J. Internet Comput. Serv.
ISSN 1598-0170 (Print) / ISSN 2287-1136 (Online)
http://www.jics.or.kr


# T-EBOW를 이용한 취업알선 챗봇용 단문 분류 연구☆

# Short Text Classification for Job Placement Chatbot by T-EBOW

김 정 래[1] 김 한 준[1] 정 경 희[2*]
Jeongrae Kim Han-joon Kim Jeong Kyoung Hee

**요 약**

최근 각종 사업 분야에서 기업들은 기존 메신저 플랫폼에 인공지능을 더하여 다양한 환경을 대상으로 챗봇 서비스 지원에 주력하고 있다. 취업알선 분야의 기관에서도 취업상담 서비스 품질 제고와 상담 인력 해소를 위해 챗봇 서비스를 요구한다. 일반적인 텍스트 기반 챗봇은 입력된 사용자 문장을 학습된 문장으로 분류하여 적합한 답변을 사용자에게 제공한다. 최근 소셜 네트워크 서비스의 활성화 영향으로 챗봇에 입력되는 사용자 문장은 단문으로 입력되는 경향이 있다. 따라서 단문 분류의 성능향상은 챗봇 서비스의 성능향상에 기여할 수 있다. 본 연구는 취업알선 챗봇을 위한 단문 분류 강화를 위해 기존 연구의 개념 정보뿐만 아니라 번역문 정보를 활용하는 방법인 T-EBOW (Translation-Extended Bag Of Words)를 제안한다. T-EBOW를 기계학습 분류 모델에 적용한 단문 분류의 성능은 기존 방법에 비해 우수한 성능 평가 결과를 보였다.

☞ 주제어 : 취업알선, 챗봇, 단문, 분류

**ABSTRACT**

Recently, in various business fields, companies are concentrating on providing chatbot services to various environments by adding artificial intelligence to existing messenger platforms. Organizations in the field of job placement also require chatbot services to improve the quality of employment counseling services and to solve the problem of agent management. A text-based general chatbot classifies input user sentences into learned sentences and provides appropriate answers to users. Recently, user sentences inputted to chatbots are inputted as short texts due to the activation of social network services. Therefore, performance improvement of short text classification can contribute to improvement of chatbot service performance. In this paper, we propose T-EBOW (Translation-Extended Bag Of Words), which is a method to add translation information as well as concept information of existing researches in order to strengthen the short text classification for employment chatbot. The performance evaluation results of the T-EBOW applied to the machine learning classification model are superior to those of the conventional method.

☞ keyword : Job placement, Chatbot, Short text, Classification

## 1. 서 론

챗봇은 “업무 프로세스를 자동화하기 위해 문자 메시지를 수신하여 마치 인간과 비슷하게 사용자와 직접 대화를 나눌 수 있는 컴퓨터 프로그램”으로 정의한다 [1]. 챗봇은 기존의 온라인 텍스트 기반 대화인 채팅 내용을 분석하여 필요한 서비스를 사용자에게 제공한다. 최근 다양한 분야에서 채봇을 통한 서비스가 상용화 되고 있으며, 취업알선 분야에서도 취업알선 상담 서비스 품질 향상과 상담 인력 해소를 위해 챗봇 서비스를 위한 연구가 요구되고 있다 [2].

챗봇에 입력되는 사용자 문장(질문)의 특성은 SNS (Social Networking Service)활용 증가를 배경으로 단문 (short text) 형태의 문장이 많다. 또한 일상 대화를 포함하는 구어체의 문장, 특정분야의 전문 어휘를 포함하는 문장이 많다. 특히, 신조어, 축약어, 이모티콘 등을 자연스럽게 사용하고 있는 실태이다. 대표적인 텍스트 기반 챗봇 방식은 입력된 사용자 문장을 챗봇이 학습한 문장과 비교하여 유사한 문장으로 분류하여 정해진 답변을 사용자에게 제공한다 [3]. 그럼으로 단문의 분류 성능 향상은

1 School of Electrical and Computer Engineering, University of Seoul, Seoul, 02054, Korea.
2 Powder Research Institute, EngTech CO., Chuncheon-Si Gangwon-Do, 24252, Korea.
* Corresponding author (unikhee@gmail.com)
[Received 26 December 2018, Reviewed 2 January 2019(R2 11 March 2019), Accepted 01 April 2019]
☆ 본 연구는 2018 년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업(No. NRF-2018R1D1A1A02086148)이며, 또한 과학기술정보통신부 및 정보통신기술진흥센터의 대학 ICT 연구센터지원사업의 연구결과로 수행되었음 (IITP-2019-2018-08-01417).

