카드사 고객 상담 Speech-To-Text 기반 상담 카테고리 분류

Jeomwoo Kang
Dept. of Applied Data Science
Sungkyunkwan University
Seoul, South Korea
jeomwoo@skku.edu

ABSTRACT

대부분의 기업은 고객 대상 상담 콜센터를 운영하고 있으며, 상담 품질 관리를 위해 노력하고 있다. 상담 품질로 인한 고객 민원만이 아니라 기업 이미지에도 중요하다. 상담 품질 관리를 위해서는 상담 카테고리 분류가 필수이지만, 상담 녹취 음성을 이용한 상담 카테고리 분류는 너무 많은 비용이 발생한다. A 카드사에서는 이를 해결하기 위해 업무 내용에 따라 상담 카테고리를 나누었고, 상담 도중 업무 시스템 사용 로그 통계를 이용하여 상담원에게 상담 카테고리를 추천하고 상담원은 알맞는 상담 카테고리를 선택해서 간편하게 상담 카테고리를 입력할 수 있도록 했다. 그러나 상담 카테고리 추천 화면이 나오고 선택 후 등록되는 시간이 10 초가 걸렸고 이 시간에 대한 효율화도 필요했다. 평일 약 10 만건의 상담이 이루어지고 있으며, 상담 카테고리 등록에 총 277.78 시간이 소요된다. 해당 시간을 효율적으로 관리한다면 고객의 대기시간을 줄이거나 상담원의 휴식시간을 제공할 수 있다. 해당 시간을 효율화 하기 위해 상담 카테고리 입력에 대해 자동화가 필요하며, 이를 해결하기 위해 기계학습 기법 적용을 연구하게 되었다. 본 연구에서는 상담 녹취 음성에 Speech-To-Text 변환기를 적용하고 이를 통해 얻은 텍스트를 활용하여 텍스트 분류(Text Classification)를 수행하였다. 상담 카테고리는 656 개이며, 전통적인 기계학습 알고리즘부터 심층신경망 알고리즘을 적용하면서 분류 성능을 확인하였고, Sequence to Class 형태의 Attention 기법을 활용한 Bi-LSTM 모델이 F1-Score 기준 0.801 을 나타내며 가장 좋은 결과를 보였다. [1, 2, 8, 9]

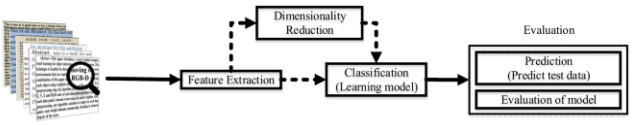
KEYWORDS

Text Classification; Speech-To-Text; Attention; Bi-directional LSTM; Customer Service

1 INTRODUCTION

텍스트 분류는 텍스트를 입력으로 받아, 텍스트가 어떤 종류의 카테고리에 속하는지 구분하는 작업이다. [1] [그림 1-1]과 같이 텍스트에서 특성을 추출하고 기계학습 모델을 학습시켜서 분류를 수행한다. 차원 축소를 통해 특성을 추출하기도 하지만, 이번 연구 내에서는 수행하지 않는다. 텍스트 분류 결과에 따라 이진 텍스트 분류 또는 멀티-레이블 텍스트 분류로 나눌 수 있다. [1]

[그림 1-1] 텍스트 분류



출처: https://github.com/kk7nc/Text_Classification

A 카드사의 고객 상담 텍스트에 대한 상담 카테고리는 멀티-레이블 텍스트 분류 문제로 볼 수 있다. 주어진 데이터셋은 2018 년 1 월부터 2019 년 6 월까지 총 2,065 만건이다. 해당 데이터는 상담 일자, 상담 시간, 상담 텍스트, 상담 카테고리로 되어 있으며, 이번 연구에서는 상담 텍스트를 이용하여 분류를 수행한다. Speech-To-Text 변환기의 인식률이 80%정도 밖에 되지 않아 상담 텍스트에는 노이즈가 많이 존재하고 있으며, 단어 수는 평균 4,000 단어에 이른다.

