFlow) 프레임워크로 구성한다.

임베딩 레이어에서는 텍스트 벡터화 기법을 사용하여 주소 요소들을 딥러닝 모델에 사용하기 전에 벡터화한다[11]. 텍스트 벡터화는 단어 수준과 문자 수준으로 구분할 수 있는데 본 연구에서는 문자 수준을 적용하였다. 그 이유는 철자오류 등이 포함된 주소 요소도 유사성 정도에 따라 적절한 행정구역 수준으로 분류할 수 있게 하기 위함이다.

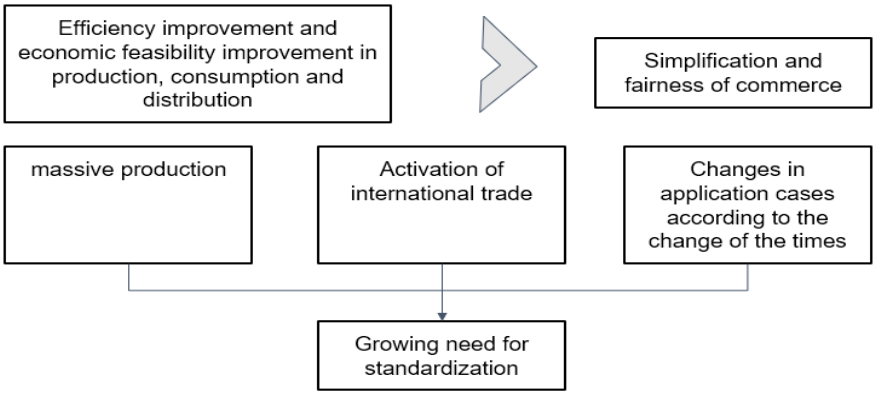


그림 4. 주소 표준화의 필요성
Fig. 4. Needs for address standardization

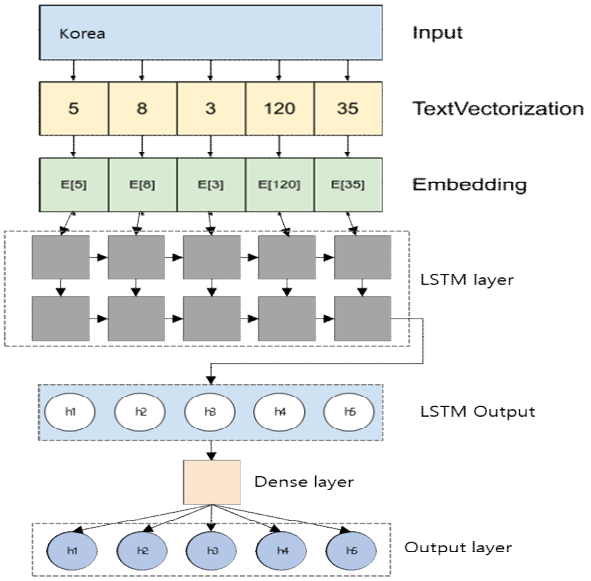


그림 5. 딥러닝 기반 주소 검증 및 분류 모델 아키텍처
Fig. 5. Deep learning based address verification & classification model architecture

또한, 약어나 코드 등이 사용되는 경우에도 학습 데이터에 포함시켜 학습모델을 생성하는 방식으로 정확도를 높일 수 있다.

LSTM레이어는 시퀀스 예측 또는 분류 문제에서 순서 의존성을 학습할 수 있는 특수한 유형의 순환 신경망이다[12]. LSTM레이어는 2개 층을 적용하였으며, 각 LSTM 레이어마다 128개의 은닉 유닛을 정의하였다.

Dense 레이어는 Softmax 활성화함수를 적용하여 분류 결과를 국가 수준, 레벨 1(Region level 1), 레벨 2(Region level 2), 레벨 3(Region level 3)과 같은 4가지 다른 클래스(행정구역 수준)로 분류한다.

## 3.2 주소분류 모델

그림 6은 딥러닝 분류모형을 사용하여 사용자가 입력한 주소를 검증하고 분류하는 과정을 도식화한 것이다. 사용자가 입력한 주소는 지오코딩을 이용하여 구성요소들로 파싱(Parsing)한다(국가수준, 수준1-광역시도, 수준2-시군구, 수준3 - 구읍면동). 파싱된 구성요소들은 딥러닝 기반 분류 모델을 사용하여 4개의 수준으로 분류하며, 입력한 주소에 포함된 철자오류나 약어를 사용한 경우에도 알맞은 행정구역을 매칭하고 주소의 정확도를 반환한다.