챗봇 서비스의 품질을 향상시킬 수 있다.

본 연구는 챗봇에 입력되는 단문의 분류 성능을 향상하기 위해 단어의 정보를 풍부하게 방법으로 접근한다. 제안한 논문에서는 문장의 단어 출현 빈도를 활용하는 기존 BOW(Bag Of Words)의 한계점을 극복하기 위해 새로운 T-EBOW (Translation-Extended Bag Of Words)을 제안한다. BOW는 문장에서 사용된 단어만을 활용하기 때문에 제한된 정보로 문장 분류 성능 향상에 한계가 있다. 그래서 본 연구는 제한된 정보를 확장하는 방법으로 접근하였다. T-EBOW는 단문의 정보를 개념 정보와 번역문 정보를 활용하여 단어의 정보를 풍부하게 확장하는 방법이다. 개념 정보 활용은 단어의 의미를 활용하는 방법으로 단어 유의어 사전을 사용하여 대표 단어로 변환한다. 번역문 정보 활용은 번역기의 정보를 활용하는 방법이다. 번역기는 단문을 포함하여 많은 문장의 정보를 갖고 있다. 그리고 문장의 의미를 내포하여 외국어로 번역하기 때문에 의미정보를 포함하는 외부정보로 활용할 가치가 있다고 판단한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구로 챗봇의 정의와 활용 분야에 대하여 살펴보며, 3장에서는 T-EBOW에 대한 설계, 4장에서는 T-EBOW을 분류 모델에 적용하여 성능평가 실험 및 실험 결과를 설명한다. 마지막 5장에서는 결론 및 향후 연구 과제를 제시한다.

## 2. 관련연구

### 2.1 챗봇 활용 분야

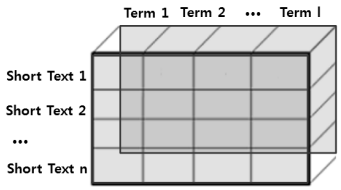
최근 기업들은 챗봇 플랫폼 기술과 API (Application Programming Interface)를 개발함으로써 챗봇 응용 솔루션 개발에 주력하고 있다. 또한, 기존 메신저 플랫폼에 인공지능을 더하여 각종 사업 분야에서 다양한 환경을 대상으로 서비스를 지원한다. 챗봇을 통한 정보제공은 웹을 기반으로 하고 있으며, 스마트 기기에서도 서비스가 가능한 응용프로그램으로 확장되어 일반화되고 있다. 국내외의 대표적인 ICT (Information & Communication Technology) 기업들은 챗봇 서비스 개발 플랫폼(예를 들어, 페이스북 메신저(Facebookmessenger), 위챗(Wechat), 텔레그램(Telegram), 봇샵(Botshop), 알로(Allo), 라인(Line), 카카오톡 (Kakaotalk) 등)을 제공하고 있다. 다양한 챗봇 플랫폼 지원으로 챗봇 서비스는 금융, 보험, 의료, 교육 등 다양한 분야에서 적용되며, 서비스 범위는 확대되고 있는 실태이다. 그리고 챗봇은 일관된 품질의 서비스를 제공하고, 근무 시간과 근무 여건에 관계없이 서비스를 제공할 수 있다. 따라서 다양한 사업 분야에서 정보 제공의 수단으로 챗봇의 활용 가치는 높다.

### 2.2 챗봇의 원리

텍스트 기반의 챗봇은 패턴인식 (pattern recognition), 자연어처리 (natural language processing), 시멘틱웹 (semantic web), 텍스트마이닝 (text mining), 상황인식 컴퓨팅 (context aware computing)과 같은 주요 기술을 필요로 한다 [4]. 텍스트 기반 정보 제공 챗봇은 사용자 질문 적합한 정보를 제공하기 위해 질문과 답변으로 구성된 데이터세트를 학습한다. 챗봇은 입력된 사용자 질문을 학습된 데이터세트에서 가장 유사한 질문으로 분류하고, 분류된 질문의 답변을 선택하여 사용자에게 제공한다 [3]. 그럼으로 챗봇이 사용자 질문에 적합한 답변을 제공하기 위해서는 우선적으로 사용자 질문의 분류 (classification) 문제를 해결하여야 한다.

