JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 22, No. 3, pp. 501-508, Mar. 2021

Check for updates

수입물품의 HS 코드 자동 분류를 위한 자연어처리 기반의 딥러닝 모델 개발

이 종 권[1] · 최 근 호[2] · 김 건 우[3*]
[1](주)데이터월드 운영사업본부 이사
[2]한밭대학교 융합경영학과 조교수
[3*]한밭대학교 융합경영학과 부교수

Development of a Natural Language Processing based Deep Learning Model for Automated HS Code Classification of the Imported Goods

Jong-Kwon Lee[1] · Keunho Choi[2] · Gunwoo Kim[3*]

[1]Director, Headquarter of Operation Business, Dataworld Co. Ltd., Daejeon 35240, Korea
[2]Assistant Professor, Department of Business Administration, Hanbat National University, Daejeon 34158, Korea
[3*]Associate Professor, Department of Business Administration, Hanbat National University, Daejeon 34158, Korea

[요 약]

관세법에는 수입물품에 대해 물주가 직접 품목분류를 하고 신고한 HS코드의 세율에 따라 관세를 납부하게 되어 있다. 하지만 급격한 산업환경의 변화와 무역팽창, 융복합 신상품의 출현 등으로 인해 품목분류에 대한 물주의 지식이 부족해졌고 이에 따른 오류신고로 국내외에서 많은 마찰이 발생하고 있다. 이에 본 연구는 자동으로 HS코드를 분류할 수 있는 자연어처리 기반의 딥러닝 모델을 구축하였다. 본 연구에서 제안하는 모델은 수입물품의 품명 정보만을 바탕으로 워드 임베딩과 딥러닝 기법을 통해 수입물품의 HS코드를 물주에게 추천해줌으로써, 손쉬운 품목분류 가능하게 하여 물주의 부대 경비 감소 및 정확한 수입신고를 통한 국가세수 재정의 안정적 확보에 큰 도움을 줄 것으로 기대된다.

[Abstract]

The Korean Customs Law requires the owner to directly classify items on imported goods and pay customs duties at the tax rate of the declared HS code. However, the rapid changes in the industrial environment, the expansion of trade, and the emergence of new convergence products have caused the lack of the owners' knowledge on the classification of the imported goods. And this leads to the trade friction in both domestic and abroad. Therefore, this study aims to establish a deep learning model based on natural language processing that can classify HS codes automatically. The proposed model in this study recommends the HS code of imported goods using word embedding and deep learning techniques based soley on an item name, which is expected to help the owner's cost reduction and accurate import declaration, thus helping to secure the national tax finance.

색인어 : 텍스트 마이닝, 문서분류, HS 코드, 딥러닝, 기계학습
Key word : Text Mining, Documents Classification, HS Code, Deep Learning, Machine Learning

http://dx.doi.org/10.9728/dcs.2021.22.3.501

Received 16 December 2020; **Revised** 27 January 2021
Accepted 27 January 2021

***Corresponding Author; Gunwoo Kim**

Tel: +82-42-821-1290
E-mail: gkim@hanbat.ac.kr

Ⅰ. 서 론

관세청 홈페이지에 따르면 2018년 우리나라의 수출입 물동량이 약 3,000만 건, 1조 1,400억 달러(1,272조)을 달성하였는데, 2008년(1,108만 건, 8,572억 달러) 대비 물동량 건수는 3배, 금액으로는 33% 급상승하였다. 특히, 수입 건수는 4배 가량 급증하였는데 이는 인터넷, 모바일 등 컴퓨팅 기술의 발달에 따른 온라인 시장이 전 세계로 넓어지면서 개인의 수입 물품 거래가 많아졌고 기업은 이윤 극대화를 위해 글로벌 생산기지를 확대함에 따라 국가 간 원자재 및 부품 수급을 위한 교역량이 증가한 것으로 보인다. 또한, 2018년도 기준 수입업자가 물품을 수입하면서 과세관청에 납부한 징수금액은 약 62.9조 원으로 수입금액(597.1조 원)의 약 10.5%를 차지하고 있다.

10여 년 전부터 우리나라는 자국 시장 보호와 국가경쟁력을 높이기 위해 세계 여러 나라와 FTA를 체결하고 있으며, 세계 무역흐름도 FTA를 중심으로 경제 블록화가 가속화되고 있다. 이러한 WTO 체제에서 자국 시장을 보호하기 위한 수단으로 활용되는 것이 관세이다. 국가 간 물품교역을 하는 데 있어 물품에 대한 관세는 국가가 정한 관세율표에 따라 부과하는데 이때 국제적으로 통일시킨 HS(Harmonized Commodity Description and Coding System) 협약을 적용한다.

[1]은 HS코드 분류가 국제무역에서 수입 물품의 세율 결정, 원산지 증명, 무역 수송, 보험 등에서 핵심요소로 사용 중이라고 분석했다. 수입 물품 통관에 관한 절차는 수입 물품에 대해 물주가 물품의 과세표준 및 납부세액을 스스로 결정하여 관세청에 신고하게 된다. 따라서 수입 물품의 관세율표상 세율로 세금이 납부되게 되므로 물주는 물품의 지식과 품목분류 체계에 대한 상당한 수준의 지식과 정보를 필요로 하는데 이는 명확한 품목분류 신고의 장애 요인이 된다.

[2]는 품목분류 결과에 따라 수입의 납부세액과 수출의 환급액이 달라지기 때문에 물주와 과세관청 간 이견 마찰의 원인이 된다고 분석했다.

