인공지능 학습을 위한 자연어 데이터 가공단계의 품질검수 기준

박정하 TTA 데이터품질평가팀 책임

1. 머리말

인공지능 학습용 데이터의 구축 및 활용이 증가함에 따라 데이터 구축 과정을 기반으로 데이터의 생애주기(Life Cycle) 관점에서 데이터 품질관리가 필요하다. 인공지능 학습용 데이터 구축 공정 과정에서 확보된 품질은 학습데이터 전체의 품질을 결정한다. 특히 데이터 가공단계에서 기계학습에 활용할 수 있도록 기능이나 목적에 부합하는 라벨링 작업을 수행하는데, 인공지능 모델의 성능 평가에 중요한 영향을 미치는 참값(Ground Truth) 등의 라벨링 정보의 정확성이 확보되는 수준으로 가공 작업이 이루어져야 한다.

또한 인공지능 전문 연구 기관들의 데이터 가공 및 검수 기준 도입을 통해 데이터 품질수준을 한층 높은 수준으로 끌어올릴 수 있으나, 전문연구 기관들의 데이터 품질검증 기준은 연구 목적에 부합하도록 가공 및 검수의 기준을 정하고 있어 일반 다수 사용자들이 데이터를 활용하거나 적용하기에는 품질기준의 복잡도가 높다.

더구나 인공지능 학습용 데이터는 특정 인공지능 모델의 학습을 목적으로 생산되는 데이터이므로 임무 정의에 따라 구축되는 특성을 가지며, 그 특성은 가공 단계의 라벨링 작업을 통해 데이터에 반영된다. 따라서 라벨링의 의미적인 측면이 인공지능 모델의 임무 정의에 정확하게 부합하고 있는지 검증해야 하며, 이는 인공지능 학습을 위해 유효한 데이터와 더불어 인공지능 모델의 임무 수행에 대한 성능 결과와 직결된다.

본고에서는 한국어 음성 및 텍스트 데이터를 대상으로 인공지능의 기능과 목적에 부합하는 요구사항을 충족할 수 있도록 품질검증 항목을 중심으로 데이터 가공 단계에서의 품질검수 기준을 제시한다.

2. 인공지능 학습을 위한 자연어 데이터 가공단계의 품질기준

2.1 인공지능 학습용 데이터 생애주기별 품질지표

인공지능 학습용 데이터 구축 공정 과정은 구축계획수립, 데이터 획득/수집, 데이터 정제, 데이터 가공, 데이터 학습과 같이 5단계의 생애주기를 가진다. 각 생애주기별 품질검사 활동 및 품질지표는 <표 1>과 같다.

인공지능 모델의 학습 수준 및 평가 성능을 높이기 위하여 인공지능 학습용 데이터를 생성하는 구축 공정 과정별로 품질관리가 잘 이루어져야 한다.

그중 데이터 가공단계에서는 인공지능 학습에 필요한 어노테이션 작업을 수행하는데, 이때 어노

테이션 작업은 인공지능 서비스의 임무 및 모델의 기능과 목적에 부합하는 정보를 분석한 후 원천데이터에 태깅(tagging)하는 작업을 말한다. 따라서 학습에 요구되는 어노테이션 정보 설정 및 라벨링 기준은 데이터 품질검증의 기준이 되며, 특히 인공지능 모델의 정확한 임무 수행 및 성능 향상을 위하여 고품질의 라벨링 작업은 필수적이다.