### 3.2.1 주소 파싱(Address parsing)

주소 파싱은 입력된 주소를 구글 지오코딩 서비스를 이용하여 주소구성요소들로 파싱하는 과정이다[13]. 주소의 각 구성요소는 국가, 지방, 도시, 지구 또는 거리 세부 정보 등 다수 레벨을 갖는 행정주소체계(행정구역레벨)로 나눌 수 있다. 그림 5의 하단 좌측에서 입력된 주소 문자열에 대하여 지오코딩 결과로 생성된 주소요소들을 보여주고 있다.

### 3.2.2 분류모델 (Classification model)

주소 문자열에 대하여 지오코딩 결과로 생성된 각 주소구성요소를 딥러닝 기반 분류모델에 입력하여 각 구성요소별로 해당하는 행정구역 수준을 생성한다. 이 때 분류모델은 주소를 외국어로 번역할 때 발생하는 발음차이로 인한 표기법의 상이함이나 단순오타 등을 감안하여 행정구역 수준을 분류하고, 분류의 정확도 점수를 반환한다.

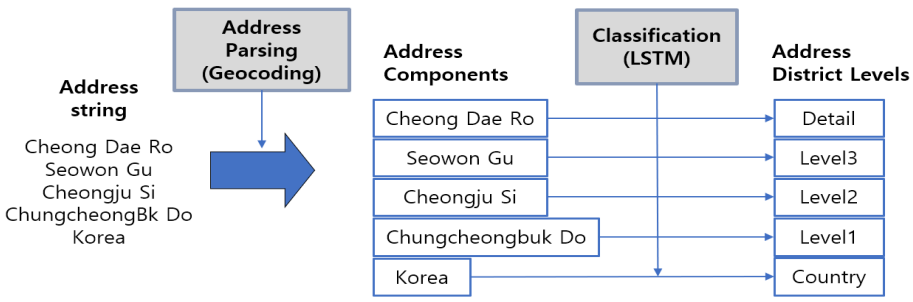


그림 6. 주소검증 프로세스(예시)
Fig. 6. Address verification process(example)

주소구성요소의 지명이 학습된 행정수준의 지명과 정확하게 매칭된다면 행정구역 수준별 부분정확도 최대치를 반환하고, 정확하게 일치하지 않는다면 지명의 불일치정도에 따라 낮은 부분정확도를 반환한다. 또한 학습데이터에 동의어나 약어 등을 포함시킨 경우에는 이를 감안한 분류가 가능하게 된다. 분류모델은 기본적으로 각 국가별로 생성하며, 주소체계가 유사한 국가들을 하나의 그룹으로 묶어서 모델을 생성할 수도 있다.

#### 3.2.3 모델의 평가

본 연구에서는 제안된 행정구역 레벨분류 기법의 정확도를 평가한다. 먼저, 중국에 대한 행정구역 데이터셋을 8:2로 분할하여 훈련하고, 생성된 분류모형을 검증한 결과는 표 3과 표 4에 나타냈다. 각 레벨의 분류정확도는 모두 90% 이상으로 나타났다.

표 3. 혼돈행렬
Table 3. Confusion matrix

| Actual predicted | Country | Level_1 | Level_2 | Level_3 |
|---|---|---|---|---|
| Country | 1788 | 0 | 0 | 0 |
| Level_1 | 0 | 1802 | 17 | 0 |
| Level_2 | 0 | 18 | 1740 | 7 |
| Level_3 | 0 | 4 | 117 | 1582 |

표 4. 모델성능예측
Table 4. Model performance prediction

| Class | Precision | Recall | F1-Score |
|---|---|---|---|
| Country | 100.00% | 100.00% | 100.00% |
| Level_1 | 98.79% | 99.07% | 98.93% |
| Level_2 | 92.85% | 98.58% | 95.63% |
| Level_3 | 99.56% | 92.89% | 96.11% |

다음으로 식품의약품안전처의 수입식품정보 포털사이트에서 수집한 3개국(캄보디아, 중국, 일본)의 다양한 실제 사업장 주소를 이용하여 주소를 검증한다. 각 국가의 주소체계가 상이함을 감안하여 국가별로 별도의 분류모델을 구축하여 사용한다.