품목분류에 대한 선행 연구를 보면, 품목분류제도의 이론적 고찰과 주요국의 품목분류제도의 운용현황, 그리고 품목분류 분쟁사례와 판례 분석을 통해 시사점을 도출하고 법적 제도적 측면의 개선방안을 제안하거나, 또는 품목분류가 모호한 특정 부분품이나 완성품의 분류사례 및 분류기준을 제안하는 연구가 주를 이룬다[3]. 앞선 선행 연구들은 품목분류의 분쟁사례와 이슈사항에 대해 정책적, 제도적, 절차적 개선방안을 제안하는 것으로 의의가 매우 높다.

하지만 최근 인터넷, 컴퓨팅 기술이 발전함에 따라 끊임없이 발생하는 방대한 데이터를 기반으로 인공지능 등 기계학습을 통한 문서분류를 시도하는 성공사례가 여러 산업군에서 나오고 있으나, HS품목을 자동으로 분류하는 사례는 그 필요성에도 불구하고 아직까지 찾아볼 수 없는 실정이다. 또한 혼자 생활하는 1인 가정의 증가, 저성장·저물가 추세에 따른 소비 패턴의 변화, 그리고 인터넷 기술의 발전과 4차 산업혁명으로 이어지는 산업 환경 속에서 기업들은 생존을 위해 글로벌 제품출시와 대외무역을 가속화하고 있다.

표 1. 최근 5년 수출입 물동량과 관세 징수 현황
Table 1. The present of import and export in recent 5 years

	2014	2015	2016	2017	2018
Number of import	16,302,570	14,459,780	15,517,151	18,691,086	21,559,704
Amount of import (US dollar)	525,514,506	436,498,973	406,192,887	478,478,296	535,202,428
Number of export	7,083,325	7,438,562	8,273,729	8,423,120	8,950,209
Amout of export (US dollar)	572,664,607	526,756,503	495,425,940	573,694,421	604,859,657
Tariff (100 million won)	581,467	522,548	495,283	573,534	629,495

이에 따라 앞으로 생산·유통되는 제품이 점점 다양해지고, 융·복합 상품의 출현으로 품목분류는 더욱 어려워지며 무역 분쟁도 증가할 것으로 예상된다.

따라서 본 연구의 목적은 품목분류 분쟁사례를 바탕으로 체계개선 등 정책을 제시했던 선행 연구와는 달리, 관세 전문가들에 의해 정확하게 분류된 품목분류 사례 데이터를 활용하여 HS코드를 자동으로 분류하는 자연어처리 기반의 딥러닝 모델을 제안하는 것이다.

본 연구의 구성은 다음과 같다. 2장에서는 HS품목분류 및 딥러닝 기술 등에 대한 선행연구들을 다루고, 3장에서는 본 연구에서 사용한 데이터와 모델 개발 방법론 등에 대해 설명한다. 4장에서는 실험 결과에 대해 설명하고, 마지막으로 5장에서는 요약 및 시사점, 그리고 한계점에 대해 다룬다.

Ⅱ. 관련 연구

2-1 HS 품목분류에 관한 연구

1) 품목분류표의 결정기준

HS코드는 전 세계에서 거래되는 개별 물품을 분류하기 위해 WCO(세계관세기구)가 제정한 국제통일상품분류체계(HS)로서 개별물품에 품목번호를 부여하는데, 체약국은 국제협약에 따라 HS 체계원칙에 맞춰 품목분류를 수행하게 된다[4].

“통일상품명 및 부호체계에 관한 국제협약(제7조, 제8조에)”에 따라 HS위원회가 작성하고 관세협력이사회가 승인한 HS해설서(Harmonized Commodity Description and Coding System Explanatory Notes)에 따르면 무역거래의 대상이 되는 모든 상품은 단일 HS코드 분류 원칙에 따라 분류될 수 있도록 품목분류 통칙이 제시되고 있는데 HS 해석에 대한 통칙(General Rules

for Interpretation of Nomenclature, GRI)은 어떠한 품목도 품목표상 하나의 호에 분류되고 다른 호에 중복 분류되지 않도록 보장하는 것을 목적하고 있다. 통칙은 1호부터 7호까지 규정되어 있는데 1호를 최우선적으로 적용하고 품목분류가 어려운 경우에는 2호에서 4호가 적용되며 5~6호는 필요시 적용하게 된다. HS 품목분류표는 국제무역에서 거래되는 물품을 체계적으로 표시하고 있는데, 부·류·절의 표제에 상품의 범주나 형태를 계통적으로 분류하고 있다. HS 품목분류표는 동일한 원재료부터 얻어지는 모든 물품은 하나의 류(Chapter)에 같이 분류하고 그 각 류(Chapter) 내에서는 생산 가공단계를 기준으로 원재료 > 반제품 > 완제품 순으로 수직배열의 방식을 택하고 있다. [4]는 상이한 원재료를 사용하여 단일 제품으로 생산되는 물품의 경우에는 생산 및 가공단계별 분류체계를 적용하지 않는다고 분석하였다.

2) 품목분류표의 기능

국제무역에 있어서 HS품목분류의 HS코드는 통관 및 승인요건, 수입관세율을 결정하고 FTA 원산지를 증명하는데 활용되며, 수입물품을 가공하여 수출할 경우 관세환급의 기초 자료로도 활용된다.