### 2.3 EBOW (Extended Bag Of Words)

문장의 유사도 측정을 위해 문장의 단어를 기준으로 BOW (Bag Of Words)를 형성하고 단어의 출현빈도를 활용하여 유사도를 계산한다 [5, 6]. 그림 1은 BOW 을 표현한 것으로 문장의 단어 출현빈도를 기준으로 구성한 DTM (Document Term Matrix)이다. 행은 문서 (document)로서 단문 (short text)을 의미하고, 열은 단어 (term)로서 단문에 포함되는 단어 차원을 의미한다. 그리고 DTM 행렬의 각 셀은 해당 단어의 출현빈도 값을 가진다.




(그림 1) BOW 매트릭스
(Figure 1) BOW matrix

EBOW (Extended Bag Of Words)는 BOW의 제한된 정보를 극복하기 위해 단어를 확장하는 방법이다. 그림 2는

단문의 정보 확장을 위한 EBOW 방법의 이해를 위한 그림이다. 외부정보는 지식 (Knowledge)을 활용하여 단문의 정보를 확장하는 방법이다. 기존 EBOW를 위한 방법으로 문장의 다양한 정보를 활용하기 위하여 지속적인 연구가 진행되고 있으며, 이러한 사례는 WordNet, Wikipedia, Open Directory Project를 통하여 제시되기도 하였다. 이전 연구는 단어의 유사어, 반의어, 상.하위어를 포함하는 외부정보를 활용하여 문장분류의 성능을 향상시켰다. 단어의 의미를 고려한 이전 연구는 개념 (concept) 정보를 적용하는 모델을 보였다 [7]. 본 연구의 제안기법과 성능 비교를 위하여 개념 정보를 적용하는 모델을 C-EBOW (Concept-Extended Bag Of Words)으로 정의하여 구현하였다. C-EBOW는 고용정책 관련된 전문용어 사전을 외부정보를 활용하여 개념 처리를 통해 단어의 정보를 확장한 방법이다. 본 논문은 문장의 전체 의미를 고려하여 번역문을 외부정보로 활용하는 방법을 제안하고자 한다.


(그림 2) EBOW 방법
(Figure 2) EBOW method

## 3. T-EBOW 기반 취업알선 챗봇

취업알선 챗봇에 적용되는 단문의 분류 강화를 위한 T-EBOW (Translation-Extended Bag Of Words)에 대하여 설명한다. 그리고 T-EBOW에 적용되는 모델에 대하여 기술한다.

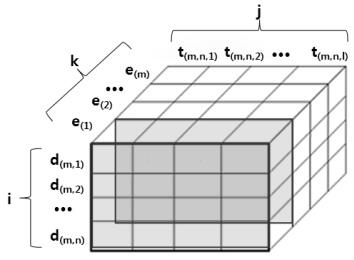
### 3.1 취업알선 텍스트 환경

취업알선 분야에 적용되는 챗봇에 입력되는 단문에는 다양한 의미의 단어를 포함하고 있다. 이러한 문장은 취업알선의 직종 및 직무에서 활용되어지는 무한한 단어를 포함할 수 있도록 알고리즘 개선이 요구된다. 단문의 분류 문제점을 살펴보면, 단문의 유사성을 측정하기에 통계적 단어 정보가 미흡하며, 사전에 없는 두문문자가 존재하며, 단어는 다르지만 의미가 같은 동의어, 단어는 같지만 의미가 다른 다의어, 오탈자, 비문법 등이 존재함을 알 수 있다 [8].

제안한 논문에서는 단문의 단어를 풍부하게 만들면 문장 분류의 성능이 향상됨을 가설로 놓았으며, 고용관련 전문용어 사전과 번역문을 반영하여 단문 분류의 성능이 향상됨을 실험 결과로 증명한다.

### 3.2 T-EBOW 정의

제안한 논문에서는 단문의 정보를 확장하기 위해 다수 개의 외부정보를 활용하여 단문의 정보를 풍부하게 하였다. T-EBOW는 기본 문장의 정보를 확장하기 위하여 단어의 개념 정보와 번역 문장을 활용한다. 활용된 외부정보는 지역 명칭, 직종 구분, 대학 명칭, 자격증 종류, 고용정책 관련된 전문용어 사전과 단문의 영어 번역문이다. 그림 3은 T-EBOW의 구조를 보여준다. $i$는 단문들을 의미하고, $j$는 단문에 포함되는 단어들을 의미한다. 그리고 $k$는 단문의 단어 정보를 풍부하게 하는 방법으로 개념 정보 지식과 번역문들을 의미한다. 개념 정보 지식은 고용 정보관련 전문용어 사전들이 될 수 있다. 번역문을 활용하여 단어의 정보를 확장하는 방법으로 영어, 일본어, 독일어 등으로 확장할 수 있다.




