2019년 대표성과

[1] 기본 정보

<table>
<tr><td>후보추천 Track</td><td colspan="4">미래선도 (　　), 산업육성 (○), 국가·사회문제해결 (　　)</td></tr>
<tr><td>협약과제명 (협약과제번호)</td><td colspan="4">(엑소브레인-1세부) 휴먼 지식증강 서비스를 위한 지능진화형 WiseQA 플랫폼 기술 개발 (19HS3200)</td></tr>
<tr><td>총연구기간</td><td colspan="4">2013년 5월 ~ 2023년 2월</td></tr>
<tr><td>총연구비</td><td colspan="2">총 42,770 백만원</td><td colspan="2">정부: 32,070 백만원
민간: 10,700 백만원</td></tr>
<tr><td rowspan="2">성과책임자 정보</td><td>연구자 성명</td><td>직할부서</td><td>연구본부/연구실</td><td>직위/직급</td></tr>
<tr><td>김현기</td><td>인공지능연구소</td><td>지능정보연구본부/
언어지능연구실</td><td>책임연구원</td></tr>
<tr><td>연구사업계획서 관련 성과목표명</td><td colspan="4">[성과목표1-1] 스스로 보고 듣고 읽으며 성장하는 범용 인공지능 원천기술</td></tr>
</table>

[2] 2019년 우수성과 내용

1. 성과명

자연어 이해 기반 엑소브레인 심층질의응답 기술

2. 성과내용

기술개발 목표달성도

□ **기술적 선점이 필요한 분야**

○ 언어를 이해하고 지식을 학습하여, 전문가 수준의 지식을 서비스하는 언어지능

□ **기술개발 목표**

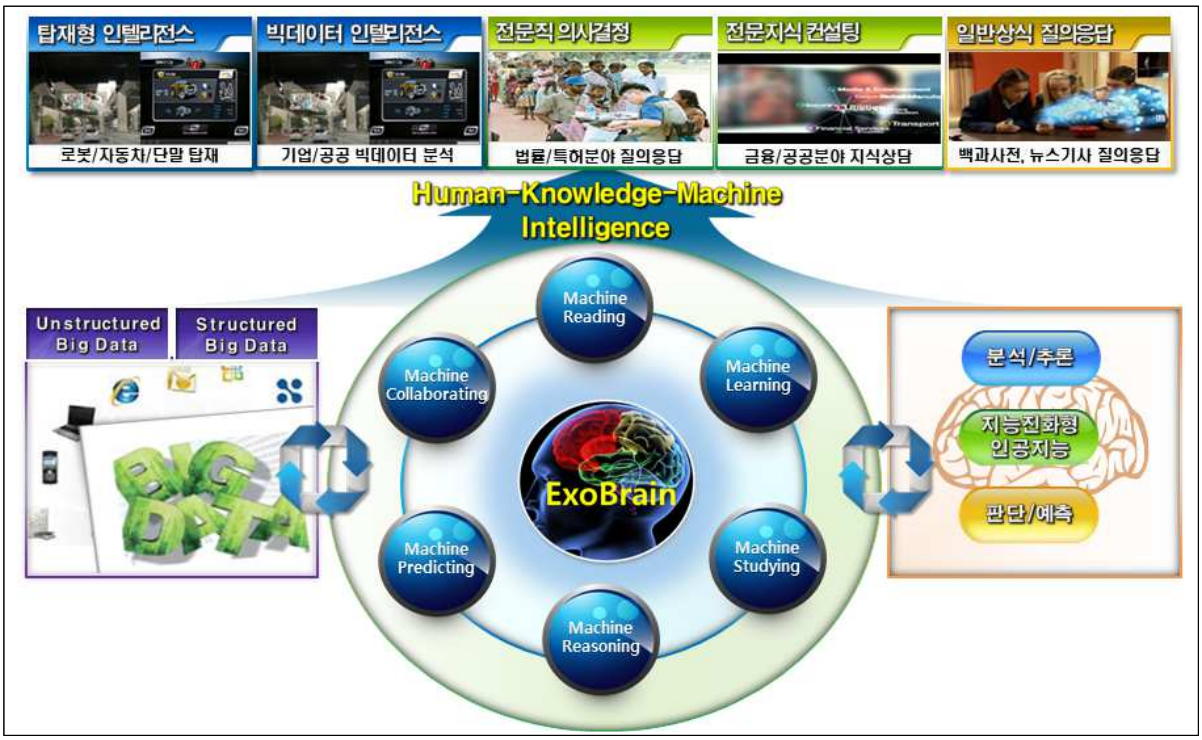
○ **자연어로 기술된 텍스트의 의미를 이해하고 지식을 학습하여, 사용자의 질문에 대한 정답을 추론하여 제공하는 심층질의응답 기술**