HS품목분류는 세계 여러 나라와 체결되는 FTA 양허관세율을 결정하는데도 활용된다. FTA가 체결된 국가로부터 물품을 수입하게 되면 관세인하 혜택을 받을 수 있는데 이것이 양허관세율이며 수출입업자들은 수출입물품의 HS코드에 따라 정해진 원산지결정기준을 충족시켜야 혜택을 받을 수 있게 된다. 관세환급은 과세관청이 징수한 조세를 특정한 요건이 충족되었을 때 되돌려 주는 제도로 환급특례법에 의한 관세환급은 수출을 목적으로 관세를 납부하고 수입한 원재료를 가공해 수출한 경우 수출자(생산자)에게 관세를 반환해주는 제도를 말한다. 실제 관세청에서 고시한 관세환급실적을 보면 매년 약 1만여 업체가 11여만회 1800여억 원을 환급받고 있다.

3) 품목분류 마찰

WCO HS협약 제10조에는 품목분류의 분쟁해결방법과 절차가 규정되어 있는데, 품목분류분쟁 조정제도와 정책 수립을 HS 이사회에서 담당하고 있다. 실제로 우리나라의 수출물품이 수입국 세관의 자의적인 품목분류로 인해 부당한 관세를 부과 받거나 통관승인이 어려워져 과세관청에 호소하는 사례가 많다. 특히, 많은 국가와 FTA를 체결하면서 한국산 물품에 대한 관세가 낮아지게 되자 수입국의 품목분류심사가 엄격해지고 있다.

우리나라에서는 수출기업의 부당한 피해를 막기 위해 전담조직을 운영하고 있는데, 2019년 보도에 따르면 외국세관과의 품목분류를 둘러싼 국제분쟁에서 기업들은 약 3,833억 원의 해외 관세비용을 절감하였다. 또한, 관세법령정보포털 결정례에 따르면 품목분류 및 각종 행정절차와 관련해 국내에서도 마찰이 지속적으로 발생하는 것으로 나타났다.

4) 품목분류 사전심사 제도

관세청은 수출입신고를 하기 전에 수출입업자가 물품정보 부족으로 HS분류가 어려운 경우 관세평가분류원장에게 품목분류 사전심사를 신청하면 품목번호를 결정하여 회신하도록 하는 민원제도(관세법 시행령 제106조)를 운영하고 있다. 신청권자는 수출입업자 또는 수출물품 제조자, 관세사를 포함한 관세법인(통관취급법인)이면 누구나 가능하다. 신청방법은 품목분류사전심사 신청서와 증빙서(물품 견본 등)를 첨부하여 인터넷, 우편, 방문 신청하고 수수료를 납부하면 결정서를 통보받게 된다. 신청서에는 신청인정보, 수출입업자정보, 신청물품정보, 신청 사유와 결과 공개여부, 물품설명서(물품품명, 구조 및 형태, 기능 및 용도, 제조공정, 물품사진과 신청인의 분류의견)을 작성하도록 되어 있다. 또한, 품목분류 사전심사 신청내역은 수출입 신고된 물품과 사전심사 신청 물품이 같은 경우 통지내용에 따라 세관장이 품목분류를 적용하도록 법제화되어 있다. 또한, 현재 품목분류사전심사로 처리건 중 공개 신청한 4만 6천여 건이 관세정보법령포탈에서 제공 중이다.

5) HS 코드 분류 연구

HS 코드의 자동 분류에 대한 연구는 분야의 특수성과 데이터 수집의 어려움 등으로 인해 그 중요성에도 불구하고 거의 이루어지지 못했는데, 최근에 수행된 HS 코드 분류를 위한 CNN 기반의 추천 모델 개발 연구[5]가 유일하다. 이 연구는 관세청 품목분류 결정 사례를 이용하여 각 사례에 첨부된 품목의 이미지를 이용하여 HS 코드를 자동 추천하는 연구를 수행하였다. 이 연구는 품목의 이미지 정보를 입력 데이터로 사용하고 CNN 알고리즘을 활용한 반면, 본 연구는 품명 정보를 임베딩한 값을 입력 데이터로 사용하고 순환 신경망 알고리즘 중 하나인 LSTM과 BLSTM을 활용한다는 점에서 차이를 보인다. 따라서, 본 연구는 품목의 이미지 정보가 없는 경우에도 적용이 가능하다는 장점을 지닌다. 또한, [5]는 타겟 변수의 클래스를 빈도수가 높은 5개의 HS 코드로 구성하였지만, 본 연구에서는 모델의 일반화 가능성을 높이기 위해 230개의 HS 코드로 구성하였다는 점에서 차이를 보인다.

2-2 임베딩과 딥러닝 관련 연구

1) 워드 임베딩

텍스트 분류는 분석용도에 맞게 대상 텍스트에 대한 표현 방법을 선택하는 데에서 시작한다. 단어를 표현하는 데에 있어서 해당 단어만 표현하는 국소표현 방법(One-hot Encoding)과 주변 단어를 참조해서 단어를 표현하는 분산표현 방법(Word2Vec)이 있다[6]. 국소표현은 여러 개의 단어집합이 있을 때 해당 단어에만 1을 가지도록 n차원의 벡터로 표현한다. 이러한 표현 방식은 처리할 단어수가 많을수록 벡터 저장 공간이 상대적으로 많이 필요로 하고 처리시간도 길어지는 단점이 있다. 또한 단어의 존재유무를 벡터화하기 때문에 단어간의 의미(맥락)를 파악하기 어려운 단점을 지니고 있다.

