vol.01

Intent Classification

AIQ.TALK STORY

Skelter Labs

사용자에게 실용적 '가치'를 제공할 수 있도록 대화형 인공지능이 고도화될 때, 사용자의 일상이 더욱 편리하고 윤택할 수 있도록 돕는 서비스가 비로소 탄생할 수 있습니다.

Skelter Labs CTO, 조성진

Introduction

어느날 갑자기 '인공지능'은 우리의 일상에서 쉽게 들려오기 시작했습니다. 인공지능은 더이상 개발자나 연구원들에게만 쓰이던 낯선 용어가 아니라, 매일 쓰는 가전기기에도, 모바일 애플리케이션에도, 학교에서도 차 안에서도 심심치 않게 들리는 단어가 됐습니다. 알파고와 이세돌의 바둑 대결 이후 많은 이들이 인공지능 기술의 수준에 놀랐고, 누군가는 두려워하고, 누군가는 찬사를 표하기도 했습니다. 영화 <마이너리티 리포트>에 등장하는 가상 스크린과 유사한 제품은 이미 출시되었고, <아이언맨>의 인공지능 비서 자비스(JARVIS)가 곧 등장할 것이란 이야기도 들려옵니다.

하지만, 여전히 일각에서는 대중을 위한 인공지능은 아직 먼 미래의 일이라고 말합니다. 때로는 기계에게 명령하는 것보다, 사람의 손이 빠르기도 합니다. 심지어 '인공지능' 스피커라 불리는 제품들도 '잘 알아듣지 못했어요'라는 말을 반복합니다.

스켈터랩스는 생각했습니다.

우리의 미션인 '언제, 어디서나 우리의 일상을 이해하고, 도와주고, 더 나아지게 하는 머신 인텔리전스'를 무엇으로, 어떻게 구현할 수 있을지 말입니다. 현재 우리는 사람의 대화 방식에 가장 가까운, 혹은 사람보다 더 예민하게 어감과 문장을 분석하고, 대화를 이해할 수 있는 인공지능의 구현을 위해 기술 고도화에 집중하고 있습니다. 스켈터랩스의 ASR(Automated Speech Recognition) 기술 성능은 글로벌 탑티어 수준에 도달했으며, 실제 현업에서 활발하게 쓰일 수 있도록 소음 환경에서의 정확도를 더욱 개선하는 작업을 진행 중입니다. 또한 이번 글에서 중점적으로 이야기하려는 대화형 인공지능 기술의 핵심, 인텐트 분류(Intent Classification)에 대해서는 2년 여간 다양한 실험 및 개발을 시도했고 누구보다 뛰어난 한국어 성능을 보유하고 있습니다. 앞으로의 보고서 내용은 인텐트 분류가 진행되는 과정을 개략적으로 설명하고, 스켈터랩스가 거둔 유의미한 성능 결과를 담고 있습니다.

Hello!

우리는 컴퓨터를 사용할 때, 버튼을 클릭하고 명령어를 입력해왔습니다. 그러나 지금, 인공지능 기술이 발전하며 사람과 대화를 나누듯 컴퓨터와 소통할 수 있는 다양한 방법이 속속들이 등장하고 있습니다. 이러한 배경에는 자연어 이해(NLU; Natural Language Understanding)를 중심으로 한 자연어 처리(NLP; Natural Language Processing) 기술이 자리잡고 있습니다.

자연어 처리는 인공지능이 인간의 언어를 이해하고 활용하는데 필요한 다양한 기술을 망라하는 개념입니다. 여기에 속하는 세부 기술로는 인텐트 분류(Intent Classification), 개체명 인식(Named Entity Recognition), 기계 번역(Machine Translation), 감성 분석(Sentiment Analysis) 등이 주로 언급됩니다.

스켈터랩스는 그동안 자연어 처리, 특히 자연어 이해 기술 개발을 위해 다양한 머신러닝, 딥러닝 기법을 활용해 왔고, 최근에는 BERT(Bidirectional Encoder Representations from Transformers)의 발빠른 도입으로 괄목할만한 성능 개선을 이루었습니다.

What is Intent Classification

인텐트 분류는 인간의 언어를 이해하고 인간과 대화하는 가상 에이전트(Virtual Agent) - 익숙한 사례로는 챗봇 - 를 개발할 때 근간이 되는 기술입니다. 문장이나 음성 등 사람이 하는 말(자연어)이 어떤 의도를 가지는지, 즉 어떤 인텐트에 속하는지 분류하는 것을 의미합니다. 예를 들어, 특정 산업 분야에서 나올 수 있는 대화의 인텐트 분류가 잘 이루어질 때 직원과의 실제 대면 없이도 높은 수준의 고객 지원 서비스가 가능해집니다. 다시 말해, 고성능 인텐트 분류 엔진의 도입으로 인간과 기계의 대화는 더욱 정교하고 자연스럽게 구현될 수 있습니다.

<그림1>의 스켈터랩스와 타사의 인텐트 분류 성능 결과를 살펴보면, 정확성을 판단하는 지표인 정밀도(Precision)와 재현율(Recall), 그리고 이를 복합적으로 고려한 F1 Score 모두 스켈터랩스가 타사 엔진에 비해 뛰어난 결과를 보입니다. 특히 600개 인텐트로 70% 이상의 스코어를 기록했다는 것은 실제 사람 간 대화에서 쓰이는 만큼 다양한 인텐트에 대처할 수 있는 인공지능임을 의미합니다. 이는 수많은 테스트 및 트레이닝 등 다양한 시도 끝에 스켈터랩스가 자체적으로 축적한 노하우와 개발 모델로 이루어낸 성과입니다.