(그림 3) T-EBOW 구조
(Figure 3) T-EBOW structure

· Word extending methods (concept, translation)
$$e_{(k)}(k \in [1,m])$$
· Short text (document)
$$d_{(k,i)}(i \in [1,n])$$
· Term
$$t_{(k,i,j)}(j \in [1,l])$$
· Short text corpus
$$D_k = \{d_{(m,1)}, d_{(m,2)}, d_{(m,3)}, ..., d_{(m,n)}\}$$

그림 3은 T-EBOW에서 사용하는 심볼의 의미를 설명한다. $e_{(k)}$는 단어의 정보를 풍부하게 하는 정보 확장 방

법으로 $m$개의 방법을 고려할 수 있다. 즉, 지식 사전의 개수와 다수의 외국어에 따른 번역문 개수이다. $d_{(k,i)}$는 단문이며, 단어의 정보 확장 방법 $k$에서의 단문 $i$를 의미한다. $t_{(k,i,j)}$는 단문의 단어이며, 단어의 정보 확장 방법 $k$에서의 단문 $i$의 단어를 의미한다. $D_k$는 단어의 정보 확장 방법 $k$에서의 단문들의 집합이다.

본 연구의 DTM은 tf*idf (term frequency inverse document frequency)를 기반으로 한다. tf*idf는 문장에서 출현 빈도가 높은 단어에 가중치를 높이고, 많은 문장에서 출현하는 단어는 반대로 가중치를 낮추는 방식이다 [9]. 단어의 출현 빈도가 높지만 많은 문장에서 출현하는 단어는 중요도가 낮은 단어로 계산된다. 이를 통해 중요도가 낮은 단어가 출현빈도가 크다고하여 그것의 가중치가 높아지는 단점을 보완할 수 있다. tf*idf의 수식은 식 1과 같다.

$$tf*idf(d_k,t_k) = tf(d_k,t_k) \times \log(\frac{|D_k|}{df(t_k)}) \qquad \text{식 1}$$

여기에서, $tf(d_k,t_k)$는 단어의 정보 확장 방법 $k$에서의 단어 $t_k$가 단문 $d_k$내에서 발생한 빈도이다. $df(t_k)$는 $t_k$가 나타난 단문의 개수이다. $|D_k|$는 단어의 정보 확장 방법 $k$에서 데이터세트의 모든 단문의 개수이다.

### 3.3 T-EBOW 적용 분류 모델

단문 분류에 적용하는 분류 모델로서 DT (Decision Tree), KNN (K-Nearest Neighbor,), SVM (Support Vector Machine), RF (Random Forest), DNN (Deep Neural Networks), CNN (Convolution Neural Networks), RNN (Recurrent Neural Networks) 지도 기계학습 모델 [10, 11]을 적용한다. T-EBOW의 성능을 비교하기 위해 기본 단어로 구성된 BOW와 개념 처리를 통해 단어의 정보를 확장한 C-EBOW (Concept-Extended Bag Of Words)를 구현하여 비교한다. 표 1은 단어 정보 확장 방법에 따른 BOW, C-EBOW, T-EBOW 방법의 구분을 보인다. BOW는 기본 문장의 명사와 어근을 추출하였다. C-EBOW의 개념정보는 지역 명칭, 직종 구분, 대학 명칭, 자격증 종류, 고용정책 관련된 전문용어를 포함하는 사전을 생성 및 활용하여 유사의미를 갖는 단어들의 대표단어로 변경하였다. T-EBOW의 번역문 정보는 구글 문서 번역기(https://translate.google.com)를 사용하여 생성하였다.

(표 1) BOW, C-EBOW, T-EBOW 방법
(Table 1) BOW, C-EBOW, T-EBOW method

| 구 분 | 단어 정보 확장 방법 |
|---|---|
| BOW | 기본문장 활용 |
| C-EBOW | 기본문장 + 개념 정보 |
| T-EBOW | 기본문장 + 개념 정보 + 번역문 정보 |

T-EBOW를 이용한 챗봇의 단문 분류 흐름도를 그림 4에 표현하였다. 사용자 질문은 단어의 정보를 확장하여 tf*idf를 기준하여 DTM으로 생성된다. 학습된 분류 모델은 사용자 질문을 유사한 질문 유형으로 분류하고 분류된 질문에 대한 답변을 제공한다.