분산표현은 국소표현과과 달리 단어를 실수값으로 이루어진 일련의 벡터로 표현한다.

또한, 분산표현은 차원의 크기를 임의로 지정할 수 있는데, 모든 단어의 벡터 표현을 임의의 밀집된 차원에서 표현한다고 하여 밀집벡터라고도 하는데, 이 과정을 워드 임베딩(Word Embedding)이라 하고 이 임베딩 과정을 통해 나온 결과를 임베딩 벡터(Embedding Vector)라 한다. 잘 알려진 워드 임베딩 방법론은 Word2Vec, FastText, GloVe 등이 있다. [7]은 분산표현은 문장 단어들의 앞뒤 단어를 학습하여 벡터 값으로 수치화하게 되며 단어의 벡터를 학습하면서 특정한 문맥에서 유추할 수 있는 확률을 최대화하는 방법으로 학습하므로 비슷한 단어들은 비슷한 벡터값을 가지며, 벡터 공간상에 짧은 거리로 표현할 수 있다고 하였다. 2000년대 초 NNLM(neural network based language model) 방법론에서도 신경망 모델을 기반으로 단어를 벡터로 바꾸는 방법론이 주목을 받기 시작했고 RNNLM(recurrent neural network language modeling)은 NNLM 기능을 업데이트 하여 타겟 단어를 이용해 인접 단어를 예측하는 CBOW(continuous bag of words) 방식과 인접 단어를 이용해 타겟 단어를 예측하는 Skip-gram 방식으로 발전하였으며 현재의 Word2Vec과 같은 임베딩 방법론으로 진화했다[6]. 따라서 워드 임베딩은 작은 차원으로 자료를 표현하고 단어 간의 의미를 표출하는데 용이하여 신경망 텍스트 분석에 많이 활용되고 있다.

2) LSTM(long short term memory) 모델

RNN(Recurrent Neural Network)은 Hidden Layer의 노드가 방향을 가진 엣지로 이어지는 순환 구조를 가진 인공 신경망의 한 종류로 음성, 텍스트, 센서 등 시계열적으로 등장하는 데이터 처리에 주로 사용된다[7]. 또한 RNN은 은닉계층의 활성화 함수를 통해 나온 값이 출력계층에도 전달되기도 하고 다음 은닉계층의 입력값으로 보내진다. 즉 RNN의 현 시점 은닉계층은 현 시점의 입력값과 이전 은닉계층에서 전달해 준 값을 입력값으로 받기 때문에 RNN의 은닉계층을 메모리 셀로도 표현한다.

하지만 RNN은 노드와 노드 사이의 길이가 길어질수록 가중치조정을 위한 역전파 시 점차적으로 이전의 정보가 사라져 오래 전에 발생한 정보와의 연관성을 찾지 못하는 단점이 있다. 이러한 장기간 의존성 문제(The Problem of Long Term Dependencies)를 개선하고자 [8]은 LSTM(Long Short Term Memory) 네트워크를 제안하였다. <그림 1>과 같이 LSTM은 기존 RNN 유닛에 장기간 메모리 역할을 하는 cell state와 정보 전달 강도를 조절하는 3개 gate(forget, input, output)를 추가하여 장기 의존성 문제를 해결했다.

Cell state가 현재 유닛의 정보를 다음 유닛에게 정보를 전달하는 컨베이 벨트와 같은 역할을 하기 때문에 state가 오래 돼도 전파가 잘 되는 특성이 있다. 따라서 전체 체인에서 실행되며 gate에 의해 값을 없애거나 더할 수 있다. LSTM의 첫 시작은 cell state로부터 이전에 받은 값을 보존할지 버릴지를 정하는 것으로 이 gate를 'forget gate layer'라 한다. 이 단계에서는 이전 LSTM 유닛의 출력값인 h_{t-1} 과 현재의 입력값인 x_t 를 받은 후 sigmoid를 입혀서 0과 1 사이의 값을 메모리 C_{t-1} 에 보내주게 된다.

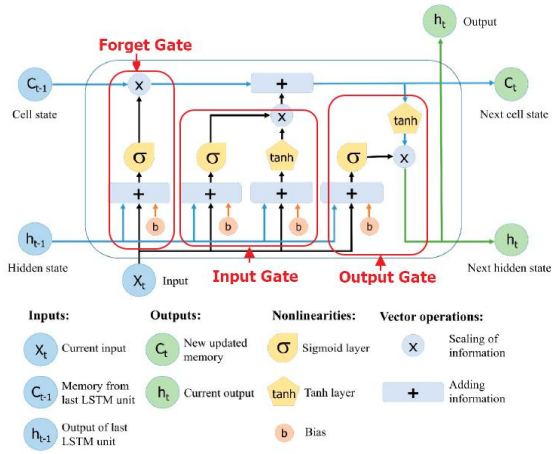

그림 1. LSTM 구조[9]
Fig. 1. Architecture of LSTM[9]

그 값이 1이면 C_{t-1} 값을 보존하고 0이면 삭제하게 된다. 즉, cell_state는 현재 정보를 가지고 있어서 새로운 주제가 왔을 때 기존 정보는 필요 없게 된다. 다음 단계는 새로 들어온 정보를 cell state에 저장할지를 정하는 단계로 먼저 'Input layer'라 불리는 sigmoid layer가 어떤 값을 갱신할지 정한다. 그 후 과거 state인 C_{t-1} 를 업데이트해서 새로운 cell state인 C_t 를 만든다. 마지막 단계는 무엇을 출력할지를 정하는 부분인데 이 출력값은 cell state를 바탕으로 최종 필터링 된 값이 된다.