600 Intents dataset performance comparison

단위: %, 시행일: 2019.06.28

Engine	Precision	Recall	F1 Score
Skelter Labs	64.64	65.30	64.97
글로벌 A 사	64.12	66.54	65.31
글로벌 B 사	64.86	65.72	65.28

*2019년 6월 21일 한국정보화진흥원(NIA)의 인공지능 데이터 허브 AIHub 웹사이트에 공개된 챗봇 훈련용 데이터 중 600 인텐트를 훈련시키고, 성능 평가 진행.

그림 1 600개 인텐트 분류 성능 비교 결과

The Process of Intent Classification

데이터 준비

높은 성능의 인텐트 분류를 구현하는 과정은 크게 세 단계로 구성됩니다. 먼저, 인공지능이 학습할 데이터를 준비합니다. 당연하게도, 데이터의 질이 좋을수록 인공지능의 성능은 높아집니다. 여기서 좋은 데이터란 무엇을 의미하는지, 그리고 사용자가 더 좋은 데이터를 준비할 수 있도록 스켈터랩스의 대화형 인공지능 제품 AIQ.TALK 이 어떠한 기능을 제공하는지를 'STEP1. 데이터 준비' 단계에서 알아보고자 합니다.

특징 추출

두 번째 단계는 특징 추출입니다. 인텐트 분류에 쓰이는 인공지능은 대부분 데이터를 그대로 학습하지 않고, 데이터에서 추출된 특징을 학습합니다. 데이터 원본을 무작정 학습시키는 것보다 데이터끼리 서로 구별되는 고유한 특징으로 학습을 진행하면 인공지능의 성능 개선에 더 효과적이기 때문입니다. 여기서 특징 추출이 잘 이루어지도록 특징 추출 알고리즘을 정교하게 설계하는 것이 중요합니다.

모델 학습

마지막으로는 모델 학습을 진행합니다. 인텐트 분류에 쓰이는 모델은 여러 종류가 있습니다. 높은 성능의 인텐트 분류 엔진을 만들기 위해서는 좋은 데이터를 준비하고, 적절한 특징을 추출하며, 이에 맞는 모델을 선택하고 튜닝하는 과정이 필요합니다. 자료의 'STEP3. 모델학습'을 통해 학습에 필요한 준비와 좋은 모델을 결정짓는 기준을 함께 설명하겠습니다.

STEP1.
데이터 준비

좋은 데이터란 무엇인가

데이터의 '질(Quality)'은 인공지능 성능을 위한 중요한 지표 중 하나입니다. 데이터의 질은 현재 목표가 무엇인지, 그 목표를 위해 어떤 방법이 쓰이는지 등 다양한 요소에 의해 평가됩니다.

인텐트 분류에 쓰이는 데이터는 자연어로 이루어진 문장이며, 각 문장이 어떤 인텐트에 해당하는지에 따라 라벨링(Labelling)이 이루어져 있습니다. '라벨링'은 머신러닝 알고리즘 중 '지도 학습(Supervised Learning)'에 속하는 알고리즘으로 풀고자 할 때 정답을 데이터에 연결해 주는 작업을 말합니다. 여기서 '지도 학습'이란 주어진 데이터에 대해 맞춰야 하는 어떤 정답이 이미 존재하는 과제에 대하여, 그 정답을 예측하는 인공지능을 만드는 알고리즘을 뜻합니다. 인텐트 분류는 대표적인 지도 학습 과제 중 하나입니다.

어떤 이가 당신에게 반가움을 표현하는 인사의 메시지를 보낸다고 가정합시다. 정확하게 '안녕'이라는 단어를 입력할 수도 있지만, '헬로우' 라던가 '하2'와 같은 변형어가 사용되기도 합니다.

그렇기 때문에 사람이 직접 '안녕'이라는 말의 인텐트는 '인사'라는 것을 인공지능에게 라벨링하면(즉, 지도 학습시키면), '하2'를 입력했을 때도 이것의 인텐트가 '인사'임을 추론할 수 있어야 합니다. 언어는 항상 다양한 형태로 변주되고 때로는 그 의미도 바뀔 수 있기에, 현재 인텐트 분류 기술의 목표는 다양하고 방대한 자연어가 어느 인텐트에 해당하는지를 가장 정답에 가깝게 추론하는 것이라고 볼 수 있습니다.

따라서 인텐트 분류 기술에서 말하는 '좋은 데이터'란, 하나의 인텐트를 뜻하는 다양한 문장으로 구성되어 있고, 각기 다른 인텐트는 최대한 비슷하지 않은 문장들로 구성된 데이터를 의미합니다. 예를 들어, '안녕'이라는 문장이 첫 인사 인텐트와 작별 인사 인텐트 모두에 들어있다면, 인공지능은 '안녕'이라는 말이 어느 인텐트에 해당하는지 정확히 추론할 수 없습니다. 이는 하나의 데이터를 2개 이상의 인텐트로 라벨링 하는 과정이 되어버립니다. 본 보고서에서는 빠른 이해를 위해 멀티 인텐트 라벨링을 제외하고 하나의 문장을 하나의 라벨로 분류하는 과정만을 다루었습니다.