표 5는 검증에 사용된 2021년 6월 기준 식약처에 등록된 각 국가의 식품제조업소 주소 개수이며, 표 6은 국가별로 분류모델이 분류한 행정구역 분류 정확도를 보여준다. 중국과 일본의 경우 각각 24,401개의 업체와 5,968개의 실제 해외업체 주소가 검증에 사용되었다.

분류 결과 캄보디아를 제외하고는 90% 정도의 정확도로 행정구역 레벨을 분류하였다. 캄보디아의 경우 도시지역은 비교적 정확하게 분류하지만 시골지역은 정확도가 낮게 나타났으며, 이는 수입업자가 등록한 주소에 오류가 포함되었기 때문으로 판단된다.

표 5. 국가별 제조업소 주소 개수
Table 5. Number of manufacturing address by country

| Country | Number of addresses |
|---|---|
| Cambodia | 76 |
| China | 24,401 |
| Japan | 5,968 |

표 6. 국가별 제조업소 주소 평균 정확도
Table 6. Average accuracy of manufacturer addresses by country

| Country | Average accuracy |
|---|---|
| Cambodia | 77.81% |
| China | 90.18% |
| Japan | 89.44% |

## Ⅳ. 분류모형을 이용한 주소검증 시스템 구축과 활용

본 장에서는 제안된 주소검증 기법을 실제 해외 식품제조업소 주소에 적용하여 정확성을 평가하고, 현업에서 사용할 수 있는 주소검증시스템으로 구축한 결과를 소개한다.

### 4.1 웹 기반 주소 검증 시스템

웹 기반 검증 시스템은 파이썬 기반 웹 프레임워크인 플라스크(Flask)를 사용하여 주소검증이 필요한 사용자가 웹에서 편리하게 주소를 검증할 수 있도록 구현하였다. 웹 기반 시스템은 사용자가 특정 국가를 선택하고 주소를 입력하면 시스템에서 행정구역별로 분류하고 정확도를 검증[13]할 수 있는 간단한 인터페이스로 구성되어 있다. 그림 7부터 그림 9은 주소검증시스템의 화면이다.

그림 7에서와 같이 사용자가 국가를 선택하고 제조업소의 주소를 입력한 뒤에 확인 버튼을 누르면 해당 국가 주소에 대해 학습한 딥러닝 기반 모델을 이용해 주소 구성요소의 분류하여 각 구성요소의 정확도와 평균 정확도를 반환한다.

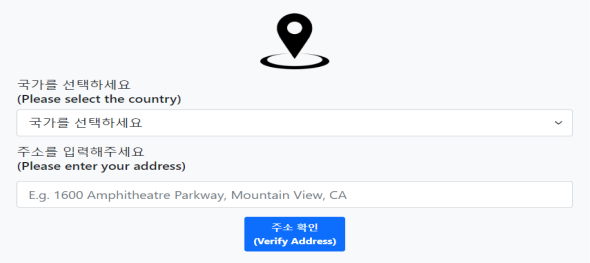


그림 7. 웹 기반 검증시스템 사용자인터페이스
Fig. 7. User interface of web-based verification system

그림 8과 그림 9는 중국의 주소를 입력하여 행정구역 레벨을 확인하고 입력한 주소를 지도에서 확인한 사례이다. 그림 8은 입력한 주소가 중국의 행정구역체계에서 어느 지역에 해당하는가를 매칭한 후 매칭확률까지 제시하며 표준화된 주소로 변환하여 반환한다. 그림 9는 구글 맵(Google Map) 서비스를 이용하여 입력한 주소의 위치를 지도에 표시한다.