(그림 4) 취업알선 챗봇의 단문 분류 흐름도
(Figure 4) Flow chart for short text classification of job placement chatbot

## 4. 성능 평가

제안 기법의 성능 평가를 위한 데이터세트, 평가 지표, 실험환경에 대하여 설명한다. 그리고 실험 결과에 대하여 논의한다.

### 4.1 데이터세트

실험에 적용한 데이터세트는 취업알선 챗봇에 입력되는 단문과 그 단문의 의미가 유사한 유형으로 구분한 클래스로 구성된다. 취업알선 챗봇에 입력되는 단문은 설문을 통해 생성되었다. 설문은 남·여. 연령 모두 랜덤하게 1,406명을 대상으로 하였으며, 설문지역은 춘천지역으로 취업관련 챗봇에 질문하고 싶은 내용이 무엇인지, 고용관련 챗봇의 요구사항 등을 주요사항으로 하였다. 설문 예로 “취업알선 챗봇에게 질문하고 싶은 문장을 간략하게 적어보세요”이며, 설문결과 챗봇에 대한 사용자 질문 유형을 분류 (전문가 집단에 의해 13개 유형)하였다. 사용자 질문의 유형은 표 2와 같다. 데이터세트는 2개 속성 (사용자 질문, 클래스)과 1,173개의 레코드 (1,406개 설문에서 무응답 233개 제외)로 구성된다. 13개 클래스의 레코드

개수는 다양한 분포 (6개부터 293개까지)로 불균형하다. 사용자 질문은 대체적으로 6개 이하의 단어로 구성된 단문이다.

(표 2) 사용자 질문 유형으로 구성된 클래스
(Table 2) Class by user question type

| Class | 사용자 질문 유형 | 개수 |
|---|---|---|
| 1 | 일자리 문의 (지역) ; OO 지역에 일자리가 있나요? | 238 |
| 2 | 일자리 문의 (직종) ; OO 관련된 일자리가 있나요? | 39 |
| 3 | 일자리 문의 (경력) ; 신입, 경력, 인턴 | 58 |
| 4 | 일자리 문의 (기업별) ; 대기업, 중견강소기업, 외국계, 공기업, 상장, 공무원 | 8 |
| 5 | 일자리 문의 (전공계열별) ; 공학, 자연, 사회, 인물, 예체능, 교육 등 | 6 |
| 6 | 취업 우대조건 ; 국가유공자, 청년층, 장년층, 여성, 장애인 | 89 |
| 7 | 직업훈련 관련 문의 | 31 |
| 8 | 직업 정보 관련 문의 | 91 |
| 9 | 맞춤서비스 취업 지원 관련 문의 | 57 |
| 10 | 취업성공패키지 취업 지원 서비스 관련 문의 | 27 |
| 11 | 취업 준비 관련 문의 | 293 |
| 12 | 취업 지원 사이트 요구사항 | 191 |
| 13 | 취업알선과 관련 없는 문의 | 45 |

## 4.2 평가 지표

본 연구의 제안 기법을 평가하기 위해 혼합 행렬을 이용하여 정확도 (Accuracy), $F_1$, 정확률 (Precision), 재현률 (Recall) 평가지표를 활용한다. 불균형적 분포를 가지는 13개 클래스를 고려하여 5-fold 교차검증 (cross validation)을 적용하였다. 본 연구의 데이터세트는 다중 클래스임으로 매크로 평균 (Macro averaging)을 활용하여 $F_1$, 정확률 (Precision), 재현률 (Recall)을 계산한다. 표 3은 혼합 행렬 (confusion matrix)이며, 테스트 데이터에 대한 예측 시행을 수행할 때 예측 (분류) 결과의 클래스 개수에 대한 정오 평가표를 의미한다. 여기서 TP (True Positive)와 TN (True Negative)는 올바르게 분류한 경우에 해당하며, FP (False Positive)와 FN (False Negative)는 틀리게 분류한 경우에 해당한다 [12].