다다익선(多多益善), 인공지능에게 학습시키는 데이터는 많을수록 좋습니다. 세상에 존재하는 모든 자연어를 이해하는 인공지능을 만들기란 아직 쉽지 않습니다. 따라서 우리는 인공지능이 어떤 주제의 대화, 어떤 종류의 문장을 이해할 수 있게 만들 것인지 목표를 정하고, 이에 따라 데이터를 준비합니다. 다시 말해, 모든 말을 이해하지는 못하더라도, 하나의 목표에 대해서는 잘 이해할 수 있도록 인공지능을 학습시킵니다.

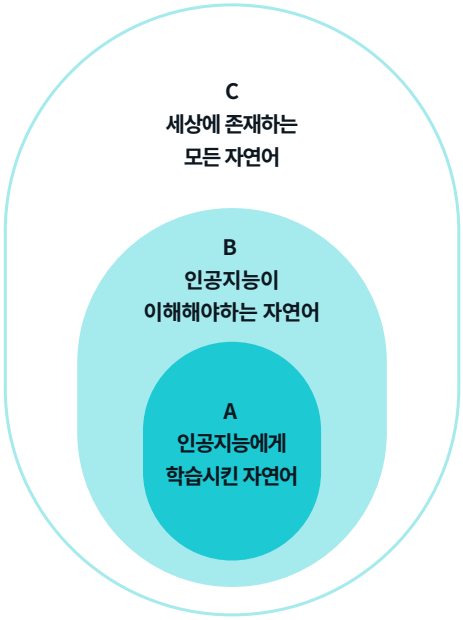


그림 2 인공지능 학습을 위한 자연어 분류

하나의 목표를 잘 이해하기 위해서는, 해당 주제에서 존재할 수 있는 모든 종류의 자연어 데이터를 준비하는 것이 이상적입니다. 하지만 이것은 현실적으로 어려운 일이므로 결국 우리가 인공지능에게 학습시키는 데이터는 해당 인공지능이 필수적으로 이해해야하는 자연어의 일부를 담게 됩니다. 그리고 인공지능이 데이터를 얼마나 효과적으로 학습하고 어떻게 예측 알고리즘을 만드느냐에 따라서, 학습하지 않은 데이터인 '인공지능이 이해해야 하는 자연어:<그림2>'의 영역은 더욱 넓어지게 됩니다.

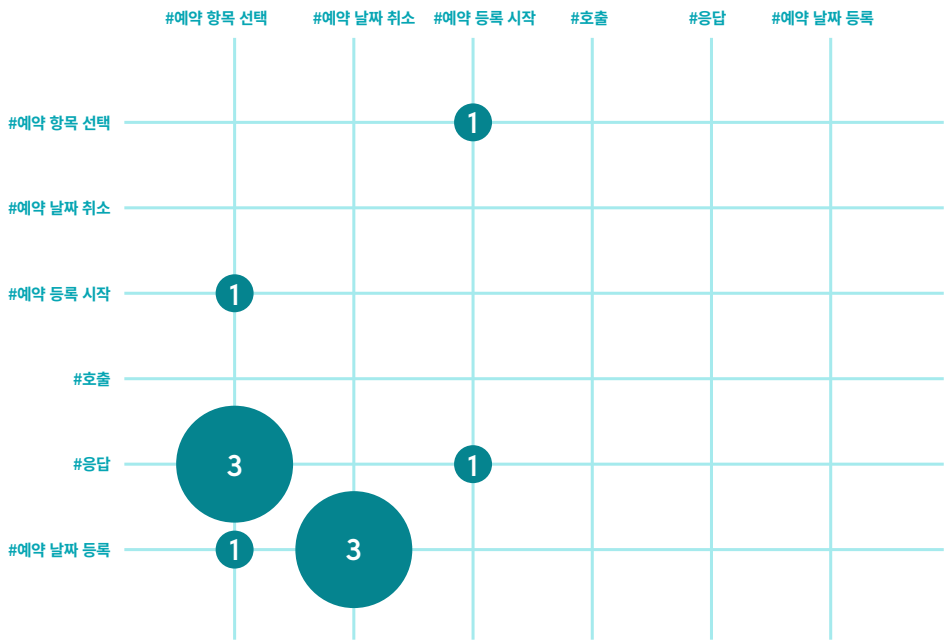


그림 3 스켈터랩스의 AIQ.TALK이 제공하는 '인텐트 인식률 평가하기' 기능의 평가 결과

어떻게 데이터의 품질을 판단하는가

좋은 데이터가 갖추어야 하는 성질을 이해하고 준비했다고 하더라도, 실제로 그 데이터가 좋은 데이터라는 보장은 없습니다. 스켈터랩스의 인텐트 분류 엔진은 사용자가 준비한 데이터의 질을 확인하고, 이를 바탕으로 데이터를 효과적으로 개선할 수 있도록 '인텐트 인식률 평가' 기능을 제공합니다.
스켈터랩스의 대화형 인공지능 엔진, AIQ.TALK의 인텐트 인식률 평가 기능은 주어진 인텐트와 테스트 데이터를 바탕으로 인텐트 오차 행렬을 분석해 사용자에게 직관적으로 정보를 전달합니다.

<그림3>은 인텐트 오차 행렬의 일부입니다. 행렬의 왼쪽에서 세로 방향으로 나열된 인텐트 목록은 정답 인텐트이고 행렬의 오른쪽 위에 나열된 인텐트 목록은 인식된 인텐트입니다. 동그라미 안에 있는 숫자는 잘못 인식된 인텐트 예문의 수이며 갯수가 클수록 동그라미 크기가 커집니다. 위 테스트 결과에서 3건의 오인식 동그라미가 2개인 것을 확인할 수 있습니다. 이 중 하나는 '응답' 인텐트로 인식되어야 하는데 '예약 항목 선택' 인텐트로 잘못 인식되었다는 의미입니다. 다른 하나는 '예약 날짜 등록' 인텐트로 인식되어야하지만 '예약 날짜 취소' 인텐트로 오인식 되었음을 뜻합니다.