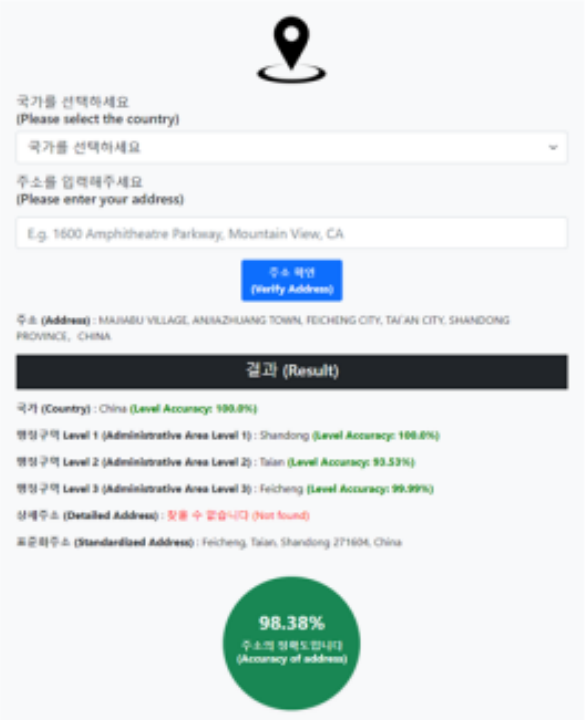


그림 8. 웹 기반 검증시스템 결과
Fig. 8. Results of web-based verification system

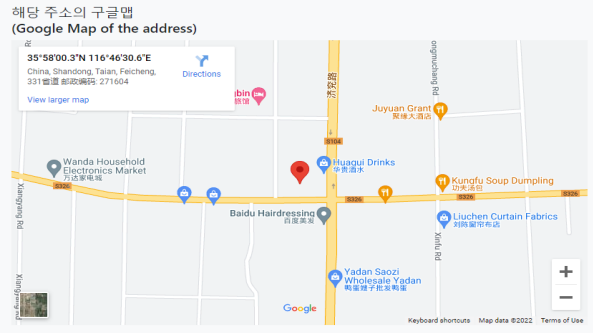


그림 9. 구글맵 서비스로 제공한 제조업소 주소 지도
Fig. 9. Manufacturers address map provided by google maps service

## Ⅴ. 결 론

본 연구에서는 해외식품제조업소의 주소를 주소 구성요소들로 구분하고, 각 구성요소에 대하여 적절한 수준의 행정구역으로 분류하는 딥러닝 기반의 분류모델을 만들고 정확도를 평가하였다. 우리나라의 수입식품 수요는 꾸준히 증가[14]하고 있으며 전 세계적인 지구온난화, 기후변화 가속화에 따라 신·변종 세균·바이러스, 내성균, 곰팡이독소 등 식품 중 위해요소가 증가하고 있다[15].

이로 인하여 식품안전의 관리영역이 확대되고 있으며, 비즈니스 영역에서는 주소 자체의 정확도가 중요하겠지만 공공행정분야에서는 주소 문자열을 이용해서 해당 국가의 행정구역을 명확하게 식별해 내는 것이 행정효율 향상에 필요하다. 분류모델을 학습시키기 위해 각 국가별 행정구역 주소체계 DB를 구축하여 사용하였으며, 모델은 기본적으로 각 국가별로 생성된다.

실제 사용시 주소가 입력되면 구글 지오코딩 서비스를 이용하여 입력한 주소의 구성요소를 파싱하고, 각 구성요소에 대하여 딥러닝 분류모델에서 국가별 행정구역 수준을 분류하고 분류정확도를 산출한다. 테스트 데이터셋에 대하여 예측한 결과 90% 이상의 정확도를 보였다. 특히, 구글 지오코딩의 한계점으로 지적되는 주소를 영어로 번역할 때 발음차이 등으로 인한 오류나 동의어 및 약어 등의 사용으로 인한 문제점을 해결할 수 있도록 하였다. 본 연구의 목적은 조직이나 기업이 상용 서비스에 의존하지 않고 자체적으로 주소를 검증할 수 있는 시

스템을 구축할 수 있도록 딥러닝 기반의 기술을 제공하는 것이다.

## References


[1] S. H. Kim, et al., "The trend of international address standardization and implications - with a focus on ISO 19160-2 -", Journal of Cadastre & Land InformatiX(JCLI), Vol. 52, No. 1, pp. 57-68, 2022. http://doi.org/10.22640/lxsiri.2022.52.1.57.