<표 1> 생애주기별 품질검사 활동 및 품질지표

지표		생애주기(Life Cycle)				
		구축 계획 수립	데이터 획득/수집	데이터 정제	데이터 가공	데이터 학습
구축 공정	준비성	●	●	●	●	
	완전성		●	●	●	
	유용성					●
데이터 적합성	기준 적합성	●	●			
	기술 적합성	●	●	●	●	
	통계적 다양성	●	●			
데이터 정확성	의미 정확성				●	
	구문 정확성				●	
유효성	알고리즘 적정성					●
	학습 모델 유효성					●

출처: 인공지능 학습용 데이터 품질관리 가이드라인 및 구축 안내서 v3.0 – 제1권 품질관리 가이드라인 v3.0, 2023

2.2 자연어 데이터 학습 임무별 주요 품질검증 항목

인공지능 학습을 위한 자연어 데이터는 음성과 텍스트 데이터의 형태로 구분된다. 언어는 각 국가마다 다른 형태(figure)로 존재하고 인간 사회의 문화와 관계의 시간적 변화가 누적되어 사용되고 있으며, 또한 현재도 지속적으로 변화하고 있다. 그러한 변화는 언어를 사용하는 사람들의 음성과 텍스트를 통해 전달되며, 이를 통해 언어의 형태와 의미가 발전하고 있기 때문에 형태적 측면과 의미적 측면을 구분하여 자연어 데이터를 바라볼 필요가 있다.

음성데이터는 발화시 발생하는 음향정보와 발화 내용에 포함된 의미정보를 내포하고 있다. 음향정보는 음성신호(파형정보 등) 분석을 통해 언어를 인식하는데 필요한 정보이며, 음성을 통한 의미정보는 언어의 이해에 필요한 텍스트 정보추출을 위한 전사 정보이다.

인공지능 학습을 위한 음성 데이터를 구축하는 과정은 도메인과 화자 구성에 기반하여 발화 스타일에 따라 음성 발화 내용을 수집하고, 발화 정보를 전사 텍스트로 가공하는 과정이다. 이 과정에서 발생하는 음성데이터와 전사데이터에 대한 품질검증은 양질의 음성데이터 구축을 위해 필수적인 검증 대상이다.

텍스트 데이터는 지도학습을 위한 인공지능 서비스 임무에 따라 정제 및 가공되는 라벨링 대상이 명확하게 구분된다. 질의응답을 목적으로 하는 기계독해 데이터인지, 챗봇과 같은 대화를 끌어내려는 목적인지, 전문 도메인 분야의 요약, 번역과 같은 특수 목적을 수행하고자 함인지에 따라 가공 단계의 라벨링이 달라진다.

인공지능 학습을 위한 음성과 텍스트 데이터의 주요 임무를 중심으로 음성인식, 번역, 요약, 질의 응답, 대화, 주제분야 말뭉치로 구분할 때, 한국어 음성 및 텍스트 데이터의 인공지능 학습 임무별 주요 품질검증 항목은 <표 2>와 같다.

<표 2> 한국어 음성 및 텍스트 데이터의 인공지능 학습 임무별 주요 품질검증 항목

AI 학습 목적별 주요 품질검증 항목	음성	텍스트				
	음성 인식	번역	요약	질의응답	대화	주제 분야 말뭉치
메타데이터 분류 정확성	●					
전사 정확성	●					
주제 분류 정확성		●	●	●	●	●
번역 정확성		●				
요약 정확성	○		●	○	○	
질의 응답 정확성				●		
대화 의도 정확성	○			○	●	
NLP태깅 정확성		○				●
문법 정확성		●	●	○	○	○

● 주요 필수 검증항목 ○ 선택적 검증항목

2.2.1 음성인식 데이터

음성 인식 데이터는 메타데이터의 분류 정확성과 음성 전사의 정확성이 주요 품질검증 항목이다. 메타데이터는 화자구성, 발화스타일, 도메인, 발화환경 정보로 구성된다. 인공지능 학습을 위한 음성 데이터를 구축하는 과정은 계획된 도메인과 화자 구성에 기반하여 발화 스타일에 따라 음성 발화 내용을 수집하고, 발화 정보를 전사 텍스트로 가공하는 과정이다. 따라서 메타데이터의 분류가 정확하게 라벨링되었는지 그리고 전사 정확도를 검증하는 것이 가공단계의 주요 품질 검수 기준이 된다.