본 연구에서는 상담 텍스트에 대하여 데이터 전처리를 적용하고 특성을 생성한 후 기계학습 알고리즘을 통해 텍스트 분류를 수행하였다. 데이터 전처리에서는 Konlpy 형태소 분석기를 이용하여 품사 (POS, Part-Of-Speech) 태깅을 통해 명사, 동사 외 품사를 제거하였다. 상담 카테고리는 전체 건수

빈도에 따라 Top-50, Top-88, All 로 구분하여 분류하였다. 특성 추출을 위해 TF-IDF, Word2Vec 을 적용하였고, Binary Relevance, Classifier Chain, Hierarchical Classifier, TextCNN, Bi-LSTM with Attention 을 이용하여 분류를 수행하였다. [3, 4, 5, 6, 7, 8, 9] 분류에 대한 평가는 F1-Score 를 이용하였고, 전통적인 기계학습 알고리즘 보다 심층신경망이 더 효과적이었다. 특히, 심층신경망의 알고리즘 중에 Sequence 를 학습하는 Bi-LSTM with Attention 모델이 0.801 로 가장 좋은 성능을 보여주었다. [2, 8, 9]

2 DESIGN OF EXPERIMENT

2.1 Dataset

A 카드사의 고객 상담 데이터로 2018 년 1 월부터 2019 년 6 월까지의 총 2,065 만건 데이터이다. 데이터셋의 구성은 상담 일자, 상담 시간, 상담 텍스트, 상담 카테고리로 되어 있으며, 상담원에게 추천하고 선택한 상담 카테고리이다. 상담 카테고리는 총 656 개이며, 상담 유형 Top 5 의 내용을 살펴보면 [표 2-1]과 같다.

[표 2-1] 상담 카테고리 TOP 5

상담 카테고리 TOP 5	건수 비중
기타단순문의	16.81%
이용/결재내역 관련 문의	9.66%
바로출금 결제 요청	6.99%
(총)한도확인/단순문의	5.01%
카드사용등록 관련 문의	4.90%

상담 텍스트는 음성 녹취 파일에서 Speech-To-Text 변환기를 적용하면서 의미 없는 잡음 변환도 포함되어 있으며, 띄워쓰기 문제로 인하여 의미 판별이 어렵기 때문에 관련 처리가 필요하다.

[표 2-2] 상담 텍스트 내 이슈

이슈	내용
잡음 변환	음음는, 음음음만, 음음음음음음음음, 저, 어 등
띄워쓰기 문제	행복 한, 원 리금 등

체 건수 2,065 만건 중 상담 카테고리 수에 따른 학습 성능을 확인하기 위해 전체 건수 커버리지 기준 3 가지 그룹을 만들었다. 또한, 업무 내용에 따른 상담 카테고리 생성으로 인해 유사 상담 카테고리가 존재할 수 있기 때문에 그룹별 확인이 필요하다. 세부 그룹은 [표 2-3]과 같이 분류하며, 해당 그룹을 기준으로 학습을 수행한다.

[표 2-3] 상담 카테고리 그룹

그룹	전체 건수 커버리지
Top-50	80%
Top-88	93%
All	100%

상담 텍스트의 단어는 평균 4,000 단어로 이루어져 있으며, 초기 대화에서 대부분의 의도가 있기 때문에 학습 효율성을 위해 초기 300 단어만 사용한다.

2.2 Preprocessing

데이터셋에서 살펴본 상담 텍스트 내 가장 큰 이슈는 잡음 변환과 띄워쓰기 문제이다. 두 가지 문제를 해결하기 위해서 형태소 분석기의 품사 태깅(POS) 를 통해 불필요한 잡음을 제거하고 띄워쓰기 문제를 해결한다. 품사는 [그림 2-1]에 있으며, 전체 중 명사, 동사만 추출하고 나머지는 제외한다.

[그림 2-1] 품사

Universal tag	Description	설명
VERB	All verbs	동사
NOUN	Common and proper nouns	명사
PRON	Pronouns	대명사
ADJ	Adjectives	형용사
ADV	Adverbs	부사
ADP	Prepositions and postpositions	전치사
CONJ	Conjunctions	접속사
DET	Determines	?
NUM	Cardinal numbers	숫자
PRT	participles	분사
X	Other	
.	Punctuation	구두점

제외된 명사, 동사 단어 리스트에서 초기 300 단어만 필터링 하여 특성을 만드는데 사용하였다.