3) BLSTM(bidirectional LSTM) 모델

BLSTM(Bidirectional LSTM)은 GRU(Gated Recurrent Unit)와 함께 LSTM 신경망을 개선한 알고리즘이다. RNN과 LSTM은 은닉계층에 과거의 데이터 정보를 기억하기 때문에 순차적인 시계열 예측에 적합하다. 하지만 LSTM 신경망은 입력값을 시간 순서대로 입력받기 때문에 결과물이 직전 패턴을 기반으로 수렴한다는 한계가 존재한다[10]. 이 문제점을 개선하기 위해 [11]은 BRNN(Bidiretional Recurrent Neural Networks, BRNN)을 제안하였는데 BRNN은 순방향과 역방향 2개 형태로 신경망을 학습하게 된다. BRNN도 일반 순환신경망 보다 높은 성능을 보였지만 데이터가 길어질수록 장기의존성 문제를 지닌다.

그에 반해 BLSTM은 <그림 2>와 같이 기존 순방향으로 처리하는 LSTM 계층에 역방향으로 처리하는 LSTM 계층을 추가한다. 마지막 Hidden Layer의 상태는 두 LSTM 계층의 은닉 사이를 연결한 벡터를 출력한다. 연결 이외에도 더하거나 평균을 내는 방법도 적용할 수 있다.

BLSTM도 오차보정(Back Propagation Through Time, BPTT)과 학습을 위해 RNN이나 LSTM과 마찬가지로 one-to-one, one-to-many, many-to-one, many-to-may, TimeDistributed()을 사용한다.

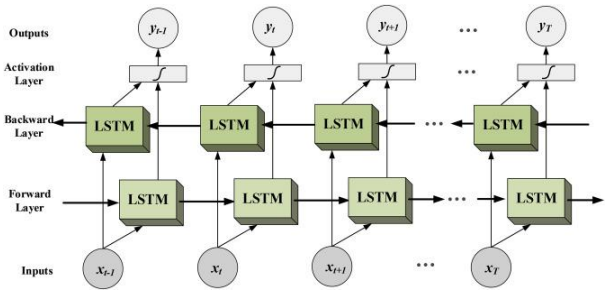

그림 2. BLSTM 구조[10]
Fig. 2. Architecture of BLSTM[10]

Many-to-one는 여러 스텝의 x값을 순차적으로 입력받아 단일 y값을 예측하는 학습법이며 many-to-many는 여러 스텝의 x값을 순차적으로 입력받아 여러 차원의 y값을 예측하는 학습법이다. TimeDistributed()는 모델 레이어 다음 레이어에 사용되고 모델 레이어 각 스텝마다 나온 결과의 Cost(오류)를 계산해서 다음 스텝에 전파하겠다는 것이다. 일반적으로 순차적인 여러 y값을 예측하는 many-to-many 학습법에서 TimeDistributed() 레이어를 적용한다.

Ⅲ. 자동 HS 품목분류 모델 개발

3-1 데이터 수집

본 연구는 품목분류문서의 자동 HS코드 분류를 위해 관세법령정보포탈에서 제공하고 있는 '품목분류 국내사례'를 크롤링하여 데이터를 수집하였다. 수집한 데이터는 1988년부터 2019년까지 21년 기간 동안의 품목분류 결정사례로 총 70,035건이다.

수집한 데이터는 신청인이 입력한 품명, 물품설명, 관련 이미지 등 물품 정보와 이를 바탕으로 관세청에서 결정한 결정세번, 결정사유 등 검토결과 정보, 그리고 자료관리를 위한 참조번호, 시행일자, 유효일자, 시행기관 등으로 구성되어 있으며, 텍스트는 국문과 영문이 혼용되어 있다. 본 연구의 텍스트분류에 활용할 데이터는 '결정세번'과'품명'으로 두 항목을 중심으로 전처리 작업을 수행하였다.

3-2 데이터 전처리

본 연구에서는 1988년부터 2019년까지 21년 동안 품목분류를 결정한 국내사례 데이터를 대상으로 자연어처리 기반의 딥러닝 학습을 통한 HS 코드 자동 분류 모델의 성능을 높이기 위하여 데이터탐색을 통해 결측치를 정비하고 가공하는 3단계의 데이터 전처리 과정을 수행하였다.

1단계에서는 크롤링을 통해 수집한 데이터에서 개행문자 등 불필요한 자료를 제거하는 작업을 수행하였다. 크롤링 수집 단계에서 html 결과물을 확인한 결과, 추출할 컨텐츠 내에'
' '\t', '\n', 'r' 등 개행문자와 html 태그가 포함되어 있어 크롤링 시 개행문자 등의 특수문자를 제거하고 파싱될 수 있도록 하였다. 파싱한 결과는 URL호출 연결이 끊기는 현상방지를 위해 일정 주기로 수집하여 저장한 후, 하나의 파일로 통합하였다.