(목표 ①) 한국어 최첨단 딥러닝 언어모델 개발

(목표 ②) 세계 최초 단답형·서술형 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] 한국어 최첨단 딥러닝 언어모델 개발

➡ **(달성성과) 구글 언어모델 대비 4.5% 우수한 한국어 언어모델 개발**

➡ **(핵심기술 확보)**

- 한국어의 특징인 내용어(명사, 동사 등)와 기능어(조사, 어미 등)이 결합하여 어절을 구성하는 교착어를 의미단위인 형태소로 분리한 언어모델 설계
- 신문기사·백과사전 등 대용량 텍스트(23GB)를 대상으로 47억개의 형태소를 비지도 학습방법을 적용하여 한국어 최고 언어모델 'KorBERT' 개발('19.6)
- 5개 분야 한국어 응용 태스크 대상 **구글 언어모델 대비 4.5% 우수** (KorBERT 86.19% vs. 구글 언어모델 81.69%)

[개발목표 ②] 세계 최초 단답형·서술형 심층질의응답 기술 개발 및 산업화

➡ **(달성성과) 전이학습 기반 도메인 확장이 가능한 심층질의응답 기술 개발**

➡ **(핵심기술 확보)**

- 특정 도메인에 대한 종속성을 탈피하고, 특정 문제에 대한 과적합 방지를 위해 **자가 학습과 가중치 학습이 가능한 기계독해 기술 개발**

※ 자가 학습: 학습 시 동적으로 난이도를 상향시킨 신규문제를 생성하여 학습

※ 가중치 학습: 데이터 편향 및 과적합 방지를 위해 커리큘럼 학습방법 적용

- 일반상식과 법령지식 대상으로 **단답형 답변뿐 아니라 서술형 답변이 가능한 세계 최고 수준의 심층질의응답 기술 개발 성공**
  ※ 구글은 서술형 질의응답 기술 개발을 위해 Google Natural Questions 챌린지를 '19년 1월에 시작함
- **상식 심층질의응답 기술 산업화('19.10)**: **'한컴오피스 2020'에 지식검색으로 탑재**
  ※ ㈜한글과컴퓨터사는 일반상식 분야 문제를 대상으로 엑소브레인을 구글 지식그래프 검색과 비교한 결과, 엑소브레인이 10% 이상 높은 성능을 보였다고 평가함
- **법령 심층질의응답 기술 실증 성공('19.9)**: 국회도서관과 국가과학기술연구회로부터 실증테스트를 거쳐 우수성을 인정받아 **양 기관에서 ' 20년도에 상용화 예정**

[개발목표 ③] 세계 최고 한국어 분석 기술 개발 및 산업화

➡ **(달성성과) 기계학습 기반 세계 최고 한국어 분석 기술 6종 개발 및 보급**

➡ **(핵심기술 확보)**

- 기계학습을 적용하여 한국어에 최적화된 자연어 이해 기술 6종 개발(형태소 분석, 동음이의어 분석, 다의어 분석, 개체명 인식, 의존 구문분석, 의미역 인식 기술)
  ※ IBM 왓슨의 한국어 분석 기술(SK 에이브릴) 대비 성능우위 확보(개체명 인식 9%, 의미역 인식 24% 우위)
- 미래에셋대우, 하나금융티아이, 이큐포올, 어반데이터랩 등에 **기술이전을 통한 국내 산업 경쟁력 제고 (기술이전 7건, 기술료 7억1천만원)**
  ※ IBM 왓슨의 한국어 API(SK에서 에이브릴로 국내 사업화 중), 구글의 한국어 API 등 외산 상용 솔루션의 국내 시장 잠식 저지에 기여
- 연구 결과물과 데이터는 산·학·연이 연구개발 용도 범위내에서 자유롭게 활용할 수 있도록 **오픈 API 방식으로 제공 (누적 사용건수: 1천9백만건)**