2.2.2 번역 데이터

번역 데이터는 텍스트 포맷의 원천 데이터와 번역된 데이터가 한 쌍으로 구성되어 있으며, 이 한 쌍의 데이터는 특정 도메인의 어휘 특성을 반영한 텍스트 데이터여야 한다. 따라서 해당 도메인에서 수집되어야 하며, 그 도메인 영역 내의 데이터인지 검증하기 위하여 주제 분류 정확성을 검증한다. 또한 해당 도메인의 용어 및 어휘로 번역이 정확하게 되었는지를 검증하기 위하여 번역 정확성을 검증한다. 또한 인공지능 학습 임무가 번역 문장을 정확하게 생성하는 것이므로 생성된 번역 문장의 문법 정확성을 검증하는 것이 가공단계의 주요 품질 기준이 된다. 구체적인 품질 검수 기준은 <표 3>과 같다.

2.2.3 요약 데이터

요약 데이터는 원천 데이터와 요약 데이터로 구성되어 있으며, 요약문은 추출요약문과 생성요약문 2가지 유형으로 구분된다. 인공지능 모델 학습시 입력 데이터가 원문과 추출요약문의 쌍인지, 원문과 생성요약문의 쌍인지에 따라 인공지능 모델의 요약문 생성 결과가 달라진다. 입력 데이터가 원문과 추출요약문의 쌍일 경우, 학습용 데이터는 추출요약문으로 구축해야 하며 추출요약문의 정확성을 검증해야 한다. 추출요약문은 원문의 문장 중, 핵심 정보를 반영한 문장이 정확하게 추출되었는지를 검수한다. 한편 입력 데이터가 원문과 생성요약문의 쌍일 경우, 학습용 데이터는 생성요약문으로 구축해야 하며 생성요약문의 정확성을 검증해야 한다. 생성요약문은 원문의 핵심 정보에 대한 문맥적 의미를 얼마나 정확하게 요약했는지 검수해야 하며, 원문의 문장과

는 비슷한 의미를 가졌으나 다른 어휘로 구성된 문장이어야 한다.
추출요약문과 생성요약문 데이터의 가공단계 품질 검수 기준은 <표 4>와 같다.

<표 3> 번역 데이터 품질검증 항목의 세부 검수 기준

검증 항목	품질검수 세부 항목			오류 유형
번역 정확성	형태	단어/문장	문법적 오류, 구두점 오류	•오탈자, 시제/조사 오류 등 문장부호 위치가 부적절한 경우
	의미	단어	오역, 누락, 직역	•오역: 잘못된 단어 표기 •직역: 사전적 의미 그대로 번역하는 경우
		문장	오역, 의미누락, 과한직역, 부자연스러움	•오역: 잘못된 문장 번역 •과한직역: 사전적 의미 그대로 번역되는 경우, 접속사 등을 불필요하게 사용하는 경우 등으로 원문의 의미를 제대로 전달하지 않는 경우 •부자연스러움: 전체 번역문의 의미가 논리적이지 않은 경우
	표현	문장	스타일, 유창성, 가독성	• 이해하기 쉽고 명확하며, 자연스러운 문체로 번역하였는지를 판단하는 척도로써, 원문이 축약형이거나 구어체 특성이 강하여 오역이 발생하는 경우, 복잡한 한국어 통사구조로 인하여 1:1 문장 번역 시 번역문의 가독성이 저하되는 경우

