

## 2021년도 ETRI 10대 대표성과 후보 추천서(요약)

성과Track	기초·미래선도	산업육성	국가·사회문제해결	
		○		
<b>협약(세부)과제명</b>	(엑소브레인-총괄/1세부) 휴먼 지식증강 서비스를 위한 지능진화형 WiseQA 플랫폼 기술 개발			
<b>과제번호</b>	<b>협약(세부) 과제번호</b>		<b>NTIS 과제번호</b>	
	20HS3200		1711103077	
<b>성과목표</b>	[1-1] 범용 인공지능 원천기술			
<b>총 연구기간</b>	2013년 5월 1일~ 2023년 2월 28일			
<b>총 연구비</b>	총 45,110 백만원		정부: 33,825 백만원 민간: 11,285 백만원	
	<b>연구자 성명</b>	<b>직할부서</b>	<b>연구본부/연구실</b>	<b>직위/직급</b>
<b>연구책임자</b>	임 준 호	인공지능연구소	지능정보연구본부 / 언어지능연구실	책임연구원
<b>성과 정보</b>				
<b>성과 내용</b>	<b>① 언어 이해에 기반하여 언어를 생성하는 최첨단 딥러닝 언어모델 기술 개발</b> - 언어이해(BERT)와 언어생성(GPT)의 동시 사전학습 기술 개발 (구글 배포 모델 대비 문장유사도 추론 7.12%, 기계독해 1.40%, 형태소 분석 1.43% 우수 평가) - KT 기술이전(2억원, '21.10.22)을 통하여 GPT-3 수준 초거대 규모 언어모델 학습 및 AI스피커, AI고객센터 대상 상용화 (진행 중)			
	<b>② 세계 최초 오피스문서 대상 심층질의응답 기술 개발 및 산업화</b> - 사내 규정, 메뉴얼, 온라인 공고 등 다양한 오피스문서로부터 사용자의 질문에 정답을 추론하는 심층질의응답 기술 확보 - 국가과학기술연구회, 공정거래위원회 등 공공분야 지능화(기술이전 3건, 기술료 4.8억원) 및 추가 실증테스트 2건 추진 중 (한글과컴퓨터, ETRI)			
	<b>③ 세계 최고 한국어 분석 기술 개발 및 산업화</b> - 소량 학습데이터 환경에서 도메인 전이가 가능한 메타러닝 기반 언어분석 기술 개발 (소량 학습데이터 개체명인식 평가 시 +2.36% 향상) - 기술출자/이전 및 상시 기술지원을 통한 마인즈랩 코스닥 상장 ('21.11.23) . 현물출자 지분 50% 매각을 통한 기술이전료 68억원 달성 (복합지능실 공동) - 은행, 보험, 제조, 법률, 공항, 온라인 광고 등 다양한 분야의 지능화 기술로 적용 및 기술이전을 통한 국내 산업 경쟁력 제고 . 기술이전 38건, 32.6억원 및 사업화 26건, 58.2억원 (최근 5년) 성과 달성			
	<b>④ 오픈API 기술 공개: 누적 4천9백7십만건 사용 ('21.10월 1백3십7만건 사용)</b>			
<b>대표성과 1</b>	- '21년 기술이전 9건 및 기술료 9.4억원 실적 . 한국어 이해생성 언어모델 사전학습 기술, 2억원, 2021년 10월 . 일반 문서 기반 심층 질의응답 기술, 1.8억원, 2021년 07월 . 법률 및 공정거래 분야 한국어 심층 질의응답 기술, 3.0억원, 2021년 01월, 11월 . 한국어 언어분석 및 언어모델 기술, 2억원, 2021년 06월 . 교육부 자소서 검증을 위한 한국어 언어분석 세부 기술, 0.2억원, 2021년 06월 . 고객센터 상담 분석을 위한 한국어 언어분석 세부 기술, 0.4억원, 2021년 10월			
<b>대표성과 2</b>	- 기술 공개 실적: 공공 인공지능 오픈API-데이터 서비스 현황(21년 10월 기준) . 공개 기술 누적 사용건수: 4천9백7십만건, '21.10월 1백3십7만건 사용			

# 1. 성과명

엑소브레인, 세계 최고 한국어 인공지능 기술로 AI산업을 견인하다!