2.3 Feature Extraction

전처리가 완료된 데이터를 이용하여 TF-IDF, Word2Vec 을 수행하였다. [3] TF-IDF(Term Frequency-Inverse Document Frequency)는 단어의 빈도와 역 문서 빈도를 사용하여 Document Frequency Matrix 내의 각 단어들마다 중요한 정도를 가중치로 주는 방법이다. 수식은 [그림 2-2]와 같다.

[그림 2-2] TF-IDF 수식

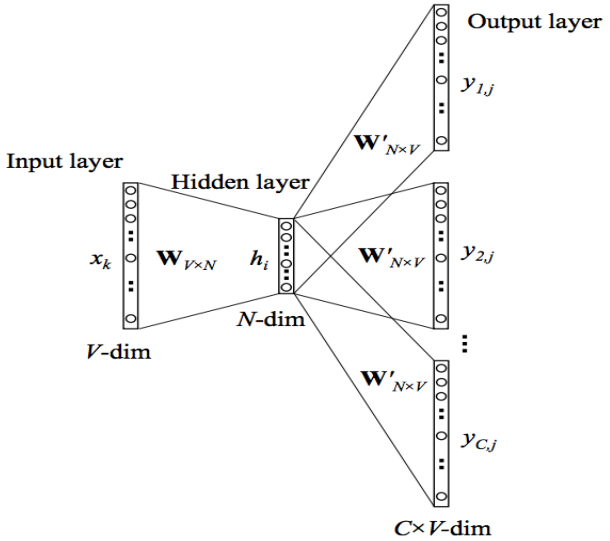
$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF
Term x within document y

$tf_{x,y}$ = frequency of x in y
df_x = number of documents containing x
N = total number of documents

Word2Vec 은 단어를 벡터로 바꿔주는 알고리즘으로 Neural Network Language Model 을 계승하면서 비약적인 발전을 한 알고리즘이다. Word2Vec 은 중심단어로 주변 단어를 예측하는 Skip-gram 과 주변 단어로 중심단어를 예측하는 CBOW 가 있다. CBOW 는 주변 단어를 가지고 중심 단어를 맞추기 때문에 중심 단어는 한 번의 업데이트만 이루어지지만, Skip-gram 의 경우 중심 단어를 여러 번 업데이트가 가능하기 때문에 Skip-gram 을 많이 이용한다. [그림 2-3]은 Skip-gram 의 아키텍처이다.

[그림 2-3] Skip-gram 아키텍처



TF-IDF 는 sklearn, Word2Vec 은 gensim 을 이용하였고, 심층신경망에 이용할 피처는 단어에 대한 Word Sequence 를 만들어서 이용하였다. 학습과 테스트 데이터 생성을 위해 카테고리별 1 만개씩 랜덤 샘플링을 하였다. 그룹별 55 만, 88 만, 656 만개의 데이터를 이용하였다.

2.4 Algorithm

사용한 알고리즘은 5 종류이며, 전통적인 기계학습 알고리즘 3 개와 심층신경망 2 개이다. 세부 알고리즘은 [표 2-4]와 같다. 단어의 문맥에 따라 상담 카테고리를 학습할 수 있다고 판단되며, Sequence 학습 모델인 Bi-LSTM with Attention 이 가장 높은 결과를 보일 수 있다고 기대된다. [2, 8, 9]

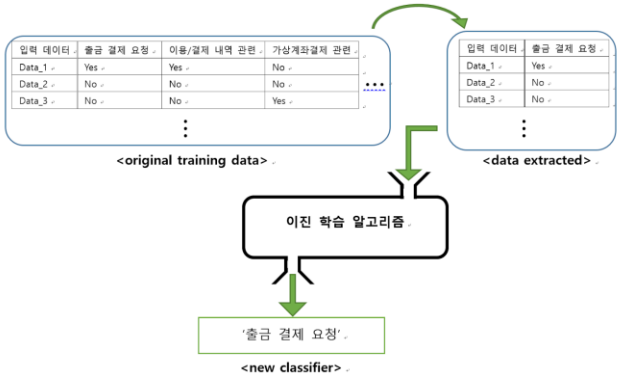
[표 2-4] 활용 알고리즘

구분	알고리즘
전통적인 기계학습	Binary Relevance with Linear SVM
	Binary Relevance with Logistic Regression
	Classifier Chain
	Hierarchical Classifier
심층신경망	TextCNN
	Bi-LSTM with Attention