(표 3) 혼합 행렬
(Table 3) Confusion matrix

| 구 분 | | Predicted class | |
|---|---|---|---|
| | | True | False |
| Actual class | True | *TP* | *FN* |
| | False | *FP* | *TN* |

각 평가지표에 대한 수식 정의는 다음 식 2-5와 같다.

$$Precision(P) = TP/(TP+FN) \quad \text{식 2}$$

$$Recall(R) = TP/(TP+FP) \quad \text{식 3}$$

$$F_1 = 2PR/(P+R) \quad \text{식 4}$$

$$Accuracy = \frac{(TP+TN)}{(TP+FP+FN+TN)} \quad \text{식 5}$$

매크로 평균 (Macro averaging)을 적용한 수식 정의는 다음 식 6-8과 같다 [13].

$$Precicion_{Macro\ averaging} = \frac{1}{|C|}\sum_{c\in C} Precision_c \quad \text{식 6}$$

$$Recall_{Macro\ averaging} = \frac{1}{|C|}\sum_{c\in C} Recall_c \quad \text{식 7}$$

$$F_{1\,Macro\ averaging} = \frac{1}{|C|}\sum_{c\in C} F_{1\,c} \quad \text{식 8}$$

## 4.3 실험 환경

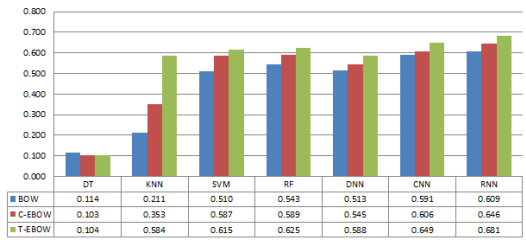
본 연구의 실험 환경은 표 4와 같으며, Ubuntu 운영 체제와 R 기반의 여러 머신러닝 라이브러리가 사용되었다.

(표 4) 실험 환경
(Table 4) Experimental environment

| 구 분 | Equipment | Content |
|---|---|---|
| Hard ware | OS | Ubuntu 16.04LTS |
| | CPU | 3.40GHz |
| | Memory | 16G |
| Soft ware | R (ver.3.4.2) libraries | rpart, e1071, gbm, randomForest, cluster, h2o |

### 4.4 실험 결과

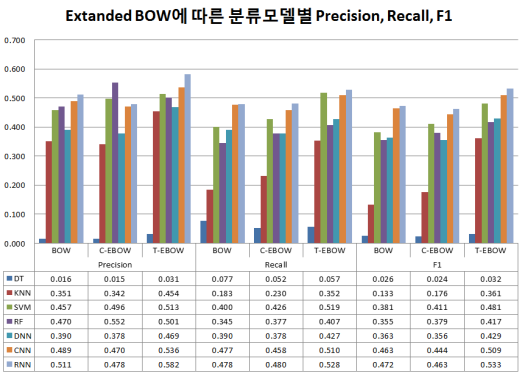
그림 5는 단어의 정보 확장 방법에 따른 분류 모델별 정확도 실험 결과이다. x축은 분류 모델들이며, y축은 BOW, C-EBOW, T-EBOW별 정확도이다. 각 분류모델에서 개념 정보만을 활용한 C-EBOW는 기본문장을 활용한 BOW보다 우수한 성능을 보였다. 그리고 개념 정보와 영어 번역문 정보를 활용한 T-EBOW는 BOW와 C-EBOW보다 우수한 정확도 성능을 보였다. RNN 분류 모델의 경우 정확도 측면에서 C-EBOW는 BOW보다 약 4% 성능향상을 보였으며, T-EBOW는 C-EBOW보다 약 4% 성능향상을 달성하였다.




| | DT | KNN | SVM | RF | DNN | CNN | RNN |
|---|---|---|---|---|---|---|---|
| BOW | 0.114 | 0.211 | 0.510 | 0.543 | 0.513 | 0.591 | 0.609 |
| C-EBOW | 0.103 | 0.353 | 0.587 | 0.589 | 0.545 | 0.606 | 0.646 |
| T-EBOW | 0.104 | 0.584 | 0.615 | 0.625 | 0.588 | 0.649 | 0.681 |