[2] J. Tetler, "Why address verification is essential for your business", https://www.edq.com/blog/why-address-verification-is-essential-for-your-business [accessed: Jun. 24, 2021]

[3] M. Yassine, D. Beauchemin, F. Laviolette, and L. Lamontagne, "Leveraging Subword Embeddings for Multinational Address Parsing", in 2020 6th IEEE Congress on Information Science and Technology (CiSt), Vol. 5, No. 3, pp. 353-360, Jun. 2020, http://doi.org/10.1109/CiSt49399.2021.9357170.

[4] Mellisa, https://www.melissa.com/address-verification [accessed: Jun. 21, 2022]

[5] C. Peter and D. Belacic, "Automated probabilistic address standardisation and verification", Australasian Data Mining Conference, 2005.

[6] N. Abid, A. ul Hasan, and F. Shafait, "DeepParse: A Trainable Postal Address Parser", 2018 Digital Image Computing: Techniques and Applications (DICTA), Canberra, ACT, Australia, pp. 1-8, Dec. 2018. http://doi.org/10.1109/ DICTA.2018.8615844.

[7] K. H. Min, et al., "A Method for Detecting Location Information using Attention-based Deep Learning Model and Word Embedding", Journal of Korean Society for Geospatial Information Science, Vol. 27, No. 5, pp. 33-39, Sep. 2019. http://dx.doi.org/10.7319/kogsis.2019.27.5.033.

[8] S. Saravit, et al., "Deep-learning based global address data quality improvement", Korean patent application, Oct. 2022.

[9] Jin Kim, et al., 2015 International Standardization Countermeasures for Locating Means, Ministry of Government Administration and Home Affairs/Korea Land Information Corporation Geospatial Data Research Report, Dec. 2015.

[10] TensorFlow, "Text classification with an RNN", https://www.tensorflow.org/text/tutorials/text_classification_rnn [accessed: Jun. 21, 2022].

[11] J. Brownlee, "How to Prepare Text Data for Deep Learning with Keras", Machine Learning Mastery, https://machinelearningmastery.com/prepare-text-data-deep-learning-keras [accessed: Jun. 20, 2022]

[12] J. Brownlee, "A Gentle Introduction to Long Short-Term Memory", Machine Learning Mastery, https://machinelearningmastery.com/gentle-introduction-long-short-term-memory-networks-experts [accessed: Jun. 21, 2022].

[13] S. Saravit, et al., "Overseas Address Data Quality Verification System using Deep Learning Technique", Patent pending, Sep. 2022.

[14] Imported Food Status, http://index.go.kr/potal/stts/idxMain/selectPoSttsIdxMainPrint.do?idx_cd=3055&board_cd=INDX_001 [accessed: Jun. 25, 2021]

[15] The 5th Food Safety Management Basic Plan, Ministry of Food and Drug Safety, Feb. 2021.


## 저자소개


Soeng Saravit

2017년 5월 : B. S. in Department of Computer Science, Royal University of Phnom Penh
2022년 2월 : 충북대학교 빅데이터학과(공학석사)
2022년 3월 ~ 현재 : 충북대학교 빅데이터학과 박사과정
관심분야 : 빅데이터, 딥러닝, IoT, 데이터과학, NLP

배 진 현 (Jin-Hyun Bae)

2020년 8월 : 충남대학교
감성인지소프트웨어(공학사)
2021년 ~ 현재 : 충북대학교
빅데이터협동과정 석사과정
관심분야 : 빅데이터, 인공지능

이 경 희 (Kyung-Hee Lee)

1997년 2월 : 충북대학교
전자계산학과(이학석사)
2004년 2월 : 충북대학교
전자계산학과(이학박사)
2016년 4월 ~ 2020년 2월 :
충북대학교 대학원 빅데이터학과
초빙교수
2020년 4월 ~ 2021년 12월 : ㈜힐링소프트 이사
2022년 1월 ~ 현재 : 충북대학교 경영정보학과 초빙교수
관심분야 : 빅데이터분석, 알고리즘, 데이터연계

조 완 섭 (Wan-Sup Cho)

1987년 2월 : KAIST
전산학과(박사)
1996년 ~ 현재 : 충북대학교
경영정보학과 교수
관심분야 : 빅데이터, 비즈니스
인텔리전스, 인공지능, 블록체인,
빅데이터거버넌스