2.4.1 Binary Relevance

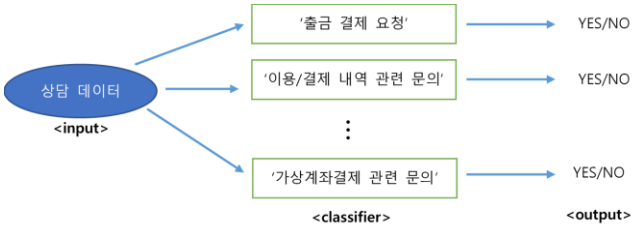
Binary Relevance 는 멀티-레이블 분류에서 가장 직관적인 방법으로, 멀티-레이블 문제를 독립적으로 나눠서 해결할 수 있는 특징을 지닌다. 이 알고리즘을 활용하여 각 항목에 해당 데이터의 부합 여부 확인이 가능하다. Classifier 로 다양한 알고리즘을 사용할 수 있다. 알고리즘의 훈련과정을 살펴보면 [그림 2-4]와 같다.

[그림 2-4] Binary Relevance Classifier 훈련 과정



[그림 2-5]는 작동 원리를 이해 할 수 있다. 이러한 작동 원리로 인해 Classifier 만드는 시간과 사용 시간을 줄일 수 있으며, 레이블 조작의 용이성, 병렬화 수행이 있다. 그러나 레이블 간 상관관계 파악은 어렵다.

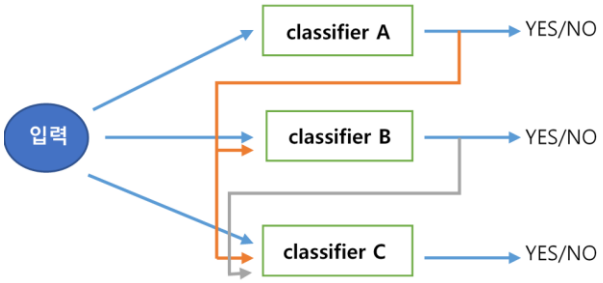
[그림 2-5] Binary Relevance 작동 원리



2.4.2 Classifier Chain

Classifier Chain 은 Binary Relevance 를 보완한 방법으로 레이블 간의 상관관계를 고려할 수 있게 구축되었다. 첫 번째 Classifier 가 계산을 하면 두 번째 Classifier 에게 전달되고, 두 번째 Classifier 가 계산을 끝내면 세 번째 Classifier 에게 전달된다. 마지막에 이르러서는 그 이전의 모든 계산 결과와 입력 데이터를 가지고 출력한다. 작동 과정은 [그림 2-6]과 같다. Binary Relevance 의 구조적 문제를 해결했으나, 복잡한 데이터에 대한 취약성을 지니고 있다.

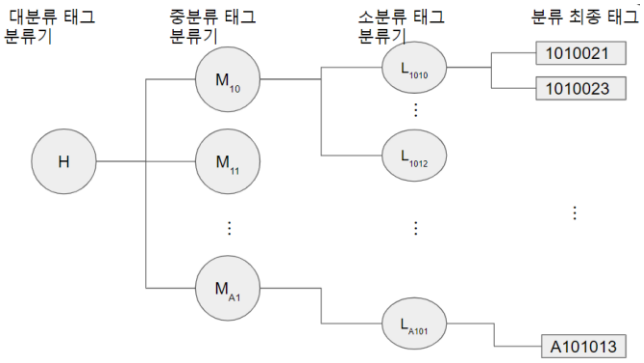
[그림 2-6] Classifier 의 작동 과정



2.4.3 Hierarchical Classifier

Hierarchical Classifier 는 레이블 간의 상관관계를 고려하기 위해 계층 구조를 적용한 것으로, 계층 구조를 직접 설정할 수 있으며 이에 따라 학습을 진행할 수 있다. 치명적인 단점은 사람이 분류 가능한 계층 구조가 없다면 사용할 수 없다. [그림 2-7]은 계층 분류기의 도식화된 모습이다.