(그림 5) EBOW 방법에 따른 분류 모델별 정확도
(Figure 5) Accuracy by classification models

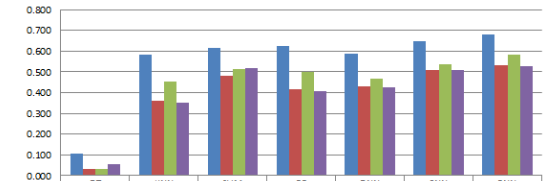
그림 6은 단어의 정보 확장 방법에 따른 분류 모델별 정확률 (Precision), 재현률 (Recall), $F_1$ 실험 결과이다. x축은 분류모델별 평가지표이다. y축은 평가지표별 결과값이다. 정확도와 같이 정확률, 재현률, $F_1$의 경우에도 T-EBOW가 일관되게 우수한 경향을 보임을 확인할 수 있었다.




| | BOW | C-EBOW | T-EBOW | BOW | C-EBOW | T-EBOW | BOW | C-EBOW | T-EBOW |
|---|---|---|---|---|---|---|---|---|---|
| | Precision | | | Recall | | | F1 | | |
| DT | 0.016 | 0.015 | 0.031 | 0.077 | 0.052 | 0.057 | 0.026 | 0.024 | 0.032 |
| KNN | 0.351 | 0.342 | 0.454 | 0.183 | 0.230 | 0.352 | 0.133 | 0.176 | 0.361 |
| SVM | 0.457 | 0.496 | 0.513 | 0.400 | 0.426 | 0.519 | 0.381 | 0.411 | 0.481 |
| RF | 0.470 | 0.552 | 0.501 | 0.345 | 0.377 | 0.407 | 0.355 | 0.379 | 0.417 |
| DNN | 0.390 | 0.378 | 0.469 | 0.390 | 0.378 | 0.427 | 0.363 | 0.356 | 0.429 |
| CNN | 0.489 | 0.470 | 0.536 | 0.477 | 0.458 | 0.510 | 0.463 | 0.444 | 0.509 |
| RNN | 0.511 | 0.478 | 0.582 | 0.478 | 0.480 | 0.528 | 0.472 | 0.463 | 0.533 |

(그림 6) EBOW 방법에 따른 분류 모델별 정확률, 재현률, $F_1$
(Figure 6) Precision, Recall, $F_1$ by classification models

그림 7은 T-EBOW을 적용한 분류 모델들의 평가 지표별 성능 결과를 보여준다. RNN 분류 모델은 다른 분류 모델보다 우수한 정확도 성능 (0.681)을 나타냈다. RNN 분류 모델은 정확도뿐만 아니라 $F_1$ (0.533), 정확률 (0.582), 재현률 (0.528)도 다른 분류 모델들 보다 상대적으로 가장 우수한 결과를 보였다. 분류 모델의 정확도 순위는 RNN (0.681), CNN (0.649), RF (0.625), SVM (0.615), DNN (0.588), KNN (0.584), DT (0.104)이였다.




| | DT | KNN | SVM | RF | DNN | CNN | RNN |
|---|---|---|---|---|---|---|---|
| Accuracy | 0.104 | 0.584 | 0.615 | 0.625 | 0.588 | 0.649 | 0.681 |
| F1 | 0.032 | 0.361 | 0.481 | 0.417 | 0.429 | 0.509 | 0.533 |
| Precision | 0.031 | 0.454 | 0.513 | 0.501 | 0.469 | 0.536 | 0.582 |
| Recall | 0.057 | 0.352 | 0.519 | 0.407 | 0.427 | 0.510 | 0.528 |

(그림 7) T-EBOW의 분류 모델별 성능 비교
(Figure 7) Performance comparison by classification models of T-EBOW

본 연구는 실험 결과로부터 T-EBOW가 BOW와 C-EBOW보다 우수한 것으로 평가 확인하였으며, 제안한 T-EBOW는 평가의 신뢰를 확보하였다. 그리고 취업알선 챗봇을 위한 단문 분류 모델은 RNN이 우수함을 실험적으로 확인하였다. RNN이 연속된 정보를 효과적으로 학습하기 때문에 다른 학습모델보다 성능이 우수한 것으로 사료된다.

## 5. 결 론

제안 논문은 취업알선 챗봇에 개념 정보와 번역문 정보를 활용하는 T-EBOW을 적용함으로써 단문의 분류 성능을 향상시킨다. T-EBOW를 적용한 단문 분류 모델들의 성능 비교 결과 RNN 분류 모델의 성능이 상대적으로 우수한 결과를 보인다. 챗봇 사용자로부터 입력된 다양한 단문의 분류 성능을 향상시킴으로써 사용자 요구사항에 대한 적합한 답변을 자동으로 제공할 수 있는 것이다.