# 2. 성과내용

## 기술개발 목표달성도

기술적 선점이 필요한 분야

○ 언어를 이해하고 지식을 학습하여, 전문가 수준의 지식을 서비스하는 언어 인공지능

기술개발 목표

○ 자연어로 기술된 텍스트의 의미를 이해하고 지식을 학습하여, 사용자의 전문분야 질문에 정답을 추론하여 제공하는 심층질의응답 기술

(목표 ①) 언어 이해에 기반하여 언어를 생성하는 최첨단 딥러닝 언어모델 기술 개발

(목표 ②) 세계 최초 실세계 오피스문서 대상 심층질의응답 기술 개발 및 산업화

(목표 ③) 세계 최고 한국어 분석 기술 개발 및 산업화

<기술개발 개념도>



기술개발 목표의 달성성과 및 핵심기술 확보

[개발목표 ①] 언어이해에 기반하여 언어를 생성하는 최첨단 딥러닝 언어모델 개발

➔ (달성성과) 언어이해(BERT)와 언어생성(GPT)의 동시 사전학습 기술 개발

➔ (핵심기술 확보)

- 대용량 원시 텍스트로부터 빈칸 단어열 맞추기(BERT 학습 유형)와 다음 단어

맞추기(GPT 학습 유형)를 동시에 학습하는 언어모델 사전학습 기술 개발

- 구글 배포 모델 대비 문장유사도 추론 7.12%, 기계독해 1.40%, 형태소 분석 1.43% 우수 평가
- KT 기술이전(2억원, '21.10.22)을 통하여 GPT-3 수준 초거대 규모 언어모델 학습 및 AI스피커, AI고객센터 대상 상용화 (진행 중)

### [개발목표 ②] 세계 최초 오피스문서 대상 심층질의응답 기술 개발 및 산업화

➔ (달성성과) 오피스문서 내 단락 및 테이블 대상 심층질의응답 기술 개발

➔ (핵심기술 확보)

- 사내 규정, 메뉴얼, 온라인 공고 등 다양한 오피스문서로부터 사용자의 질문에 정답을 포함한 단락 및 테이블 추론 기술
- 질문에 대한 다수의 복합 근거(evidence)를 인식하고, 다수의 복합 근거로부터 정답을 추론하는 설명가능한(explainable) 질의응답 기술 개발
- 국가과학기술연구회, 공정거래위원회 등 공공분야 지능화(기술이전 3건, 기술료 4.8억원) 기술로 도입 및 추가 실증테스트 2건 추진 중
  - ※ 연구행정 분야 (국가과학기술연구회, 1.5억원, '21.01) 및 공정거래 분야 (공정거래위원회, 1.5억원, '21.11)에 기술 도입, 씨엘씨이엔에스 기술이전(1.8억원, '21.07)을 통한 행정안전부 온나라 시스템에 차세대 딥러닝 기반 검색 시범사업 추진 중
  - ※ 한글과컴퓨터, 한컴 사내 규정문서 대상 실증 테스트 ('21.10~'21.12)
  - ※ ETRI, 원규정보 대상 심층질의응답 기술 실증 테스트 ('21.11~'21.12)

### [개발목표 ③] 세계 최고 한국어 분석 기술 개발 및 산업화

➔ (달성성과) 소량 학습데이터 기반 도메인 전이 메타러닝 언어분석 기술 개발

➔ (핵심기술 확보)

- 소량 학습데이터 환경에서 도메인 전이가 가능한 메타러닝 기반 언어분석 기술 개발
  - ※ 소량학습데이터 환경의 구어체 대상 개체명인식 평가 시, +2.36% 성능 향상 (개발 기술 93.14%, 기존 기술 90.78%)
- 은행, 보험, 제조, 법률, 공항, 온라인 광고 등 다양한 분야의 지능화 기술로 적용 및 기술이전을 통한 국내 산업 경쟁력 제고
  - ※ 마인즈랩, 솔트룩스, 미래에셋대우 등 기술이전 38건, 32.6억원 (최근 5년, 부가세포함)
  - ※ 기술이전 기업의 사업화 매출 26건, 58.2억원 (최근 5년, 기술기여도 반영 매출액 기준)
  - ※ 넥서스커뮤니티, 아이와즈, 큰사람커넥트 등에 기술이전('21년 4건, 1.6억원)을 통한 교육부 나이스 시스템 및 인공지능 고객센터 지능화
- 연구 결과물과 데이터는 산·학·연이 연구개발 용도 범위 내에서 자유롭게 활용할 수 있도록 오픈 API 방식으로 제공 (누적 사용건수: 4천9백7십만건, '21.10월 1백3십7만건 사용)