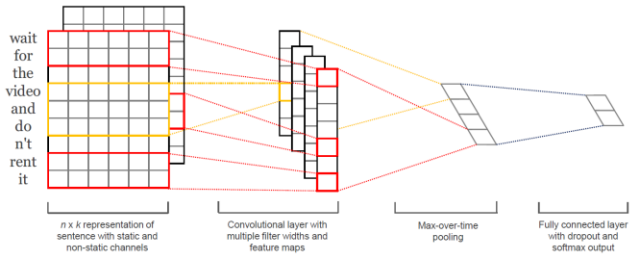
[그림 2-7] Hierarchical Classifier 계층 모습



2.4.2 TextCNN

Convolution Neural Network 를 활용한 것으로 이미지 처리를 위해 CNN 이 많이 사용되었다. TextCNN 은 이미지가 아닌 텍스트에 적용하여 이미지의 지역적인 정보를 추출하는 역할이 텍스트에서는 단어 등장순서/문맥 정보를 보존해준다. CNN 의 필터를 조절하여 다양한 N-gram 모델을 만들어낼 수 있다. 결국 TextCNN 은 문장의 지역 정보를 보존함으로써 단어/표현의 등장순서를 학습에 반영하는 아키텍처이다. [그림 2-8]은 TextCNN 의 아키텍처이다.

[그림 2-8] TextCNN 의 아키텍처

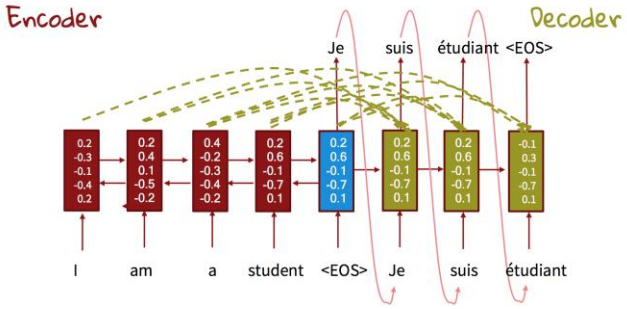


2.4.2 Bi-LSTM with Attention

LSTM(Long Short Term Memory)은 RNN 의 gradient vanishing/exploding 을 해결하기 위해 만들어졌으며, 인코더와 디코더로 구성되어 있는 Seq2Seq 모델에서 매우 좋은 성능을 보여준다. 여기서 양방향(Bidirectional) 네트워크가 적용되면 더욱 더 성능이 개선된다. Bi-LSTM 에 디코더 예측 시 가장 의미 있는 인코더 입력에 주목하게 만드는 Attention 메커니즘을 적용하면 매우 좋은 성능이 나오며, 기계번역에서 많이 활용되었다. [그림 2-9]는 Seq2Seq 기계 번역에서의 Bi-

LSTM with Attention 동작 모습이다. 상담은 대화이며 단어의 시퀀스이므로 해당 알고리즘이 좋은 성능을 보여줄 것으로 기대 한다.

[그림 2-10] Bi-LSTM with Attention 기계번역



2.5 Metric

평가에서는 멀티-레이블 분류에서 많이 이용되는 F1-Score 를 기준으로 성능을 측정한다. F1-Score 는 Precision 과 Recall 의 조화평균으로 분류에서 가장 많이 쓰인다.

[그림 2-2] F1-Score 수식

$$\textbf{Precision} = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

$$\textbf{Recall} = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

3 EXPERIMENT RESULT

실험 결과 단어의 문맥을 학습할 수 있는 Bi-LSTM with Attention 모델이 가장 높은 0.801 의 성능을 보여주었다. [8, 9] 전통적인 기계학습 모델은 학습 시간은 심층신경망에 비해 빠르지만 자연어로 구성되어 있는 텍스트를 단어 문맥 기반으로 분류하기에는 부족한 성능을 보여주었다.