최근 취업난의 해결을 위해 취업알선 분야의 기관에서도 취업상담 서비스 품질 제고와 상담 인력 해소를 위해 챗봇 서비스를 요구한다. 이러한 점에서 본 연구 결과는 학문적 가치 뿐 아니라 취업과 고용과 관련된 사회적 요구가 반영된 연구 결과로서 사회문제 해결에 적용 가능

할 것이다.

향후 연구로는 고용관련 대표 개념 정보를 추가 적용하여 보다 다양한 단어로 구성된 단문에 대처하며, 다양한 번역문 활용과 다양한 외부 정보에 대한 적용 방안을 활용할 수 있도록 한다.

## 참고문헌(Reference)


[1] SY Lee, “A text-based artificial intelligence chatbot definition and use case”, Yonsei University 4th Industrial Revolution Brief, No.6, 2018. http://4ir.yonsei.ac.kr/

[2] Ministry of Employment and Labor, “Employment and Labor Policy”, 2018. http://www.moel.go.kr/info/publict/publictList.do

[3] M Yan, P Castro, et al, “Building a chatbot with serverless computing”, Proceedings of the 1st International Workshop on Mashups of Things and APIs. ACM, 2016. http://dx.doi.org/10.1145/3007203.3007217

[4] JS Hwang and JY Oh, “Beyond the Mobile Age and into the AI Age”, IT & Future Strategy, Korea Information Technology Agency, No.7, 2010.

[5] B Sriram, D Fuhry, et al, “Short text classification in twitter to improve information filtering”, Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval. ACM, 2010. http://dx.doi.org/10.1145/1835449.1835643

[6] J Tang, X Wang, et al, “Enriching short text representation in microblog for clustering”, Frontiers of Computer Science Vol.6, No.1, pp.88-101, 2012.

[7] HJ Kim and JY Chang, “A Semantic Text Model with Wikipedia-based Concept Space”, The Journal of Society for e-Business Studies, Vol.19, No.3, pp.107-123, 2014. http://dx.doi.org/10.7838/jsebs.2014.19.3.107

[8] R Navigli, “Word sense disambiguation: A survey” ACM computing surveys (CSUR), Vol.41, 2, 2009.

[9] T Joachims, “A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization” No. CMU-CS-96-118. Carnegie-mellon univ pittsburgh pa dept of computer science, 1996.

[10] PN Tan, M Steinbach, and V Kumar, “Introduction To Data Mining”, Boston: Pearson Addison Wesley, 2006.

[11] DM Christopher, R Prabhakar, and S Hinrich, “Introduction to Information Retrieval”, Cambridge University Press, 2008. https://nlp.stanford.edu/IR-book/

[12] R Xu and DC Wunsch, “Clustering algorithms in biomedical research: a review”, IEEE Reviews in Biomedical Engineering, Vol.3, pp.120-154, 2010. https://doi.org/10.1109/RBME.2010.2083647

[13] Y Yang, “An evaluation of statistical approaches to text categorization”, Information retrieval, Vol.1, 1-2, pp.69-90, 1999.

## ◐ 저 자 소 개 ◑

**김 정 래(Jeongrae Kim)**
2000년 서울과학기술대학교 재료공학과(공학사)
2013년 국방대학교 국방대학원 국방정보관리(이학석사)
2019년 서울시립대학교 일반대학원 전자전기컴퓨터공학부(공학박사)
관심분야 : 데이터마이닝, 기계학습, 딥러닝, 프러드디텍션, 챗봇, etc.
E-mail : ceright@gmail.com

**김 한 준(Han-joon Kim)**
1994년 서울대학교 계산통계학과(공학사)
1996년 서울대학교 일반대학원 전산과학과(이학석사)
2002년 서울대학교 일반대학원 컴퓨터공학부(공학박사)
2002년~현재 서울시립대학교 전자전기컴퓨터공학부 교수
관심분야 : 데이터마이닝, 기계학습, 빅데이터분석, 지능형정보검색, etc.
E-mail : khj@uos.ac.kr

**정 경 희(Kyoung-hee Jeong)**
1998년 가톨릭관동대학교 전자계산학과(공학사)
2000년 가톨릭관동대학교 일반대학원 전자계산공학과(공학석사)
2008년 가톨릭관동대학교 일반대학원 컴퓨터공학과(공학박사)
2019년~현재 ㈜이앤지테크 책임연구원
관심분야 : 데이터분석, 인공지능, 챗봇, etc.
E-mail : unikhee@gmail.com