### 3. 우수성 및 차별성

#### 기술수준 향상 성과

- 한국어 이해생성 딥러닝 언어모델은 구글 배포 모델 대비 문장유사도 추론 7.12%, 기계독해 1.40%, 형태소 분석 1.43% 우수 평가

구분	[질의응답] 기계독해	[언어생성] 문서요약	[정보검색] 단락 순위화	[언어분석] 형태소분석	[문장의미 분석] 문장유사도추론
평가데이터 및 규격	KorQuAD v1.0 학습: 60,406건 평가: 5,773건 (dev셋)	AI Hub 요약 (뉴스-abstractive) 학습: 240,972건 평가: 30,121건(test셋)	학습: 45,521 질문 평가: 1,000 질문 (질문당 평균 8.7개 단락)	세종 말뭉치 학습: 135,238문장 평가: 4,800문장	학습: 41,465 문장쌍 평가: 3,181 문장쌍 (이진 분류체계: 유사, 무관)
평가방법	Exact Match <sup>[1]</sup> /F1 <sup>[2]</sup>	ROUGE-1/2/L <sup>[3]</sup>	Precision@Top1	F1	Accuracy
(KETI) KE-T5_ko base	91.35 / 85.40	49.15 / 21.97 / 46.78	81.8	N/A <sup>[4]</sup>	79.54
(Google) mT5 base	92.86 / 85.14	49.39 / 22.03 / 46.92	81.6	93.82	77.31
(엑소브레인) ET5 base	94.26 / 86.33	50.05 / 22.98 / 47.37	82.0	95.25	84.43

인공지능신문 | 2021.09.07.

사람처럼 문서 이해하는 인공지능 오픈소스로... ETRI, 한글 행정... 앞서 나온 행정문서QA API와 다른 한국어 AI 개발에도 쓰일 수 있는 원천 기술이다. 인공지능과 딥러닝 기술은 사람과 달리 문장이 조금만 달라져도 의미 관계를...

ETRI, 사람처럼 문서 이해하는 ... ZDNet Korea PICK | 2021.09.07. | 네이버뉴스  
ETRI, 사람처럼 문서 이해하는 AI 기술 개... 서울경제 | 2021.09.07. | 네이버뉴스  
ETRI, 사람처럼 문서 이해하는 AI 기술 개발 | 디트뉴스24 | 2021.09.07.  
사람처럼 문서 이해하는 AI(인공지능) 기술 개발 | 브레인미디어 | 2021.09.07.

관련뉴스 35건 전체보기 >

AI타임스 | 2021.09.08.

ETRI, 사람처럼 문서 찾는 AI 검색 기술 개발  
미니인터뷰(Mini Interview) ◆ API 2종 기술을 개발하게 된 연구 계기는 ETRI에서 행정문서QA와 패러프레이즈 인식 API를 개발한 목적은 널리 쓰이기 위함입니다....

ETRI, 사람처럼 문서 이해하는 AI 기술 개발 | 보안뉴스 | 2021.09.08.

데일리스포츠한국 | 2021.09.14.

ETRI, 사람처럼 문서 이해하고 정보 찾아주는 AI 기술 개발  
먼저 행정문서 질의응답(QA) API(응용프로그램개발환경) 기술은 딥러닝 언어모델을 이용해 단락과 표를 인식, 정답 및 근거 문장을 찾아준다. 예를 들어 '출장 경...




세계 최고 수준 대비 연구개발 수준

- 다수의 복합 근거를 인식하고 정답을 추론하는 질의응답 기술은 글로벌 최고 성능과 대비 99.3%로 대등한 수준의 성능 달성 (차년도에 글로벌 SOTA 대비 우위 달성 가능)
  - ➔ 글로벌 SOTA 성능은 80.52%이며, 엑소브레인 성능은 79.92% 성능 달성

Leaderboard (Fullwiki Setting)

In the *fullwiki* setting, a question-answering system must find the answer to a question in the scope of the entire Wikipedia. Similar to in the *distractor* setting, systems are evaluated on the accuracy of their answers (Ans) and the quality of the supporting facts they use to justify them (Sup).