3. 우수성 및 차별성

기술수준 향상 성과

○ 한국어 최고 딥러닝 언어모델 **KorBERT 개발 및 보급: 구글 대비 4.5% 우수**

한국말 잘 알아듣는 '코버트'…구글 '버트' 넘어섰다

<언론 보도기사(전자신문, 6.12)>

구분	의미역인식	기계독해	단락순위화	문장유사도추론	문서주제분류
평가데이터 및 규격	Korean Propbank, 학습: 19,302 문장 평가: 3,773 문장	KorQuAD 데이터, 학습: 60,406건 평가: 5,773건 (dev셋)	학습: 45,521 질문 평가: 1,000 질문 (질문당 평균 8.7개 단락)	학습: 10,874문장쌍 평가: 1,209문장쌍 (이진 분류체계: 유사, 무관)	학습: 9,301건 평가: 1,035건 (54개 분류체계)
평가 방법	F1[2]	Exact Match[3] / F1	Precision@Top1	Accuracy	Accuracy
(Google) Word Piece[4]기반 한국어 언어모델	81.85%	80.82% / 90.68% (정답 경계 구분을 위해 후처리 수행)	66.3%	79.4%	91.1%
(엑소브레인) Word Piece 기반 한국어 언어모델	85.10%	80.70% / 91.94% (정답 경계 구분을 위해 후처리 수행)	70.5%	82.7%	93.4%
(엑소브레인) 형태소 기반 한국어 언어모델	85.77%	86.40% / 94.18%	73.7%	83.4%	93.7%

<구글 모델 대비 5종 태스크의 성능 비교 결과>

○ 전 세계적으로도 단답형 정답을 추출하는 질의응답 기술 위주의 연구수준에서 나아가 **<u>범용적으로 활용가능한 단단형/서술형 질의응답 기술을 확보함</u>**

➡ 일반상식 심층질의응답 기술의 한컴오피스 탑재 및 오픈API 서비스 제공

➡ 법령지식 심층질의응답 기술은 수요기관(국회도서관, 국가과학기술연구회)의 실증 테스트를 통과하여 내년에 상용화 예정

➡ 미국의 글로벌 AI 검색 업체인 루시드웍스사에 한국어 분석 기술의 기술이전 협의 중

알파고? 우리에겐 엑소브레인…본격 상용화 시작

KBS PICK | 2019.11.01. | 네이버뉴스

지난 2016년 대전의 한국전자통신연구원(ETRI), 인간과 토종 인공지능(AI) 엑소브레인(Exobrain)의 첫 퀴즈대결이 열렸습니다. 인간을 대표한 4명의 출전자는 대학수학능력시험 만점자와 장학퀴즈 우승자 등 하나같이...

- ETRI, 한국어 최고 AI 기술 '엑소브레… 보안뉴스 | 2019.11.01.
- 한국어 최고 AI 기술 '엑소브레인' 상… YTN | 2019.11.01. | 네이버뉴스
- 서술형 문답 척척…토종 AI '엑소브레… KBS | 2019.11.01. | 네이버뉴스
- 한국어 최고 AI 기술 '엑소브레인' 상… YTN사이언스 | 2019.11.01.

관련뉴스 6건 전체보기>

ETRI, 한국어 AI '엑소브레인' 상용화 충청투데이 | 2019.10.31.

국내 연구진이 과학기술정보통신부와 정보통신기획평가원(IITP)이 추진하는 혁신성장동력 프로젝트인 '엑소브레인 사업'에서 개발한 최첨단 언어 인공지능(AI) 기술을 상용화 하는데 성공했다. 이로써 AI비서, 자연어...

- ETRI, 한국어 AI 기술 '엑소브레인' … 파이낸셜신문 | 2019.10.31.
- [리포트]지식검색 AI '엑소브레인' 상… 대전MBC | 2019.10.31.
- ETRI, 한국어 최고 AI 기술 '엑소브레… 아이티데일리 | 2019.10.31.
- ETRI, 한국어 최고 AI 기술 '엑소브레… 데이터넷 | 2019.10.31.