즉, 동그라미가 크다는 것은 정답 인텐트와 잘못 인식된 인텐트의 사용자 예문이 서로 유사하며, 그에 따라 봇(Bot)이 인텐트를 잘못 인식할 확률이 높다는 뜻입니다. 이럴 경우 두 인텐트가 담는 의미도 비슷할 가능성이 높으므로, 두 인텐트를 하나의 인텐트로 합치는 것이 전체 인식률을 높이는데 도움이 될 수 있습니다. 혹은, 별도의 인텐트이지만 사용자 예문이 잘못 들어가 있어서 인텐트 오인식이 발생할 수도 있습니다. 이 경우, 동그라미가 큰 인텐트를 중심으로 사용자 예문을 검토할 필요가 있습니다.

인공지능 모델 학습에서 좋은 데이터를 판단하는 기준은 해당 모델의 성능입니다. 서로 다른 데이터로 모델을 학습시키며 성능 증감의 추이를 살펴보고, 가장 높은 성능을 달성하는 데이터를 준비해야 합니다. 스켈터랩스는 인텐트 인식률 평가 기능을 통해 사용자가 편리하게 학습 데이터의 질을 판단하고 빠르게 개선할 수 있도록 지원합니다.

STEP2.
특징 추출

우리가 준비한 데이터는 인텐트의 목록과 각 인텐트에 해당하는 여러 문장입니다. 이러한 데이터를 인공지능이 학습할 수 있는 형태로 바꾸는 작업을 특징 추출(Feature Extraction)로 통칭합니다. 특징 추출 과정을 거치는 덕분에, 인공지능은 데이터 원본을 그대로 분석하지 않고, 각 데이터의 두드러지는 '특징'을 기반으로 빠르고 효율적인 분석을 진행합니다.

스켈터랩스의 인텐트 분류 엔진은 특징 추출을 진행하기에 앞서 '형태소 분석'과 '엔티티 추출(Entity Annotation)', 두 단계를 거칩니다. 형태소 분석이란 주어진 문장을 한국어 형태소로 나누는 과정을 말합니다. 형태소는 주어진 문장에서 해당 형태소가 담는 의미와 역할을 보여주므로 형태소 분석 후 특징 추출을 진행하면 한국어 문법에 기반하여 보다 효율적인 추출이 가능합니다. 형태소 분석 후에는 엔티티 추출이 진행됩니다.

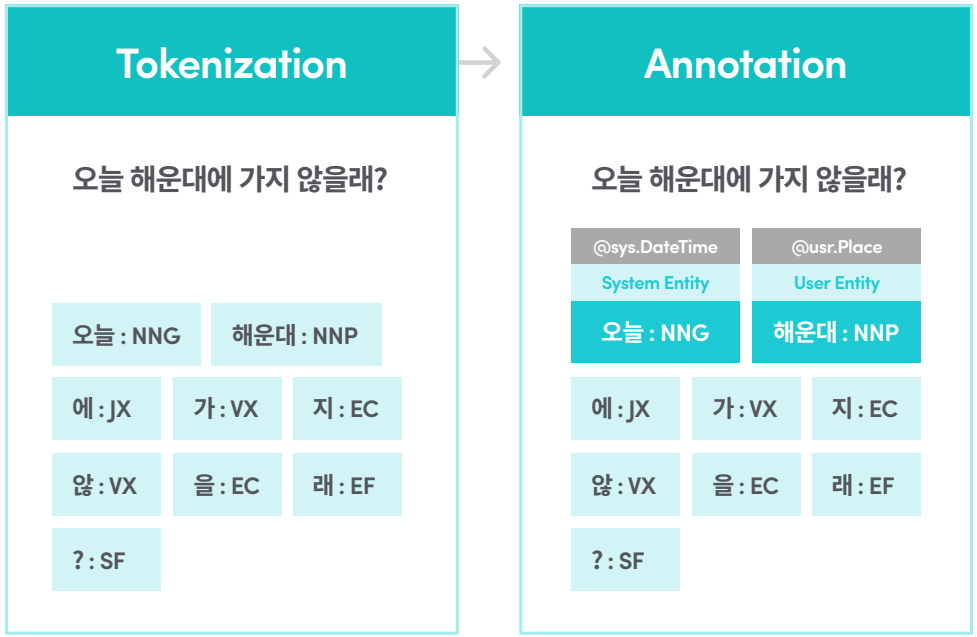


그림 4 형태소 분석과 엔티티(Entity) 추출 예시

엔티티는 문장을 구성하는 단어 중 공통점을 가지는 단어들을 묶어서 의미의 단위를 말합니다. 예를 들어 사용자는 '오늘 서울 날씨 어때?' 라고 물어볼 수도, '내일 부산 날씨 어때?'라고 입력할 수도 있습니다. 여기서 '오늘', '내일'은 날짜의 의미를 '서울', '부산'은 지역의 의미를 담습니다. 두 문장 모두 '[날짜] [지역] 날씨 어때?'라는 구조이고, 두 문장에 등장한 것 외에 날짜나 지역을 뜻하는 다른 단어로 대체해도 유사한 의도의 문장임을 알 수 있습니다.