2단계에서는 결정세번에 대한 전처리 작업을 수행하였다. 결정세번은 HS코드로 10자리 숫자형 데이터형식으로 대부분 존재하나, 전산화 이전의 업무처리와 웹기반 서비스의 편의목적으로 한글, 영문, 특수문자도 혼재되어 있었다. 또한 일부 자료에는 결정세번이 1개 이상인 데이터도 존재하였다. 딥러닝 모델에서 결정세번 항목은 타겟 변수로 사용할 항목으로 숫자 이외의 문자열을 제거하는 작업이 필요하다. 따라서 결정세번에서 숫자만 남기는 전처리 작업을 수행하였고 의미가 없는 10자리 미만의 HS 코드를 포함하는 데이터를 제거하였다.

마지막 3단계에서는 품명자료를 전처리하는 작업을 수행하였다. 품명 항목에는 신청자가 수작업으로 입력한 국문과 영문, 숫자, 특수문자 등의 텍스트가 혼재되어 있었다. 먼저 대문자를 소문자로 치환하고 한글 문자열은 품명을 설명하기 위한 부연자료로 판단하여 제거하였다. 또한, 괄호 안의 문자열이나 단순 숫자열, 그리고 영문자가 하나인 문자나 단어도 품명 정보를 대표할 수 없기 때문에 모두 제거하였다. 또한 여러 공백문자는 단일 공백으로 치환하였고 특수문자는 제거하였다.

3단계에 걸쳐 분석모델에 활용할 독립변수(품명)와 타겟변수(결정세번)에 대한 전처리 작업을 수행한 결과 66,995건의 데이터 셋이 최종 생성되었다.

3-3 데이터 설명

타겟변수인 HS 코드의 1단위를 보면 8류(22,205건), 3류(13,405건), 그리고 2류(11,572건) 순으로 품목분류 의뢰가 많은 것으로 나타났다. HS 8류는 금속류와 전자기기, 기계류의 부품 등이고 HS 3류는 정유, 프라스틱, 화학용품 등이며 HS 2류는 음료·식료품, 담배, 무기화학품 등의 품목으로, 이러한 품목의 품목분류 결정에 어려움이 많은 것으로 보인다. 반면, HS 0류(5건; 산동물, 채소 등), HS 5류(1,418건; 직물, 섬유 등), HS 4류(2,207건; 가죽, 펄프 등), HS 6류(2,586건; 의류, 모자 등) 등은 품목분류 의뢰 건수가 적게 나타났다.

3-4 모델 구축 방법

본 연구에서는 품목분류 국내사례의'품명'텍스트 항목을 이용해 품목분류문서의 데이터 특징을 추출한 후, 이 특징을 이용하여 결정세번(HS 코드)을 자동으로 분류하는 지도학습 기반의 딥러닝 모델을 제안하였다. 실험 환경은 Python 언어 기반의 Tensorflow와 Keras 프레임워크를 사용하였다.

<그림 3>에서 보는 바와 같이 모델을 학습시키기 위한 첫 단계로 본 연구에서는 전처리 작업을 마친 '품명' 항목에 있는 텍스트 데이터에 대해 임베딩 작업을 수행하였다. 우선, 문장을

단어로 토큰화하는 작업을 수행하였다. Keras의 Tokenizer()을 이용해 문장텍스트를 단어 단위로 분할한 결과, 총 31,484개의 단어가 추출되었고 하나의 문장에는 최대 600개의 단어가 존재하였다. 그 후, 각 단어를 해당 단어에 부여된 고유한 정수 값으로 인코딩한 후, 임베딩 층을 통과시켜 밀집벡터(Dense Vector)로 변환하였다. 임베딩 층은 입력 정수를 밀집벡터로 매핑하고 이 밀집 벡터는 신경망의 학습 과정에서 가중치가 학습되는 것과 같은 방식으로 학습되는데, 학습이 완료된 밀집벡터를 임베딩 벡터라 한다. 본 연구에서는 각 단어를 100차원의 벡터로 임베딩하였다.

이후, 딥러닝 모델의 Input값 길이를 한 문장의 최대 단어 수인 600차원으로 구성하고 인덱스 번호가 부여되지 않은 위치에는 pad.sequence()을 이용해 0값으로 패딩하는 작업을 수행하였다. 본 연구에서는 자연어처리와 텍스트분류에서 좋은 성능을 보여준 LSTM과 BLSTM 알고리즘을 이용하여 모델을 구축하였으며, 다중 클래스 분류를 위한 활성화 함수로 softmax를 사용하였다.

모델 작성을 완료한 후에는 모델을 기계가 이해할 수 있도록 컴파일(Compile) 작업을 수행하였다. 컴파일 작업에서는 오차함수와 최적화방법, 메트릭(모델 모니터링) 방법을 설정하였고, 설정을 완료한 후에는 모델 학습을 실시하였다. 학습이 완료된 후에는 실험데이터를 입력하여 모델의 분류 성능(정확도)을 평가하였다. 모델의 성능을 평가하기 위하여, 학습데이터와 검증데이터의 비율은 7:3 비율로 구성하여 실험을 진행하였다.