<표 4> 요약 데이터 품질검증 항목의 세부 검수 기준

검증항목	품질검수 세부 항목		오류 유형(예시)
추출요약 정확성	형태적	원문 내의 문장과 동일한가	구, 문장 일부 추출한 경우
		요약 형식(3문장 요약, 20%요약 등)을 준수하고 있는가	
		문법적 오류(오타, 오탈자, 띄어쓰기 등)를 포함하고 있는가	단, 문법적 오류까지 그대로 추출할지에 대한 여부는 수행기관과 협의
	의미적	핵심정보를 반영한 문장인가	주제와 관련 없는 문장인 경우
		추출한 문장들로 원문의 요약이 되었는가	의미가 비슷한 문장이 반복되는 경우
생성요약 정확성	형태적	원문 내의 문장과 다른가	원문내의 문장과 같은 경우
		문법적 오류(오타, 오탈자, 띄어쓰기, 비문 등)를 포함하고 있는가	
	의미적	원문의 핵심정보가 요약되었는가	•주제와 관련 없는 문장의 경우 •주관적 의견, 추측, 배경지식이 포함된 경우 • 단, 주제가 다수인 경우, 주요 주제로 요약하거나, 다수의 주제로 요약할지에 대해 수행기관과 협의
		추출요약 문장의 내용과 의미적으로 일치하는가(선택항목)	단, 추출요약 후 생성요약 데이터를 생성하는 경우에만 해당
		핵심키워드가 포함되었는가(선택항목)	단, 핵심키워드가 있는 경우에만 해당

2.2.4 질의응답 데이터

질의응답형 데이터는 원천 데이터와 질의응답데이터로 구성되어 있으며, 원천데이터는 지문 혹은 근거 단락 텍스트 데이터이다. 질의응답 데이터 품질수준을 높이기 위해서는 지문 혹은 근거 단락과 질의문장, 답변문장 간의 의미적 관계가 정확하게 매핑되어 구축되어야 한다. 가공단계에서는 원천데이터로부터 주어진 질의유형(예: 육하원칙)에 맞게 질의문장이 구축되었는지, 해당 질의에 대한 답변문장이 의미적으로 정확하게 생성되었는지를 검수해야 한다.

2.2.5 대화 데이터

대화 데이터는 다양한 형태로 수집되는 원천데이터를 통해 대화 형식의 데이터를 구축한다. 원천데이터는 대화 형식의 텍스트 데이터, 지문 형태의 텍스트 데이터, 이미지와 이미지의 설명문으로 구성된 데이터 등 여러 형태로 수집되는데, 이러한 원천데이터를 통해 대화 형식의 데이터를 구축하고 대화의 감정, 의도, 요약 등의 라벨링을 수행한다. 따라서 가공단계에서는 도메인별로 수집된 원천데이터를 통해 주제별 대화 내용이 적합한지를 검수해야 하며, 의도(intent), 감정(emotion), 감성(sensibility), 화행(speech act) 등과 같은 대화 목적이 정확하게 라벨링 되었는지를 검수해야 한다.

2.2.6 주제 분야 말뭉치 데이터

주제 분야 말뭉치 데이터는 특정 도메인에서 수집된 순수한 텍스트 데이터이다. 따라서 웹문서 형태의 텍스트부터 전문서적의 지문(단락) 형태의 텍스트까지 그 형태가 다양하다. 원천데이터는 텍스트와 텍스트 외 정보(그림, 표, 그래프 등)로 이루어져 있고, 텍스트 또한 구어체부터 문어체 등 다양한 어체를 포함한 데이터로 이루어져 있기 때문에, 순수한 말뭉치를 수집하기 위해서는 세밀한 가공기준이 요구된다. 예를 들어 텍스트와 연관된 그림, 표, 그래프의 경우 해당 설명 태그를 추가하여, 말뭉치 데이터를 구축할 때 텍스트 문장의 완전성을 검수해야 한다. 특히 구어체 데이터의 경우 특수기호를 누락하면 해당 어체의 특성이 소실되므로 특수기호, 한자, 영어 등의 말뭉치 포함 여부, 태깅 여부 등을 고려해야 한다. 한국어 말뭉치, 다국어 구어체 말뭉치 등 해당 말뭉치의 특성에 따라 기준을 설정하여 검수해야 한다. 또한 주제 분야 말뭉치 데이터는 NLP (Natural Language Process) 이해를 위한 태깅 작업을 선택적으로 수행하는데, NLP 이해를 위한 태깅 작업의 예시는 개체명(NER) 태깅, 품사(POS) 태깅 등이 있다. 따라서 주제 분야 말뭉치는 주제 태깅이 정확한지, NLP 태깅이 정확한지 그리고 띄어쓰기, 오탈자, 비문 등을 평가하는 문법적 정확성을 준수하였는지를 기준으로 가공 작업을 진행해야 한다.