	Model	Code	Ans		Sup		Joint	
			EM	F <sub>1</sub>	EM	F <sub>1</sub>	EM	F <sub>1</sub>
1 May 10, 2021	AISO (single model) <i>Institute of Computing Technology, Chinese Academy of Sciences</i> (Zhu, Pang et al., EMNLP 2021)		67.46	80.52	61.17	86.02	44.87	72.00
2 Feb 1, 2021	TPRR (single model) <i>Huawei Poisson Lab &amp; Parallel Distributed Computing Lab</i>		66.95	79.50	59.43	84.25	44.37	70.83



○ 적대적 예제에 강건한 패러프레이즈 인식 기술 OpenAPI 공개('21.09.01)



<엑소브레인 패러프레이즈 인식 OpenAPI 공개 및 적대적 예제 유형 별 평가 결과 ('21.09) >

#### 4. 성과의 활용도 및 파급효과

##### 경제 활성화 효과

##### 기업 경쟁력 향상

- 기술이전 및 출자를 통해 설립한 연구소 기업 마인즈랩(2014년 창업)은 이전 받은 기술로 'maum 오케스트라' 플랫폼을 출시하였으며, '21.11.23 코스닥 상장  
 ➔ 기업가치 약 1,600억~1,800억원 예상 (한국경제 '21.11.14.)
- OpenAI GPT-3 수준의 초거대 언어모델은 국내 대기업도 기술 추격에 어려움을 겪으나, 엑소브레인 과제에서 개발한 '한국어 이해생성 딥러닝 언어모델 학습 기술'을 기술이전('21.10)하여 언어모델 기술 경쟁력을 확보함  
 ➔ KT는 고성능 컴퓨팅 인프라를 구축하고, 기술이전 받은 언어모델 학습 기술을 이용하여 GPT-3 규모의 초거대 언어모델 학습 추진 중
- 한국어 분석 및 질의응답 기술은 마인즈랩, 아이와즈, 미래에셋대우, 하나금융티아이, 한글과컴퓨터 등 기술이전(35건, 기술료 29.9억원, 최근 5년)하여 국내 기업의 인공지능 기술 경쟁력 향상에 기여 (사업화 26건, 기술기여 매출액 58.2억원, 최근 5년)

< 엑소브레인 기술 사업화 주요 사례 >

기업	이전 기술	사업화 주요 사례
마인즈랩	한국어 분석 및 질의응답 기술	인천공항공사 스마트 항공통신 모니터링 시스템, 하나은행 금융상담 챗봇, 선진사료 AI주문봇, 한국폴리텍대학 교육용 대화 플랫폼 등의 사업화 실적 기반 <b>코스닥 상장 ('21.11)</b>
아이와즈	한국어 분석 기술	<b>교육부 나이스 시스템</b> 내 학교생활기록부 기재 내용 점검 시스템 개발 ('21.06)
KT	딥러닝 언어모델 학습 기술	엑소브레인 언어모델 학습 기술 이전('21.10)을 통하여 <b>GPT-3 수준 초거대 규모 언어모델</b> 학습 추진

씨엘씨이엔에스	질의응답 기술	오피스문서 대상 질의응답 기술 이전('21.07) 및 <b>행정안전부 온나라 시스템에 차세대 딥러닝 기반 검색</b> 시범사업 추진 중
데이터솔루션	한국어 분석 기술	개인정보비식별화 솔루션 개발로 2018 <b>SW기술대상(과학기술정보통신부 장관상) 수상('18)</b> 및 <b>법원 행정처</b> 적용 ('19)
미래에셋대우, 하나금융티아이	한국어 분석 기술	금융챗봇 및 금융기사 빅데이터 분석 기술 개발 ('19)
한글과컴퓨터	한국어 분석 및 질의응답 기술	한컴오피스 내 오피스톡 기반 지식검색서비스 솔루션 사업화 ('19)
인터웍스미디어	문맥 맞춤광고 기술	문맥 맞춤광고 기술이전('16.2)하여 조선, 중앙, 동아일보 등 온라인 언론사 82개사 대상 맞춤광고 사업화
마이다스아이티	한국어 분석 기술	AI 면접솔루션 사업화 ('19)

**산업 경쟁력 향상**

- **한국어 언어분석 OpenAPI 공개 및 한국어 언어모델 보급**을 통해 구글이 배포한 언어모델의 종속성 탈피하고, IBM 왓슨 및 구글의 한국어API 등 **의산 AI 상용 솔루션의 국내 시장 잠식 저지에 기여**
- 본 과제에서 개발한 한국어 언어분석, 질의응답 OpenAPI 및 한국어 KorBERT 언어모델을 통해 **국내의 기업, 연구소, 대학 등에서 다양한 응용개발을 추진 중**
  - ➔ OpenAPI 신청기관: 1,942개 기관
  - ➔ 누적 사용건수: 4천9백7십만건 ('21.10월 1백3십7만건 사용)
  - ➔ KorBERT 언어모델 배포: 1,153건
  - ➔ 한국어 이해생성 언어모델(ET5)은 '21년 11월 신규 배포 시작