엔티티가 추출(Annotation)되어 있으면 인텐트 분류 성능을 높일 수 있습니다. 분석해야 하는 문장에서 추출된 엔티티와 학습된 문장에서 추출된 엔티티를 대조하여 분류할 인텐트를 결정하는데 참고할 수 있기 때문입니다.

엔티티 추출까지 완료되면 그 결과 데이터를 토대로 특징 추출을 진행합니다.

자연어 이해 분야에서 사용되는 대표적인 특징 추출 방법으로는 TF-IDF, Word Embedding, Word2vec, Ngram등이 있습니다. 스켈터랩스 또한 다양한 기법을 조합하여 데이터의 특징을 추출하고 있습니다. 데이터를 대표하는 핵심적인 특징을 추출할수록 인공지능의 정확도는 점차 개선됩니다.

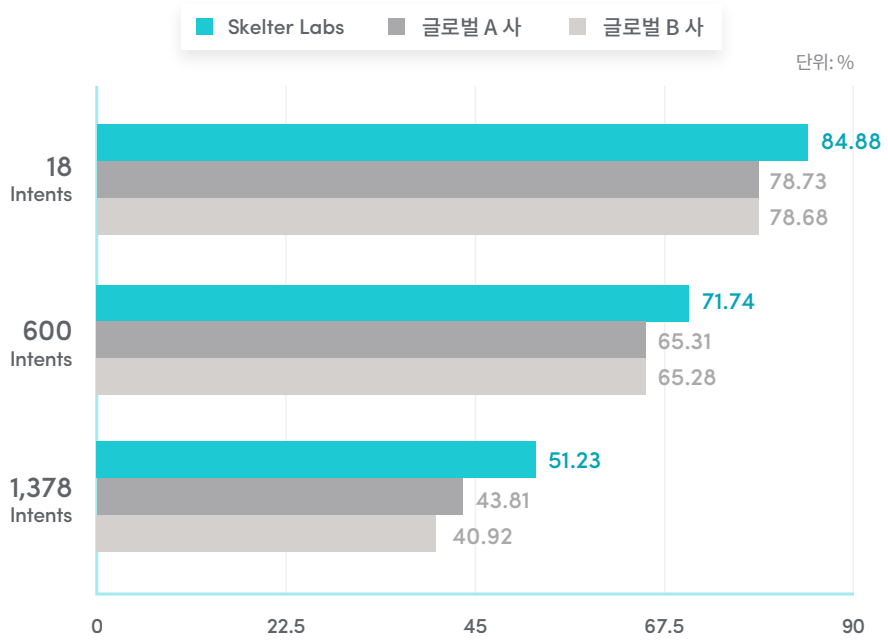
특징 추출 프로세스는, 인공지능이 도입되면서 <그림 5>의 B와 같은 형태로 바뀌었습니다. 인공지능이 도입되기 이전에는 데이터 분석에 쓰이는 모든 논리와 알고리즘을 사람이 직접 설계했습니다. 그러나 인공지능의 도입 이후, 이 과정을 인공지능에게 맡겨 자동화할 수 있게 되었습니다. 다양한 알고리즘을 통해 자연어 특징을 추출하면 이를 인공신경망 모델에 학습시킵니다. 이에 따라 A 방법보다 훨씬 방대한 데이터를 빠르고 효율적으로 분석할 수 있게 된 셈입니다.

그림 5 인공지능이 불러온 데이터 분석의 변화

STEP3.
모델학습

앞서 준비한 데이터는 일정한 비율로 쪼개져 한 쪽은 훈련용 데이터로, 다른 한 쪽은 테스트용 데이터로 활용됩니다. 훈련용 데이터만 인공지능에게 학습시킨 후, 인공지능이 학습한 적 없는 테스트용 데이터를 적용함으로써 정확도를 시험할 수 있기 때문입니다.

학습된 모델이 얼마나 뛰어난 성능을 갖추었는지는 정밀도와 재현율, 그리고 이 두 가지의 종합 값인 F1 Score 지표로 확인할 수 있습니다(각 지표의 의미 및 도출 방법은 아래 **[Appendix1. What is Precision, Recall and F1 Score]**에 기술되었습니다). 그리고 인공지능 모델의 성능이 뛰어날수록, 새로운 데이터에도 적절한 수용력을 갖춘 효과적인 인텐트 분류 엔진이 구현됩니다('적절한 수용력'에 관한 자세한 설명은 아래 **[Appendix2. What is 'Good' Intent Classification Model]**에 기술되었습니다).



*2019년 6월 21일 한국정보화진흥원(NIA)의 인공지능 데이터 허브 AIHub 웹사이트에 공개된 챗봇 훈련용 데이터 중 8문장 이상으로 구성되는 인텐트 중 샘플링한 18개 인텐트, 600개 인텐트, 1,378개 인텐트에 대해 분류 테스트 진행.

그림 6 18개, 600개, 1,378개 인텐트 분류에 대한 F1 Score 비교 분석

결국 모델 학습 과정은 적절한 수용력을 가지면서 정밀도, 재현율, F1 Score 등의 지표로 평가했을 때 우수한 성능을 보일 수 있는 최적의 모델을 찾는 지난한 여정이라고 볼 수 있습니다. 스켈터랩스는 지난 2년여간 최적의 모델을 개발하기 위한 실험을 진행해 왔습니다.

