

메타데이터 정보 (다중기입가능)	분야	데이터 유형	구축 데이터량	원천데이터 형식	라벨링 형식	라벨링 유형
	재난안전환경	텍스트, 비디오	400,000건 (각 200,000건)	txt / mp4	json / xml	형태소/비수지 주석/키포인트
	데이터 출처	데이터 구축년도	구축기관(총괄)	가공기관	검수기관	
	한국어문장: 수집 및 증강 수어영상: 촬영	2021년	(주)테스트웍스	(주)테스트웍스 강남대학교 산학협력단	한국농아인협회	
	데이터 문의처	기관명	문의답자명	전화번호 (유선전화번호기입)	메일주소	
		(주) 테스트웍스	안상훈	02 422 5178	shahn@testw orks.co.kr	
	데이터 소개	수어 번역 데이터셋은 한국어를 한국수어로 변환하기 위한 AI 학습용 데이터로써, 한국어 문장과 매칭되는 한국수어 문법이 반영된 문장, 형태소, 동작에 대한 시간 정보 그리고 얼굴 표정 등과 같은 비수지 정보들을 포함하는 말뭉치 셋임				
	주요키워드	수어번역, 형태소/비수지 한국수어, 한국어-수어, 수어영상, 수어스크립트, 재난안 전문자, 날씨정보				
카테고리 정의서		첨부의 카테고리 정의서 엑셀파일에 데이터카테고리 작성하여 제출(예시참고)				