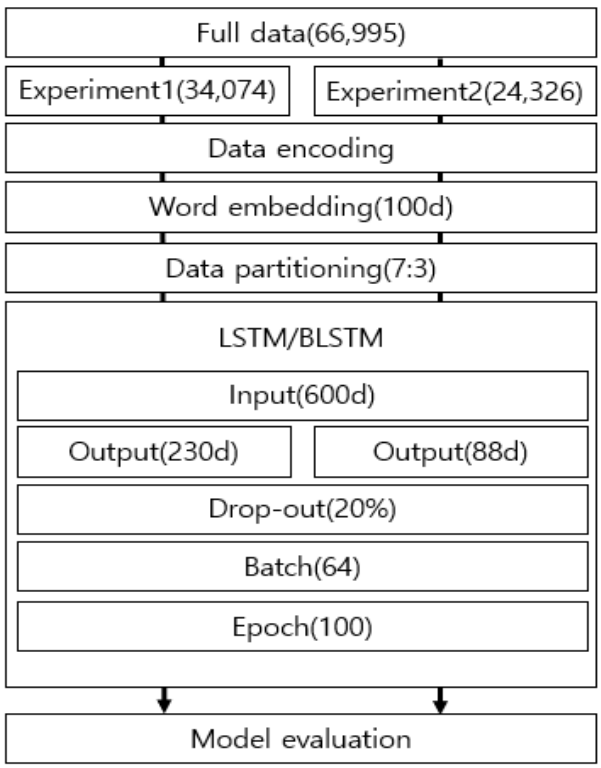

그림 3. 연구 프레임워크
Fig. 3. Research Framework

Ⅳ. 실험 결과

4-1 실험 설계

본 연구에서는 2차에 걸쳐 LSTM과 BLSTM 모델에 대한 실험을 수행하였다.

1차 실험에서는 HS 코드들 중 인스턴스가 50건 이상 존재하는 230개의 HS 코드값을 갖는 34,074개의 인스턴스를 대상으로 학습을 진행하였다.

독립변수의 인코딩 크기는 600차원 설정하고, 각 단어는 100차원의 벡터로 임베딩 하였으며, Drop-out은 20%로 설정하였다. 반복횟수(epoch)는 100회, 배치 크기(batch size)는 64, 출력 뉴런 수는 230개, 학습/검증 데이터 비율은 7:3으로 설정하였다. 1차 실험에서는 HS 코드들 중 데이터가 50건 이상 존재하는 230개 HS 코드를 대상으로 학습하였다.

2차 실험에서는 1차 실험에서와는 달리 HS 코드별 학습데이터의 양을 늘려주고, 자주 등장하는 HS 코드에 초점을 맞춘 모델을 구축하기 위해 HS 코드들 중 인스턴스가 100건 이상 존재하는 88개의 HS 코드값을 갖는 24,326개의 인스턴스를 대상으로 학습을 진행하였다. 하이퍼 파라미터는 1차 실험에서와 동일하게 설정하였다.

4-2 모델 평가

본 연구에서는 분류 정확도(Accuracy)를 측정하여 각 모델의 분류 성능을 평가하였다. 분류정확도는 전체 자료 중에 올바르게 예측한 수의 비율을 뜻한다.

본 연구는 품목분류 국내사례를 바탕으로 기계적 해석을 통해 문서분류를 시도하는 첫 사례로 선행 연구사례가 없어 많은 연구들에서 성능이 입증되어 온 신경망 및 임베딩 알고리즘을 이용하였다.

실험 결과, <표 2>와 같이 실험 1과 실험 2 모든 실험 차수에서 BLSTM 기반 신경망 모델의 분류 정확도가 LSTM 기반 신경망 모델의 분류 정확도 보다 높게 나타났다.

실험 1의 경우 각 데이터 건수가 상대적으로 적은 HS 코드가 포함되어 있고 분류해야 할 HS 코드의 수도 230개로 많아, 각 HS 코드별 데이터 건수가 상대적으로 많고 분류해야 할 HS 코드의 수도 적은 실험 2의 경우보다 낮은 분류정확도를 보여주었다.

품목의 이미지 정보를 이용하여 HS 코드 자동 분류 모델을 개발하였던 [5]의 실험 결과를 보면, 해당 모델은 5개의 타겟 클래스로 이루어져 있는데 최고 73.12%의 정확도를 보였다.

본 연구에서 개발한 모델이 타겟 클래스가 230개인 경우 최고 69.7%, 타겟 클래스가 88개인 경우 최고 71.4%의 정확도를 보이는 점을 감안할 때, 본 연구에서 제안한 모델들의 분류정확도는 상당히 양호한 수준임을 알 수 있다.

이 실험의 결과는 품명 정보를 이용한 HS 코드의 자동 분류 가능성을 보여준다고 할 수 있다.

표 2. 실험 결과
Table 2. Experimental results

Experiment No.	LSTM	BLSTM	No. of Target Classes
1	64.6%	66.1%	230
2	69.7%	71.4%	88

Ⅴ. 결 론

HS 코드는 국가 간의 무역에 있어 물품의 세율 결정, 원산지 증명, 무역 수송 비용, 보험 등에 영향을 끼치는 핵심 정보이다. 특히, 관세율표상의 세율에 따라 징수된 조세로 인해 국내외 온오프라인 시장에도 연쇄적으로 영향을 미치기 때문에 정확한 품목분류 신고는 매우 중요하다고 할 수 있다. 하지만 급격한 산업 환경 변화와 나날이 팽창하는 글로벌 무역, 그리고 고객 니즈에 따른 융복합 신상품의 출시 등은 물품 정보에 대한 이해를 어렵게 하였다. 이러한 물품 정보의 이해 부족으로 인한 잘못된 품목분류 신고는 국내외에서 끊임없는 무역 분쟁을 야기시키고 있다.

본 연구는 관세전문가들에 의해 정확하게 분류된 품목분류 국내사례를 문서분류 성능이 입증된 임베딩 기법과 딥러닝 알고리즘을 활용하여 HS 코드의 자동분류 가능성을 입증한 연구로서 제도개선, 법 절차 및 정책제안을 다룬 관세분야의 선행연구들과는 달리 실제 데이터를 기반으로 품목분류문서의 특징들을 추출하여 HS코드의 자동분류를 시도한 첫 사례라는 점에서 의의가 있다.