**경제적 파급효과**

- 국내 언어 인공지능 시장에서 딥러닝 기술 보급, 기술이전 및 산업화 지원 등을 통해 향후 5년간 500억원 이상 매출액 창출에 기여 예상
  - ➔ 세계 인공지능 SW 시장 규모는 '20년 197억달러에서 '25년에는 1,058억달러 규모로 성장 전망되며, 자연어 처리 기술의 시장 규모는 '20년 22억2,000만달러에서 연평균 32.03%씩 성장해 '25년에는 89억1,600만달러 전망(출처: Tractica)
  - < 세계 인공지능 시장 전망 : 2020-2025 (단위: 백만달러) >

구분	'20년	'21년	'22년	'23년	'24년	'25년	Total	CAGR
기계 학습	3,567	5,332	7,648	10,422	13,416	16,306	56,691	35.52%
심층 학습	11,259	17,970	27,137	38,303	52,311	67,212	214,192	42.95%
컴퓨터 비전	1,903	2,932	4,081	5,133	6,415	7,440	27,904	31.35%
자연어 처리	2,222	3,189	4,448	5,925	7,469	8,916	32,169	32.03%
기계 추론	648	1,084	1,728	2,605	3,693	4,920	14,678	50.00%
강 인공지능	161	265	408	587	782	965	3,168	43.07%
Total	19,760	30,772	45,450	62,975	84,086	105,761	348,804	39.86%

- ➔ 국내 자연어 처리 기술의 시장 규모는 '20년 451억에서 연평균 30.53%씩 성장해 '25년에는 1,709억 전망(6년 누적 전망 6,289억원, 출처: Tractica)

< 국내 인공지능 시장 전망 : 2020-2025 (단위: 억원) >

	'20년	'21년	'22년	'23년	'24년	'25년	Total	CAGR
기계 학습	629	949	1,371	1,876	2,420	2,940	10,185	36.13%
심층 학습	2,712	4,276	6,365	8,841	12,088	15,558	49,840	41.82%
컴퓨터 비전	550	841	1,156	1,439	1,818	2,136	7,940	31.17%
자연어 처리	451	638	881	1,161	1,449	1,709	6,289	30.53%
기계 추론	138	240	392	601	860	1,156	3,387	52.97%
강 인공지능	23	38	58	84	112	138	453	43.10%
Total	4,503	6,982	10,223	14,002	18,747	23,637	78,094	39.32%

### 국가·사회적 파급효과

#### ○ 해결해야 할 국가·사회문제

- GPT-3 등 언어 인공지능에 대규모 투자를 추진 중인 선진국 대비 기술적 우위를 확보할 수 있는 언어 인공지능 기술 확보 필요
- 국내 인공지능 생태계 확대를 위하여 연구성과 공개 및 최신 연구주체의 인공지능 학습데이터 개방 필요

#### ○ 성과에서 개발된 기술적 솔루션

- 세계 최초 오피스문서 대상 심층 질의응답 기술 개발 및 OpenAPI 공개
- 언어 이해에 기반하여 언어를 생성하는 최첨단 딥러닝 언어모델 학습 기술 개발 및 기술이전을 통한 국내 대기업에 보급
- 세계 최고 한국어 분석, 질의응답 기술 개발 및 오픈API 서비스 보급
  - ※ API 누적 사용건수: 총 4천9백7십만건
  - ※ 다수의 복합 근거로 정답을 추론하는 질의응답 학습데이터 공개 6만건 ('20년 3만건, '21년 3만건)

#### ○ 국가·사회적 파급효과

- 인공지능에 대한 뜨거운 관심으로 인공지능 기술을 필요로 하는 기업들이 많아졌으며, 해당 기업들에게 기반이 되는 한국어 인공지능 기술을 제공하여 인공지능 생태계를 확대함
- 엑소브레인 과제 주관기관인 ETRI는 OpenAPI 서비스를, 공동연구기관인 마인즈랩은 하나은행 등 금융권 AI 챗봇 서비스를, 솔트룩스는 NH은행 및 행정안전부 등 AI 챗봇 기술을 통한 인공지능 대국민 서비스 제공