데이터셋명	국문영문	재난안전정보 수어영상 데이터																																																	
		Korean Sign Language AI Dataset for Disaster and Safety Infomation																																																	
구축목적	<ul style="list-style-type: none">재난 안전 정보가 포함된 원천 데이터(문장, 수어영상)로부터, 형태소 및 비수지 정보와 키포인트 데이터를 추출한 고품질 AI데이터셋 구축 및 배포고품질 AI데이터셋을 기반으로 한국어-한국수어 변환 AI알고리즘 개발 및 배포한국어-한국수어 변환 AI알고리즘을 사용한 실증서비스 구현 및 서비스 구현 가능성 확인																																																		
활용서비스	<ul style="list-style-type: none">본 과제 학습데이터를 활용한 산자부 과제를 통해 수어 번역 엔진의 추가 연구개발을 통해, 비수지, 지화, 수어 동작의 자연스러운 운율과 속도까지 구현할 수 있도록 활용본 과제를 통해 구축된 학습데이터 말뭉치 데이터베이스 구조와 성과 공유를 통해 다양한 사업 모델 파생국제공동연구 (함부르크, 드폴, 이지아 등)에 학습데이터 활용																																																		
소개	<p>- 수어 인공지능 학습을 위해 필요한 수어 데이터셋의 형태는 수어 인식을 위한 데이터셋과 수어 번역을 위한 데이터셋으로 나눌 수 있음</p> <div><div><p>2020년 NIA 수어 영상 과제</p><p>수어 입력 → 사람 관절/손/얼굴 키포인트 가공 → 수어 인식 AI 모델</p><p>* 교통 관련 4000개 단어, 2000개의 문장에 대한 인식용 수어 영상 데이터 구축</p><p>* 수요기관 활용 서비스</p><p>수어 아바타 생성 및 재생 아바타 관절/손/얼굴 키포인트 데이터 베이스 매칭</p></div><div><p>2021년 NIA 수어 영상 과제</p><p>오늘 미세먼지 주의하다</p><p>비수지(표정, 운율 등) 표현, 위치 변화 등 annotation & Time stamp</p><p>Translation</p><p>오늘 미세먼지가 많으니, 외출하시는 분들은 주의가 필요합니다.</p><p>* 20만개 이상의 재난안전발씨 관련 한국어 - 수어 번역 데이터 구축</p></div><div><p>데이터의 활용/확장성을 위한 키포인트 추가 가공</p><p>공간 정보 획득을 위한 모든 수어의 얼굴/몸 키포인트 가공 수행</p><p>높은 사용성/활용성/긴급성을 가진 수어의 수지 키포인트 가공 수행</p></div></div> <p><수어 관련 학습 데이터셋 구조></p>																																																		
	데이터셋 통계 (구축 규모 및 분포)	<p>1. 데이터 구축 규모</p> <table><tr><th>세부 과제</th><th>항목</th><th>성과지표</th><th>목표</th><th>실적</th><th>달성율</th></tr><tr><td rowspan="3">학습 데이터 수집</td><td>재난정보 관련 문장 데이터(TEXT)</td><td>수집건수</td><td>150,000건</td><td>164,375건</td><td>109.5%</td></tr><tr><td>문장을 한국수어로 촬영한 영상 데이터(mp4)</td><td>수집건수</td><td>200,000건</td><td>201,026건</td><td>100.5%</td></tr><tr><td>데이터 수집 가이드(pdf)</td><td>가이드 문서</td><td>1건</td><td>1건</td><td>100%</td></tr><tr><td rowspan="3">학습 데이터 가공</td><td>수어 영상 형태소/비수지 레이블링</td><td>레이블링 완료 건수</td><td>200,000건</td><td>201,026건</td><td>100.5%</td></tr><tr><td>수어 영상 키포인트 어노테이션(30 FPS)</td><td>키포인트 추출 완료 건수</td><td>200,000건</td><td>201,026건</td><td>100.5%</td></tr><tr><td>데이터 가공/검수 가이드(pdf)</td><td>가이드 문서</td><td>1건</td><td>1건</td><td>100%</td></tr><tr><td rowspan="2">모델 개발</td><td>한국-한국수어 변환 모델 개발</td><td>모델 개발완료 수</td><td>1건</td><td>1건</td><td>100%</td></tr><tr><td>한국-한국수어 변환 모델 성능</td><td>BLEU 스코어</td><td>15%</td><td>16.33%</td><td>108.9%</td></tr></table>		세부 과제	항목	성과지표	목표	실적	달성율	학습 데이터 수집	재난정보 관련 문장 데이터(TEXT)	수집건수	150,000건	164,375건	109.5%	문장을 한국수어로 촬영한 영상 데이터(mp4)	수집건수	200,000건	201,026건	100.5%	데이터 수집 가이드(pdf)	가이드 문서	1건	1건	100%	학습 데이터 가공	수어 영상 형태소/비수지 레이블링	레이블링 완료 건수	200,000건	201,026건	100.5%	수어 영상 키포인트 어노테이션(30 FPS)	키포인트 추출 완료 건수	200,000건	201,026건	100.5%	데이터 가공/검수 가이드(pdf)	가이드 문서	1건	1건	100%	모델 개발	한국-한국수어 변환 모델 개발	모델 개발완료 수	1건	1건	100%	한국-한국수어 변환 모델 성능	BLEU 스코어	15%	16.33%
세부 과제	항목	성과지표	목표	실적	달성율																																														
학습 데이터 수집	재난정보 관련 문장 데이터(TEXT)	수집건수	150,000건	164,375건	109.5%																																														
	문장을 한국수어로 촬영한 영상 데이터(mp4)	수집건수	200,000건	201,026건	100.5%																																														
	데이터 수집 가이드(pdf)	가이드 문서	1건	1건	100%																																														
학습 데이터 가공	수어 영상 형태소/비수지 레이블링	레이블링 완료 건수	200,000건	201,026건	100.5%																																														
	수어 영상 키포인트 어노테이션(30 FPS)	키포인트 추출 완료 건수	200,000건	201,026건	100.5%																																														
	데이터 가공/검수 가이드(pdf)	가이드 문서	1건	1건	100%																																														
모델 개발	한국-한국수어 변환 모델 개발	모델 개발완료 수	1건	1건	100%																																														
	한국-한국수어 변환 모델 성능	BLEU 스코어	15%	16.33%	108.9%																																														

2. 데이터 분포

한국어 문장 대 영상 비율		1:1		1:2		1:3		계	
		계획	실적	계획	실적	계획	실적	계획	실적
한국어문장 수(문장)		115,000	124,365	20,000	21,345	15,000	18,665	150,000	164,375
수어 영상(건)	소계	115,000	114,182	40,000	35,080	45,000	51,764	200,000	201,026
	대면촬영	40,000	40,780	40,000	35,080	45,000	51,764	125,000	127,624
	비대면촬영	75,000	73,402	-	-	-	-	75,000	73,402
수어 스크립트		115,000	114,182	40,000	35,080	45,000	51,764	200,000	201,026

대분류	중분류	소분류	한국어 문장		데이터 수량 (영상)	
			계획	실적	계획	실적
재해 · 재난 · 안전 (70%)	사회재난	16	45,000	46,949	60,000	60,380
	자연재난	18	48,000	59,588	64,000	66,277
	기타재난	9	12,000	12,838	16,000	14,825
생활정보 (30%)	날씨	1	45,000	45,000	60,000	59,544
계(100%)	-	44	150,000	164,375	200,000	201,026