특히 최근에는 BERT 등의 기술을 한국어에 가장 알맞은 독자적인 방식으로 적용해 1,000여개가 넘는 인텐트 분류의 정확도 부분에서까지 유의미한 F1 Score를 확보했습니다.

현재 스켈터랩스의 인텐트 분류 엔진은 18개의 인텐트를 분류하는 과제에서는 84.88%, 600개 인텐트 분류의 과제에서는 71.74%, 마지막으로 1,378개의 인텐트 분류 과제에서는 51.23% 라는 F1 Score를 기록했습니다.

스켈터랩스의 미션은 다음과 같습니다. "언제 어디서나 우리의 일상을 이해하고, 도와주고, 더 나아지게 하는 머신 인텔리전스의 혁신을 이룬다" 우리는 당신의 언어를 제대로 '이해'하는 성공적인 첫 걸음을 디뎠고, 일상을 도와주기 위해 다양한 산업으로의 도입을 이어가고 있습니다.

기술 고도화로 보다 정확한 인텐트 분류와 인식이 가능할 때, 대화형 인공지능 기술의 성능 역시 높은 수준으로 향상됩니다. 현재 대화형 인공지능에 대해 쉽게 생각할 수 있는 적용 사례는 고객지원용 챗봇, 인공지능 스피커 정도가 떠오를 수 있지만, 앞으로 성능이 점차 향상될 수록 적용 분야는 거의 모든 산업으로 확대될 수 있습니다.

자동차와 같은 모빌리티 분야에서 드라이버, 승객에게 필요한 맞춤형 정보를 제공하는 것, 교육 분야에서의 개인 맞춤형 튜터링 서비스, 생활 밀착형 금융 자동화 등 무궁무진한 기회가 우리를 기다립니다.

글로벌 시장조사기관 가트너(Gartner)에 따르면, 2020년까지 전세계 85% 기업들이 실제 사람과의 접촉 없이도 고객 관리를 실행할 수 있을 것이라고 합니다. 스켈터랩스에서 대화형 인공지능 기술 개발을 리드하고 있는 조성진 CTO는 이렇게 말합니다. "사용자에게 실용적 '가치'를 제공할 수 있도록 대화형 인공지능이 고도화될 때, 사용자의 일상이 더욱 편리하고 윤택할 수 있도록 돕는 서비스가 비로소 탄생할 수 있습니다."

스켈터랩스는 인공지능이 사람의 역할을 대신하는 것이 아닌, 사람이 보다 창조적인 일에 시간을 사용하고, 원하는 일을 한층 효율적으로 할 수 있도록 돕고 싶습니다. 이런 철학으로 우리는 대화형 인공지능에 대한 도전을 멈추지 않을 것입니다.

Appendix 1

What is Presicion, Recall and F1 Score

정밀도(Precision), 재현율(Recall)과 F1 Score는 인공지능의 데이터 분류 성능을 평가하기 위해 가장 많이 쓰이는 지표입니다. 이를 사용하면 단순히 정확도(Accuracy)만으로는 구분하기 어려운 성능 차이를 보다 세밀하게 파악할 수 있습니다.

각 지표의 의미를 정확히 이해하기 위해, 인텐트 분류 문제를 단순화하여 주어진 문장이 '인삿말' 인텐트에 속하는가, 아닌가로 분류하는 문제(Binary Classification)를 예시로 설명하겠습니다.

		실제 정답	
		True	False
분류 결과	True	True Positive	False Positive
	False	False Negative	True Negative

그림 7 의도 분류 결과 해석을 위한 Confusion Matrix

<그림7>에서 볼 수 있듯, 인공지능의 데이터 분류 결과는 실제 정답과의 비교를 기준으로 총 4가지로 나뉩니다. True Positive(인삿말에 해당하는 문장을 인삿말로 분류한 경우)와 True Negative(인삿말에 해당하지 않는 문장을 인삿말이 아닌 것으로 분류한 경우)는 실제 정답에 맞게 분류했으므로 인공지능이 정답을 맞춘 케이스입니다. 반대로 False Positive(인삿말에 해당하지 않는 문장을 인삿말이라고 분류한 경우)와 False Negative(인삿말에 해당하는 문장을 인삿말이 아닌 것으로 분류한 경우)는 인공지능이 정답을 맞추지 못한 경우입니다.

이제 정답이 무엇인지(인삿말인지 아닌지) 알고 있는 문장들을 인공지능에게 분류시켜 각 문장이 위 4가지 경우 중 어디에 속하는지 갯수를 세면, 인공지능의 성능을 평가할 수 있게 됩니다. 예를 들어 ("안녕하세요", "인삿말"), ("내 말 좀 들어", "인삿말이 아님"), ("하이 방가방가", "인삿말"), ("수학 논문을 읽는 건 즐거운 일이다", "인삿말이 아님") 처럼 각 문장과 각 문장이 인삿말인지의 여부를 준비하고, 인공지능에게 문장을 분류시키고, 그 분류 결과가 실제 정답과 어떻게 같은지 / 다른지에 따라 4가지 경우로 나누는 것입니다.

정밀도와 재현율의 차이는 위의 네 가지의 경우를 어떤 관점으로 배분하느냐입니다. 먼저 정밀도는 다음과 같은 식으로 계산됩니다.

$$\text{Precision}_{\text{for true answer}} = \frac{\text{True Positive}}{\text{True Positive + False Positive}}$$

정밀도는 인공지능이 인삿말로 분류한 문장의 수 중에서 정답이 인삿말인 문장의 수의 비율을 말합니다. 이 인공지능이 어떤 문장을 인삿말이라고 분류했다면, 문장이 실제로 인삿말일 가능성이 얼마인지를 보여주는 지표인 셈입니다.