[표 3-1] 실험 결과

알고리즘	그룹	F1
Binary Relevance with Linear SVM	Top-50	0.396
Binary Relevance with Logistic Regression	Top-50	0.381
Classifier Chain	Top-50	0.578
Hierarchical Classifier	All	0.581
TextCNN	Top-88	0.717
Bi-LSTM with Attention	Top-88	0.768
Bi-LSTM with Attention	All	0.801

상담 카테고리 656 개에 대해 0.801 이 매우 높다고 보여질 수 있으나, 실제 상담 카테고리를 살펴보면 유사/중복 상담 카테고리가 많이 존재하고 있어 특정 상담 카테고리는 매우 낮은 성능을 보이고 있다. [이용/결제내역 관련 문의], [승인내역 문의], [해외 이용내역 문의] 등 상담원 마다 다르게 선택할 수 있는 선택할 수 있는 문제가 존재한다.

4 DISCUSSION AND FUTURE WORK

Speech-To-Text 변환기를 통해 변환한 텍스트를 활용하여 기계학습을 통한 상담 카테고리가 충분히 가능함을 알 수 있었다. 그러나 유사 상담 카테고리로 인한 분류 성능 저하는 존재하고 있다. 이를 개선하기 위해 상담 카테고리별 상담 텍스트의 유사도를 계산하여 상담 카테고리의 통폐합을 시도해볼만 하다. 또한, Google 에서 공개한 BERT 를 활용하여 분류를 수행한다면 더욱 더 높은 성능이 예상된다. [10]

5 CONCLUSION

상담과 같은 대화 텍스트에서 특정 카테고리로 분류하는 작업은 전통적인 기계학습 알고리즘에서는 아쉬운 성능이 나왔고, 심층신경망에서는 매우 잘 작동한다. 특히, 대화의 문맥을 이해하기 위해 단어 Sequence 를 학습하는 RNN 계열의 모델은 매우 우수한 성능을 보여주었다. [2, 8] TextCNN 모델도 이미지가 아닌 텍스트에서도 좋은 성능을 보여주었지만 Long-Term 에 대한 부분에서는 Bi-LSTM 을 따라가지 못하였다. [7, 9] 텍스트 분류로 카테고리를 분류하는 경우 알고리즘도 중요하지만 데이터셋이 가지고 있는 카테고리에 대한 체계가 더 중요하다. 특히, 기존 비즈니스에서 사용하고 있는 카테고리는 기계학습을 적용하기에 체계가 맞지 않을 수 있다.

이러한 경우 카테고리에 대한 검증이 사전에 필수적으로 이루어지고 가능여부 판단을 하고 진행해야 한다.

REFERENCES

[1] Kamran Kowari, Kiana Jafari Meimandi, Mojtaba Heidarysafa. Text Classification Algorithms: A Survey. Information 2019, 10, 150; doi:10.3390/info10040150

[2] Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to sequence learning with neural networks. Adv. Neural Inf. Process. Syst. 2014, 27, 3104–3112.

[3] Goldberg, Y.; Levy, O. Word2vec explained: Deriving mikolov et al.'s negative-sampling word-embedding method. arXiv 2014, arXiv:1402.3722

[4] Min-Ling Zhang, Yu-Kun, LiXu-Ying. Binary relevance for multi-label learning: an overview. Frontiers of Computer Science April 2018, Volume 12, Issue 2, pp 191–202

[5] Jesse Read, Bernhard Pfahringer, Geoff Holmes. Classifier chains for multi-label classification. Machine Learning December 2011, 85:333

[6] Carlos N. Silla Jr., Alex A. Freitas. A Survey of Hierarchical Classification Across Different Application Domains.

[7] Yoon Kim. Convolutional Neural Networks for Sentence Classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1746–1751, October 25-29, 2014, Doha, Qatar

[8] Pascanu, R.; Mikolov, T.; Bengio, Y. On the difficulty of training recurrent neural networks. ICML 2013, 28, 1310–1318

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar. Attention Is All You Need. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Google AI Language.