○ 사회재난 유형

구 분		한국어 문장		수어 영상		비고
대분류 (New)	카테고리명	목표	실적	목표	실적	
사회재난	댐붕괴	100	110	100	89	
	원전 사고	100	110	100	100	
	항공기사고	100	100	100	100	
	건축물 붕괴	100	100	100	97	
	교통사고	6,000	6,119	8,000	7,781	
	철도, 지하철, 유도선 사고	6,000	6,653	8,000	7,992	
	미세먼지	500	533	700	694	
	가축질병	500	520	700	698	
	금융전산	500	547	700	743	
	식용수	1,000	1,174	1,400	1,373	
	산불	6,000	6,319	8,000	7,863	
	정전 및 전력부족	4,500	4,735	6,000	5,350	
	화재	6,000	6,233	8,000	7,794	
	폭발	5,500	5,503	7,300	8,699	
	전기 가스 사고	4,600	4,667	6,100	5,971	
	화학물질 사고	3,500	3,526	4,700	5,036	
합 계		45,000	46,949	60,000	60,380	

○ 자연재난 유형

구 분		한국어 문장		수어 영상		비고
대분류	카테고리명	목표	실적	목표	실적	
자연재난	낙뢰	100	100	100	81	
	화산폭발	100	100	100	182	
	산사태	4,000	4,092	5,500	6,212	
	대설	5,000	5,421	7,000	6,793	
	폭염	100	100	100	100	
	한파	5,000	5,390	7,000	6,531	
	황사	100	100	100	107	
	지진	100	105	100	100	
	지진해일	100	105	100	100	
	해수면상승	100	105	100	100	
	해일	300	326	100	285	
	가뭄	500	500	100	500	
	강풍	6,500	8,061	8,700	8,755	
	침수	3,500	4,000	4,700	5,072	
	태풍	6,500	8,102	8,700	8,609	
	풍랑	3,000	4,325	4,100	4,696	
	호우	6,500	10,177	8,700	9,212	
	홍수	6,500	8,479	8,700	8,842	
합 계		48,000	59,588	64,000	66,277	

○ 기타재난 유형


구 분		한국어문장		수어 영상		비고
대분류	카테고리명	목표	실적	목표	실적	
기타재난	산행 안전 사고	300	300	300	295	
	어린이 놀이시설 안전사고	200	200	200	200	
	승강기 안전 사고	200	200	200	200	
	실종 유괴 예방	300	300	300	298	
	응급처치	300	300	300	298	
	민방공 경보	300	300	300	209	
	여름철 물놀이	200	200	200	199	
	석유제품 사고	200	200	200	199	
	감염병예방	10,000	10,838	14,000	12,927	
합 계		12,000	12,838	16,000	14,825	

○ 생활정보 유형

구 분		한국어 문장		수어 영상		비고
대분류 (New)	카테고리명	목표	실적	목표	실적	
생활정보	날씨	45,000	45,000	60,000	59,544	
합 계		45,000	45,000	60,000	59,544	