정밀도는 예측 결과 대비 실제 정답의 비율이므로, 분류 기준을 보수적으로 세울수록 더 높은 값으로 계산됩니다. 예를 들어 거의 모든 문장을 인삿말이 아닌 것으로 분류하고, 인삿말이 아닐 가능성이 거의 없는 극소수의 경우에 대해서만 - 예를 들어 인삿말일 가능성이 95% 이상인 경우에만 인삿말이라고 분류하는 인공지능의 성능을 평가한다면, 정밀도는 높을 수 밖에 없습니다. 실제로는 인삿말을 인삿말이 아닌 것으로 분류하는 약점이 있음에도 말입니다. 이런 이유로 정밀도 하나만으로 인공지능의 '정확도'를 판단하는 데에 한계가 있기에, 재현율이라는 새로운 지표도 성능 평가에서 함께 고려되어야 하는 것입니다.

$$\text{Recall}_{\text{for true label}} = \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

재현율을 계산하는 방법은 위와 같습니다. 재현율은 정답이 인삿말인 문장 중 인공지능이 인삿말로 분류한 문장의 수의 비율을 말합니다. 즉, 인삿말인 문장을 주었을 때 놓치지 않고 이를 인삿말로 분류해낼 가능성이 얼마인지를 보여주는 지표입니다. 정밀도가 모델의 예측 결과를 기준으로 성능을 판단한다면, 재현율은 실제 정답을 기준으로 성능을 판단합니다.

두 지표의 기준점이 다르므로, 인공지능의 분류 성능을 제대로 판단하려면 두 지표를 모두 고려해야 하며, 두 지표 모두 높을수록 뛰어난 인공지능입니다. 다만 두 지표는 성능 평가에 사용하는 데이터의 특성에 크게 영향을 받으므로, 두 지표를 치우침 없이 복합적으로 고려하기 위해 두 지표의 조화평균인 F1 Score를 사용하게 됩니다.

$$\text{F1 Score} = 2 \times \frac{1}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision + Recall}}$$

F1 Score는 모델의 성능을 단일 지표로 나타내기 때문에, 정밀도와 재현율 모두를 확인하는 것보다 즉각적이면서도 종합적인 성능 판단이 가능합니다.

이를 토대로 스켈터랩스와 타사 엔진의 인텐트 분류 성능 비교 평가 결과를 직관적으로 해석해 보자면, 600개 인텐트를 가지고 스켈터랩스의 대화 엔진이 분류한 의도가 실제와 맞을 확률은 72%가 넘으며, 스켈터랩스의 대화 엔진이 각각의 의도에 속하는 문장을 놓치지 않고 분류해낼 확률은 70% 에 달합니다. (소수점 이하 반올림 기준)

Appendix 2

What is 'Good' Intent Classification Model

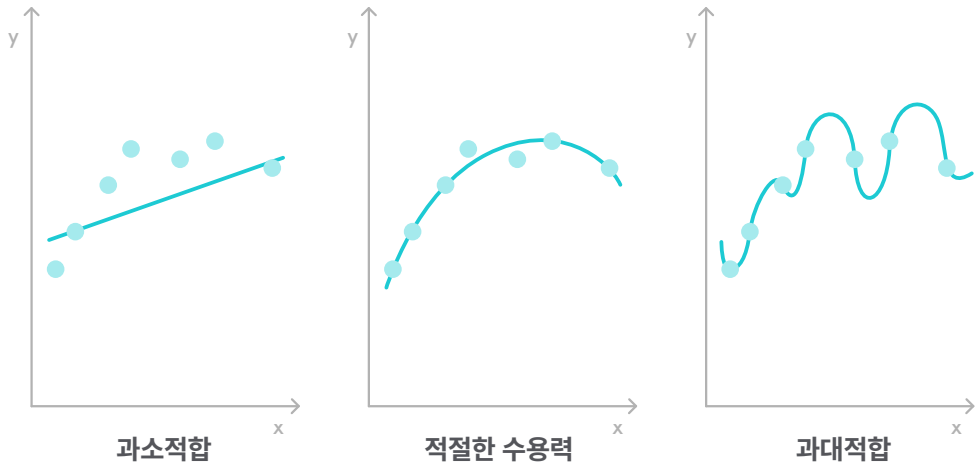


그림 8 적절한 수용력과 과소 및 과대 적합

여러 개의 점이 주어질 때 이 점들의 패턴을 가장 잘 표현하는 하나의 선을 긋는 문제를 예로 살펴보겠습니다. 점의 위치는 주어진 데이터이며, 점 간에 어떤 패턴이 숨어있는지 추론하여 수용력이 높은 선을 그리는 과제입니다. <그림 8>은 많은 점을 연결하기 위해 그은 세 개의 서로 다른 선을 보여줍니다. 가장 왼쪽의 예시는 대부분의 점이 그어진 선과 동떨어져 있으므로, 주어진 데이터를 잘 학습하지 못 했다고 볼 수 있습니다.

이러한 경우를 과소접합(Underfitting)이라고 합니다. 반대로, 가장 오른쪽의 경우 주어진 점들이 대부분 선에 걸려 있지만, 지나치게 점의 위치에 맞추어 선을 도출했습니다. 따라서 새로운 점이 주어졌을 때 그 점이 선 위에 걸리지 않을 가능성이 있습니다. 이처럼 인공지능이 주어진 데이터에만 지나치게 특화하여 학습하는 것을 과대적합(Overfitting)이라고 합니다.