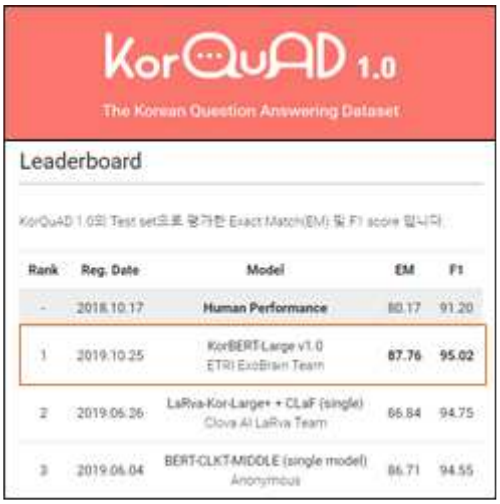
관련뉴스 28건 전체보기>

<엑소브레인 상용화 실적 및 계획 관련 보도기사('19.11)>

○ SCIE 저널 게재 2건, 국제특허 출원 2건, 국내특허 등록 8건

기술수준 공인 성과

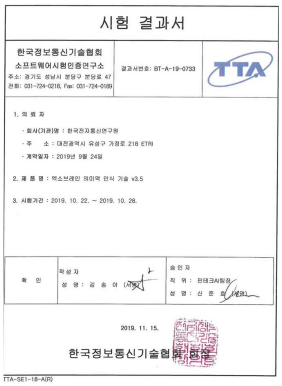
○ 한국어 기계독해 챌린지(LG CNS 주관 KorQuAD 1.0)에서 62개팀 중 1위(95.02점, ‘19.10~현재)



<KorQuAD 1.0 Leaderboard>

○ 세계적 언어처리 학회인 ACL WMT 2019 Shared Task on Quality Estimation(Task 1 Word Level) 3위(‘19.08)

○ TTA 공인시험성적에서 의미역인식(84.46% F1) 및 기계독해(94.77% F1) 성능 공인



<의미역인식 시험결과서>

<기계독해 시험결과서>

4. 성과의 활용도 및 파급효과

경제 활성화 효과

기업 경쟁력 향상

○ 한국어 분석 및 질의응답 기술의 미래에셋대우, 하나금융티아이, 마인즈랩, 한글과컴퓨터 등 기술이전(’19년도 기술료 7.1억원)과 **국내기업 산업화 지원을 통한 기업 경쟁력 향상에 기여(’19년도 기업 매출액 25억7천원 창출에 기여)**

➡ 미래에셋대우, 하나금융티아이: 한국어 분석 기술을 이전받아 금융챗봇 기술 개발 진행 중

➡ 마인즈랩: 한국어 분석 및 질의응답 기술을 이전받아 인천공항공사 스마트 항공통신 모니터링 시스템, 하나은행 금융상담 챗봇, 선진사료 AI주문봇, 한국폴리텍대학 교육용 대화 플랫폼, 국립국어원 학습데이터 구축 사업 등에 사업화

➡ 한글과컴퓨터: 한컴오피스 내 오피스톡 기능 중 지식검색서비스 솔루션 사업화

➡ 인터워스미디어: 문맥 맞춤광고 기술이전('16.2)하여 조선, 중앙, 동아일보 등 온라인 언론사 82개사 대상 맞춤광고 사업화

➡ 마이다스아이티: 형태소 분석기술을 이전받아 AI 면접솔루션 사업화

➡ 데이터솔루션: 개체명 인식기술을 이전받아 개인정보비식별화 솔루션 개발로 '2018 SW기술대상(과학기술정보통신부 장관상)' 수상 및 '19년도에 법원 행정처 적용 추진 중

❍ 기술이전·출자를 통해 설립한 연구소 기업 마인즈랩(2014년 창업)은 이전 받은 기술로 마음AI 플랫폼을 출시, 2019년도 기업가치를 930억으로 평가 받았으며, 2020년도 상장 추진 중(전자신문 2019.04.08.)