본 연구에서는 2차에 걸친 실험을 통해 두 신경망 모델을 비교하였는데, 그 결과 BLSTM 모델이 71.4%로 가장 높은 분류정확도를 보였고, LSTM 모델이 66.1%의 분류정확도를 보였다.

본 연구의 시사점은 다음과 같다. 첫째, 그 중요성에도 불구하고 관세분야의 선행연구들에서 다루어지지 않았던 HS 코드 자동분류를 위해, 데이터 기반의 딥러닝 모델을 관세분야에 적용 시켰다는 점에서 학문적 시사점이 있다고 판단된다. 이를 통해, 데이터에 기반하여 품명 정보의 특징을 분석하고 임베딩 벡터로 변환시켜 학습한 신경망 모델이 HS 코드를 자동분류 할 수 있다는 가능성을 보였다. 이는 산업 환경, 생활패턴 변화에 따라 출시되는 새로운 상품에 대해 지식이 부족해도 물품의 명칭과 같은 정보만으로도 유사 HS코드 추천에 도움을 줄 수 있다는 것이다.

둘째, 본 연구에서 재안한 모델을 활용할 경우 과세관청은 납세의무자의 정확한 수입신고로 국가세수 재정의 안정적 확보뿐만 아니라 민원인과의 품목분류 마찰 감소에 따른 행정력 절감이 이루어질 수 있을 것이며, 더 나아가 정부기관의 신뢰도을 높이는 데 도움을 줄 수 있다는 점에서 실무적 시사점이 있다고 판단된다.

본 연구는 수집 가능한 데이터의 양적인 부족과 관세품목분류 선행 연구의 부재 등으로 인해 더욱 높은 성능을 보이지 못한 한계점이 있다. 따라서, 향후 연구에서는 물품의 명칭뿐만 아니라 물품정보 등 추가적인 독립변수를 발굴하여 여러 신경망을 융합해 다층 Layer를 설계하고, 더욱 정밀한 데이터 특징 추출을 위해 업무 도메인의 특성에 맞춤화된 데이터 사전을 구축하여 모델의 성능을 더욱 향상하고자 한다.

참고문헌

[1] H. K. Jeong, "A Study on the HS Classification Dispute Settlement of Korean Export Goods: Focused on Electronic Goods," Master dissertation, SungKyunKwan University, Seoul, Korea, 2016.

[2] J. W. Chung, "A Study of Customs Friction Caused by Commodity Classification and the Principle of Prohibition of Retroactive Taxation etc.," *Korea Trade Review*, Vol. 29, No. 4, pp.51-70, 2004.

[3] W. S. Sung, J. W. Shim, and E. K. Kim, A Study on Definition and Classification Criteria of Parts and Fittings in a Tariff Rates Table, Customs Valuation and Classification Institute, 2018.

[4] J. Y. Park, M. S. Lee, and S. Lee, A Study on HS Parts Classification Criteria-Focusing on the Analysis of Machine Parts Classification, Customs Valuation and Classification Institute, 2018.

[5] D. J. Lee, G. W. Kim, and K. H. Choi, "CNN-based Recommendation Model for Classifying HS Code," *Management & Information Systems Review*, Vol. 39, No. 3, pp. 1-16.

[6] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," arXiv preprint arXiv:1301.3781.

[7] K. Y. Kim and C. J. Park, "Automatic IPC Classification of Patent Documents Using Word2Vec and Two Layers Bidirectional Long Short Term Memory Network," *The Journal of Korean Institute of Next Generation Computing*, Vol. 15, No. 2, pp. 50-60, 2019.

[8] S. Hochreiter, J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, Vol. 9, pp. 1735-1780, 1997.

[9] X. H. Le, H. V. Ho, G. Lee, and S. Jung, "Application of Long Short-Term Memory(LSTM) Neural Network for Flood Forecasting,", Water, Vol. 11, No. 7, pp. 1387-1405, 2019.

[10] Ö. Yildirim, "A novel wavelet sequence based on deep bidirectional LSTM network model for ECG signal classification.," *Computers in Biology and Medicine*, Vol. 96, pp. 189-202, 2018.

[11] M. Schuster, and K. K, Paliwal, “Bidirectional Recurrent Neural Networks,” *IEEE Transactions on Signal Processing*, Vol. 45, pp. 2673-2681, 1997.

이종권(Jong-Kwon Lee)

2020년 : 한밭대학교 창업경영대학원 빅데이터비즈니스학과 (경영학석사)

2007년～현　재: 데이터월드

※관심분야 : 데이터웨어하우스, 텍스트마이닝, 워드임베딩, 딥러닝(RNN,LSTM,GAN), 데이터 네트워크 가시화 등

최근호(Keunho Choi)

2013년: 고려대학교 대학원 (경영학 박사)

2018년～현　재: 한밭대학교 융합경영학과

※관심분야 : 추천시스템, 의료 빅데이터 분석, 딥러닝, 머신러닝, 데이터 마이닝 등

김건우(Gunwoo Kim)

2010년: 고려대학교 대학원 (경영학 박사)

2011년～현　재: 한밭대학교 융합경영학과

※관심분야 : 비즈니스 온톨로지 모델, 빅데이터 분석, 핀테크 기술 및 전략 등