효과적인 인공지능 학습 모델은 과대적합을 가능한 낮추고, 새로운 데이터에 대해서도 잘 예측하는 가운데 그림과 같은 모델입니다.

한편, 인텐트 분류와 같은 지도학습은 정답을 알고 있기 때문에 테스트 데이터에 대한 인공지능 예측 결과와 정답을 비교하여 인공지능의 성능을 판단합니다. 인공지능이 주어진 데이터를 적절히 학습해서 정답에 가깝게 예측해낼 수 있다면, 정답과 인공지능 예측치의 차이는 줄어듭니다. 이러한 인공지능 예측과 정답간의 차이를 '손실 함수'라고 표현합니다. 따라서 인공지능을 훈련시킨다는 것은 주어진 데이터를 사용하여 손실 함수를 최소화하는 모델을 구축하는 것을 말합니다.

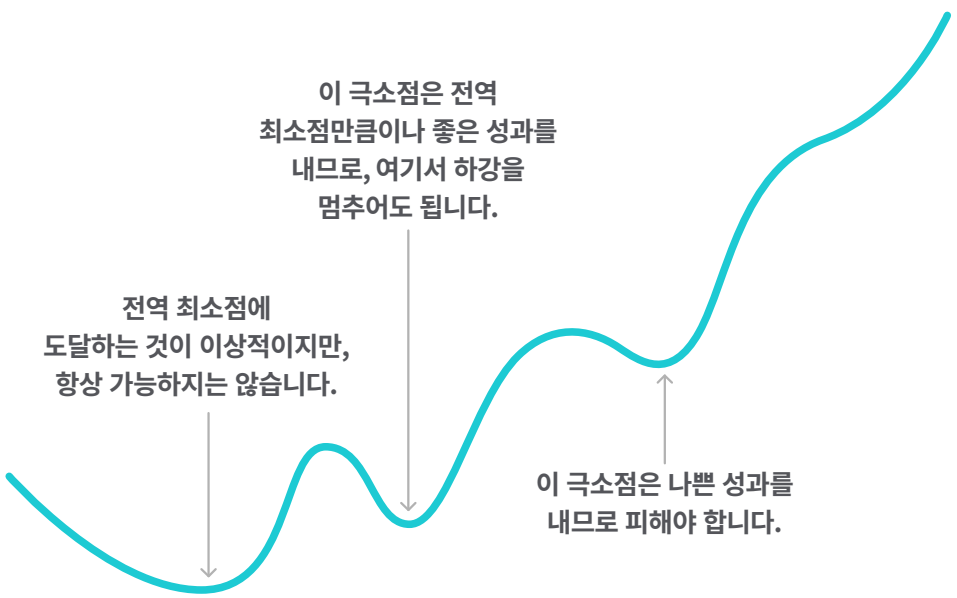


그림 9 손실함수의 최소화

<그림 9>는 손실함수를 최소화하는 과정을 쉽게 설명한 그림입니다. 그림처럼 비선형의 선으로 표현되는 함수에서 최저점의 x 값을 찾아가는 과제를 생각해 봅시다. 전체 함수를 통틀어 가장 값이 낮은 지점을 전역 최소점이라 하며, 전역 최소점만큼은 아니지만 함수의 값이 주변에 비해 낮은 지점을 극소점이라고 합니다. 손실 함수를 최소화하는 x를 찾기 위해 x의 값을 큰 값에서 작은 값으로 점점 하강시키는 전략을 따른다고 가정하겠습니다. <그림 9>의 가장 오른쪽 극소점을 만나면 x 값을 더 줄여도(하강하여도) 일정 구간 동안은 함수의 값이 오히려 증가합니다. 하지만 그 일정 구간을 지나고 나면 오히려 그 극소점보다 손실함수의 값이 더 작은 새로운 극소점을 만날 수 있습니다. 따라서, 손실 함수를 최소화하는 것은 전역 최소점에 도달하는 것을 지향하면서 충분히 함수를 최소화할 수 있는 적절한 극소점을 찾는 과정을 의미합니다.

종합하자면, 주어진 데이터에 대한 과적합을 피하면서, 배운 적 없는 새로운 데이터에서도 손실 함수를 최소화할 수 있는 인공지능 학습 모델을 구축해야 합니다. 이를 위해서는 앞서 언급한 좋은 데이터의 확보뿐만 아니라, 다양한 모델을 시도해보면서 가장 최적의 모델을 찾아나가는 과정이 필요합니다. 스켈터랩스 또한 학계 및 글로벌 기업이 선보이는 다양한 최신의 모델을 적용하며 인텐트 분류 성능을 지속적으로 개선시키고 있습니다.

See you soon!

본 보고서는 스켈터랩스의 인텐트 분류 엔진 성능에 대한 기본적인 이해를 돕기 위해 작성되었으며, 이를 통해 독자 여러분들의 대화형 인공지능에 대한 관심을 높이고자 합니다. 앞으로도 스켈터랩스의 인공지능 기반 대화 엔진을 완성하는 다양한 기술 컴퍼넌트에 대해 소개드릴 예정이니 지속적인 관심 부탁 드립니다.

AIQ.TALK STORY

vol.01 Intent Classfication

발행일 | 2019. 12.01
발행처 | 스켈터랩스
발행인 | 하미연, 변규홍, 스켈터랩스 마케팅 팀
디자인 | 이다은
메일 | contact@skelterlabs.com