❍ 한국어 분석 기술이전을 통한 연구소 기업 어반데이터랩 설립('19.11)

사 업 자 등 록 증
(법인사업자)
등록번호 : 864-86-01386

법인명(단체명) : 주식회사 어반데이터랩
대 표 자 : 안지성
개 업 연 월 일 : 2019 년 11 월 08 일 법인등록번호 : 160111-0532104
사업장 소재지 : 대전광역시 유성구 가정로 218, 3층(가정동, 에트리융합기술연구생산센터)
본 점 소 재 지 : 대전광역시 유성구 가정로 218, 3층(가정동, 에트리융합기술연구생산센터)
사 업 의 종 류 : 업태 서비스업 종목 온라인플랫폼 개발
발 급 사 유 : 신규
사업자 단위 과세 적용사업자 여부 : 여( ) 부(v)
전자세금계산서 전용 전자우편주소 :
2019 년 11 월 08 일
북 대 전 세 무 서 장
국세청

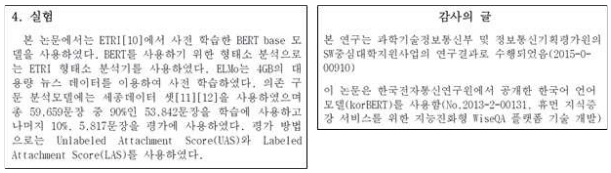
산업 경쟁력 향상

❍ 한국어 언어모델 보급을 통해 구글이 배포한 언어모델의 종속성 탈피하고, IBM 왓슨의 한국어 API(SK에서 에이브릴로 국내 사업화 중), 구글의 한국어 API 등 **외산 AI 상용 솔루션의 국내 시장 잠식 저지에 기여**

❍ 본 과제에서 개발한 한국어 언어모델 KorBERT, 오픈API 서비스를 통해, **국내의 기업, 연구소, 대학 등에서 다양한 응용개발을 추진**하고 있음

➡ 국립국어원 주관 국어 정보처리 시스템 경진대회에서 <u>KorBERT 활용하여 서강대는 의존구문분석 분야 대상 수상, KB금융지주는 일반분야 수상</u>

<국어 정보처리 시스템 경진대회>


4. 실험

본 논문에서는 ETRI[10]에서 사전 학습한 BERT base 모델을 사용하였다. BERT를 사용하기 위한 형태소 분석으로는 ETRI 형태소 분석기를 사용하였다. ELMo는 4GB의 대용량 뉴스 데이터를 이용하여 사전 학습하였다. 의존 구문 분석모델에는 세종데이터 셋[11][12]을 사용하였으며 총 59,659문장 중 90%인 53,842문장을 학습에 사용하고 나머지 10%, 5,817문장을 평가에 사용하였다. 평가 방법으로는 Unlabeled Attachment Score(UAS)와 Labeled Attachment Score(LAS)를 사용하였다.

감사의 글

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학지원사업의 연구결과로 수행되었음(2015-0-00910)

이 논문은 한국전자통신연구원에서 공개한 한국어 언어모델(korBERT)를 사용함(No.2013-2-00131, 휴먼 지식증강 서비스를 위한 지능진화형 WiseQA 플랫폼 기술 개발)

<서강대의 KorBERT 활용 의존구문분석>

<KB금융지주의 KorBERT 활용 챗봇>

➡ ETRI 인공지능 OpenAPI 활용사례 공모전에서는 엑소브레인 OpenAPI 및 KorBERT를 활용하여 금오공대에서 대상, 서울시챗봇팀에서 장려상, KB금융지주 및 경희대에서 가작 수상

- 금오공대: 식품 분석표 및 재료에 대한 텍스트에서 언어분석 API를 활용하여 성분과 재료에 대한 정보 추출, 추출한 단어에 대해 위키백과 QA API를 사용하여 상세한 정보 제공
- 서울시챗봇팀: KorBERT를 통해 언어 처리 모델을 대신하고, 형태소 분석 API를 통해 질문 문장에 대한 의도를 분석
- KB금융지주: KorBERT 모델을 활용하여 MRC 기반 금융QA챗봇 구현
- 경희대: 형태소 분석 API를 통해 주차장 안내 서비스 구현

경제적 파급효과

○ 국내 자연어 처리 시장에서 딥러닝 기술 보급, 기술이전 및 산업화 지원 등을 통해 향후 3년간 100억원 이상 매출액 창출에 기여 예상

➡ 세계 인공지능 SW 시장 규모는 '17년 54억달러에서 '25년에는 1,058억달러 규모로 성장 전망되며, 자연어 처리 기술의 시장 규모는 '17년 7억9,000만달러에서 연평균 35.4%씩 성장해 '25년에는 89억1,600만달러 전망(출처:

Tractica)

< 세계 인공지능 시장 전망 : 2017-2025 (단위: 백만달러) >

구분	'17년	'18년	'19년	'20년	'21년	'22년	'23년	'24년	'25년	Total	CAGR
기계 학습	882	1,451	2,308	3,567	5,332	7,648	10,422	13,416	16,306	61,333	44.0%
심층 학습	3,019	4,514	7,141	11,259	17,970	27,137	38,303	52,311	67,212	228,866	47.4%
컴퓨터 비젼	593	833	1,285	1,903	2,932	4,081	5,133	6,415	7,440	30,616	37.2%
자연어 처리	790	1,082	1,539	2,222	3,189	4,448	5,925	7,469	8,916	35,580	35.4%
기계 추론	96	200	370	648	1,084	1,728	2,605	3,693	4,920	15,344	63.4%
강 인공지능	18	46	92	161	265	408	587	782	965	3,324	64.5%
Total	5,398	8,126	12,735	19,760	30,772	45,450	62,975	84,086	105,761	375,063	45.0%

➡ 국내 자연어 처리 기술의 시장 규모는 '17년 168억에서 연평균 33.7%씩 성장해 '25년에는 1,709억 전망(Tractica, 3Q 2018)

< 국내 인공지능 시장 전망 : 2017-2025 (단위: 억원) >

	'17년	'18년	'19년	'20년	'21년	'22년	'23년	'24년	'25년	Total	CAGR
기계 학습	148	249	402	629	949	1,371	1,876	2,420	2,940	10,984	45.3%
심층 학습	819	1,160	1,773	2,712	4,276	6,365	8,841	12,088	15,558	53,594	44.5%
컴퓨터 비젼	188	253	381	550	841	1,156	1,439	1,818	2,136	8,762	35.5%
자연어 처리	168	226	317	451	638	881	1,161	1,449	1,709	6,999	33.7%
기계 추론	15	36	74	138	240	392	601	860	1,156	3,511	72.3%
강 인공지능	3	7	13	23	38	58	84	112	138	475	64.7%
Total	1,339	1,931	2,960	4,503	6,982	10,223	14,002	18,747	23,637	84,325	43.2%

국가·사회적 파급효과

○ **해결해야 할 국가·사회문제**

- **인공지능 기술 개발 및 산업화에 대규모 투자를 추진 중인 선진국 대비 기술우위를 선점할 수 있는 언어지능 기술 개발 필요**
- **국내 인공지능 연구 활성화 및 산업 경쟁력을 촉진하기 위해 연구성과 및 기계학습 데이터의 개방 필요**

○ **성과에서 개발된 기술적 솔루션**

- **구글(미국) 보다 우수한 최첨단 딥러닝 언어모델 기술(KorBERT) 개발 및 공개를 통한 인공지능 기술 국산화 기여 (2019.06~현재)**

※ KorBERT 공개 이전에는 국내기업이 구글에서 개발한 딥러닝 기술을 이용하여 언어 인공지능 서비스를 개발하였으나, 공개 이후 KorBERT 기술을 이용

※ KorBERT 언어모델 보급건수: 기업 및 대학에 총 355건 보급

- **세계 최고 한국어 분석 및 질의응답 기술 개발 및 오픈API 서비스 보급**

※ API 누적 사용건수: 총 1천9백만건

※ 기계학습 데이터 보급건수: 기업 및 대학에 총 427건 보급

○ **국가·사회적 파급효과**

- 알파고가 2016년 3월 이후에 촉발시킨 4차 산업혁명에 대비한 토종 AI 기술에 대한 대국민 관심도 제고 및 국내 산업화 수요 창출에 기여(‘19년 국내외 언론 보도 250여건)
- 토종 AI 기술의 우수성을 널리 홍보할 수 있는 계기가 되어 국내에서도 AI 산업화의 활성화 계기 창출, 과제 참여 기관인 마인즈랩, 솔트룩스 등은 하나은행, NH은행 등에 AI 챗봇 기술제공을 통한 대국민 서비스 제공