데이터셋 구성

구분	내용
데이터명	재난 안전 정보 전달을 위한 수어영상 데이터
구축목적	청각 및 언어장애를 가진 사람들에게 전달 되어야 하는 재난 정보(재난 문자, 재난 방송 및 날씨 정보 등)를 수어로 전달하기 위한 인공지능 기술을 개발 목적의 한국어 문장 및 영상 데이터 수집 및 가공
라벨링 방법	신체 특정점 키포인트 가공, 형태소 및 비수지 가공
데이터 종류/형식	<ul style="list-style-type: none">원시데이터 : TXT(xlsx)원천 데이터: MP4라벨 데이터: .json, XML
클래스 수량	<ul style="list-style-type: none">한국어문장 카테고리 소분류 44종신체 특징점 : 몸통 25종, 얼굴 70종, 좌우 손 21 종(표제어에 대해서 선택적 가공)형태소 및 비수지 : 우세손 총렬, 비우세손 총렬 각 글로스명, 수지/지수어 구분, 수어 발화공간, 일치동사 정보, 시작시간, 종료시간, 대응 한국어 의미
데이터 예시	<ul style="list-style-type: none">원시 데이터- 한국어 문장(재난안전정보 문자) 데이터 예시 <div><div>sentence</div><div>현재 탄천 수위 7.5m / 한강 수위 4.2m 인천, 서울시 수위 5.5m 도달시 한강 수위 상승에 따른 반포, 신정원 진입로를 통제합니다. 현재 대연천 수위 상승으로 인하여 위험하오니 대연, 정연, 고연천, 연석지 등 하천 주변에 방류 중이오니 주민들께서는 하천 밖으로 임진강 댐붕괴 위험. 댐붕괴의 교량을 통제한 뒤 하류 수위 상승으로 댐붕괴가 끊기고 있으니 안전한 곳으로 대피하여 주시기 바 태화강 수위 급상승으로 인하여 연천군 수위 7.5m도 달천,연천,가북천,연천 수위 상승이 예상되니 하천 내 낚시 및 야영 등을 자제 임진강 수위 12.17:00~12:10 임진강 주의보 발령. 하천변 저지대, 침수 우려 지역 등 위험지역 주민은 안전한 곳으로 대피하시기 바 한강 수위 상승으로 인하여 반포, 신정원, 한신아파트 앞 도로 및 한강 수위 상승으로 보행자, 차량 및 이용객은 안전에 유의하시기 7월 3일 04:10 임진강 댐산 수위 3m 도달 경보 발령, 대피명령 및 경각심을 갖고 안전에 유의하시기 바랍니다. 한탄강 수위 급상승으로 인하여 연천군 10개 읍면에 발령된 주민 대피령을 해제하오니 주민께서는 연우 피해가 없도록 유의하시기 오늘 02:00 대전 홍수경보, 홍수주의보가 발령됨에 따라, 하천변 등 위험지역 절대 접근금지 등 안전에 주의 바랍니다. 오늘 20:00 정읍전 정읍시 홍수경보 발령, 홍수 피해 발생이 우려되는 지역주민은 안전에 유의하시기 바랍니다. 5월 17일 15:10 임진강 댐산 수위 3m 도달 경보 발령, 안전지대 대피, 차량 우회 등 피해에 유의하시기 바랍니다. 한탄강 수위 급상승으로 인하여 위험하오니 하천 주변 접근금지 및 하천 내 낚시 및 야영 등 금지하여 주시기 바랍니다. 04:16:50 현재 화국천 석창천 수위 상승으로 인하여 위험하오니 하천 접근금지 및 하천변의 야영객, 어민, 지역주민 등은 안전에 주의 바랍니다. 8월11일 01:40 금강하천 임진강 댐산 수위 3m 도달 경보 발령, 안전지대 대피, 차량 우회 등 피해에 유의하시기 바랍니다. 금일 집중호우로 인해 우리시 현재 황동강 둔지 범람이 우려되오니 황동강과 주변 저지대 주민들께서는 안전에 유의하시기 바랍니다. 09:10 임진강 댐산 수위 1m 이상 임진강 수위가 상승 중임 확인. 하천변의 야영객, 어민, 지역주민 등은 안전에 유의하시기 바랍니다. 오늘 09:30 원주전 유역 홍수경보 발령으로 하천 범람 등 피해 우려되므로 인근 지역 주민께서는 경각심을 갖고 안전에 유의하시기 금일 집중호우로 인해 낙동강 수위가 지속적으로 상승하고 있습니다. 둔지 내 차량 및 이용객은 안전한 지역으로 이동해 주시기 바 오늘 09:30 임진강 댐산 수위 3m 도달 경보 발령. 하천변의 야영객, 어민, 지역주민 등은 안전에 유의하시기 바랍니다. 한탄강 수위 급상승으로 인하여 하천 범람 위험성이 매우 높으므로 하천 주변 위험지역 주민께서는 안전에 각별히 주의하시기 바 금일 5:40 임진강 연천군 수위 9.8m 도달에 따른 경보 발령, 하천 주변에 계신 분들은 즉시 안전한 곳으로 대피하시기 바랍니다. 오늘 01:50 홍천강 댐산 수위 3m 도달 경보 발령, 하천 내 거주자 및 차량들은 안전한 곳으로 대피하시기 바랍니다. 6월7일 03:30 임진강 댐산 수위 3m 도달 경보 발령, 안전지대 대피, 차량 우회 등 피해에 유의하시기 바랍니다. 섬진강 곡성군 홍수경보 발령으로 많은 물 방류 및 집중호우에 따른 피해가 예상되오니 하천 주변 접근금지, 차량 이동, 야영 등 피 7월 17일 13:50 부로 구천동마을 안동말매식당 ~ 쌍송약국 구간 차량 통제가 해제되었음을 알려드리오니 참고하시기 바랍니다. 한탄강 수위 상승으로 인하여 정연리, 이길리 저지대, 침수 우려 지역 등 위험지역에서는 안전한 곳으로 대피하시기 바랍니다. 집중호우로 인하여 하천 범람 등 피해 위험이 있어 하천 주변 점검과 사전 점검을 철저히 해주시길 당부드립니다. 오늘 02:50 낙동강 부산시 홍수경보 발령, 낙동강 하천 범람 및 침수 예방을 위하여 하천 주변 논밭 관리행위 자제 등 안전에 주의 금일 집중호우로 인해 방정 지하차도 침수로 차량 통제[교통통제] 임을 알려드립니다. 7월 29일 11:00 백석역 인근 백두 한양아파트 앞 교통 통제 발생, 인근 주민은 안전사고 발생에 주의하시기 바랍니다. 이치가 한사 수위 4m 이상 홍수경보 발령, 하천변의 야영객, 어민, 지역주민 등은 안전에 유의하시기 바랍니다.</div></div>