3. 맺음말

인공지능 학습을 위한 데이터는 원천데이터의 수집에서부터 인공지능 모델 학습에 이르기까지 데이터 생애주기 전 단계에서 품질관리가 이루어져야 한다. 특히 가공 단계의 라벨링 작업은 인공지능 모델의 학습 성능에 큰 영향을 미치므로 음성 및 텍스트 데이터의 특성과 인공지능 모델의 임무를 고려하여 가공 기준과 검수 방법을 마련해야 한다. 최근 음성 및 텍스트 데이터와 동영상 데이터를 비롯한 이미지 데이터의 특성을 함께 추출하여 학습함으로써 인공지능 모델의 성능을 올리기 위한 멀티모달 형태의 데이터 구축이 많이 이루어지고 있다. 이에 멀티모달 데이터의 가공 및 검수를 위한 품질검증 기준이 준비되어야 할 것이다.

※ 본 연구는 '인공지능 학습용 데이터 구축 사업'의 일환으로 수행됨 (상세) 본 연구는 과학기술정보통신부 인공지능 학습용 데이터 구축 사업(2600-2602-305, 2023년 인공지능 학습용 데이터 구축 사업)에 의해서 수행되었음

[주요용어풀이]

- 음성 데이터: 음성을 추출하여 녹음한 데이터이며, 음성 녹음파일과 전사 텍스트파일을 포함함
- 텍스트 데이터: 언어 형태로 기록된 문자를 문장 또는 문단 단위로 구분하여 가공한 데이터
- 임무(Task): 인공지능 시스템이 제공하는 서비스의 목적에 따라 학습데이터가 갖추어야 하는 라벨링 작업
- 전사: 언어의 음성을 일정한 규칙에 근거해 문자로 표기한 것
- 음성 인식: 음성으로부터 언어적 의미 내용을 자동으로 식별하여, 인공지능이 발화 내용을 인식하는 서비스를 수행하는 임무
- 라벨링: 인공지능에 필요한 데이터 형식에 맞게 이미지, 영상, 음성, 비디오 등의 다양한 데이터 위에 목적에 맞는 라벨(주석)을 다는 작업
- 정확성: 전체 컨텍스트에서 정답 혹은 참값이 차지하는 비율을 의미하는 품질검증 항목
- 정확도: 전체 라벨링 데이터 수량 대비 정답 혹은 참값으로 분류한 수량을 비율로 계산한 지표

[참고문헌]

[1] 인공지능 학습용 데이터 품질관리 가이드라인 및 구축 안내서 v3.0 – 제1권 품질관리 가이드라인 v3.0, 2023.
[2] 한국어 음성 및 텍스트 데이터의 의미적 정확성 품질검증 방법, TTAK.KO-10.1419, 2023.
[3] 개체명 분석 및 개체 연결 말뭉치 연구 분석, 국어국립원, 2021
[4] 메신저 대화 자료 수집 및 말뭉치 구축, 국어국립원, 2019.
[5] 오픈 도메인 자연어 질의 응답을 위한 질문 분석 메타데이터, TTAK.KO-10.1098, 2018.
[6] 대화음성 인터페이스 기술 및 응용 서비스 개발 동향, 전자공학회지, 2014

※ 출처: TTA 저널 제209호