구분	내용
	<p>- 원천데이터(수어영상 데이터) 예시</p>  <p>• 라벨링데이터</p> <p>- 형태소 / 비수지 데이터(json)</p> <pre> { "metadata":{ "id":"NIA_SL_G1_DELUGEFL000241_1_TW00.mp4", "signer":{ "id":"564321", "age":2, // 연령 "sex":1, // 성별 "deaf":1, // 농인/청인 여부 "main_communication":1, // 주된 의사소통 방법 "school_id":4, // 출신학교 아이디 "school_supplemental":""," // 출신학교 기타 텍스트 "sign_start_age":1, // 수어습득시기 "deaf_social_frequency":1, // 농인과의 정기적인 만남 여부 "occupation_id":3, // 직업 "translation_experience_category":2, // 한국어 - 한국수어 번역 경험 "residence":2// 거주 지역 }, "video_date":"2021-03-31", "video_fps":30, "hand_default":"right", "augment":true, "content_mismatch":{ "month_added": 8월 }, "filmed_in_studio":테스트웍스 촬영소 }, "korean_text":"8월 3일 04:10 임진강 횡산 수위 3m 도달 경보 발령, 대피명령 및 경각심을 갖고 안전에 유의하시기 바랍니다.", "sign_script":{ "sign_gestures_both":[{ "gloss_id":"8월", "express":"s", "position_strong":[], "position_weak":[], "direction":{ "source":""," "target":"" }, "start":0.82, "end":1.34, "sentence_loc":{ "start":""," "end":"" } }] } } </pre> <p>- 키포인트 데이터 (XML)</p>

구분		내용				
		<pre> <?xml version="1.0" encoding="UTF-8"?> - <annotations> <version>1.1</version> - <meta> - <task> <id>6084</id> <project>SL_TEST</project> <name>NIA_SL_G1_WINDWAVES000005_1_TW06_U.mp4</name> <size>742</size> <mode>interpolation</mode> <overlap>0</overlap> <bugtracker/> <flipped>False</flipped> <created>2021-08-09 13:34:59.918179+09:00</created> <updated>2021-08-09 13:39:23.407591+09:00</updated> <start_frame>0</start_frame> <stop_frame>741</stop_frame> <frame_filter>step=1</frame_filter> - <labels> - <label> <name>hand_right_keypoints_2d</name> <attributes> </attributes> </label> - <label> <name>hand_left_keypoints_2d</name> <attributes> </attributes> </label> - <label> <name>face_keypoints_2d</name> <attributes> </attributes> </label> - <label> <name>pose_keypoints_2d</name> <attributes> </attributes> </label> </labels> - <segments> - <segment> <id>5885</id> <start>0</start> <stop>741</stop> <url>http://220.85.41.207:8080/?id=5885</url> </segment> </segments> - <owner> <username>dm</username> <email>dm@testworks.co.kr</email> </owner> - <original_size> <width>1920</width> <height>1080</height> </original_size> . </pre>				
데이터셋 구축 수행기관 담당자	주관기관	기관명	책임자명	전화번호 (유선전화번호기입)	메일주소	담당업무
		(주)테스트웍스	안상훈	02 422 5178	shahn@testworks.co.kr	실무책임자
	참여기관	기관명	담당업무	기관명	담당업무	
		(주)이큐포울	마지웅 실무책임자(AI 모델 개발)	강남대학교	명혜진 실무책임자(형 태소 / 비수지 가공)	