머신러닝 기술을 활용한
국회전자도서관 지식검색 서비스 모델 연구

2022. 06. 18.

국회도서관
NATIONAL ASSEMBLY LIBRARY

머신러닝 기술을 활용한 국회전자도서관 지식검색 서비스 모델 연구

2022. 06

■ 연구 기관 : 중앙대학교 산학협력단
■ 연구책임자 : 김학래 (중앙대학교 문헌정보학과 교수)
■ 연 구 원 : 김장원 (군산대학교 소프트웨어학과 교수)
■ 연구보조원 : 김현지 (중앙대학교 문헌정보학과 석사과정)
박하람 (중앙대학교 문헌정보학과 석사과정)
선은택 (중앙대학교 문헌정보학과 석사과정)
송채은 (중앙대학교 문헌정보학과 석사과정)

국회도서관
NATIONAL ASSEMBLY LIBRARY

제 출 문

국회도서관 귀하

본 보고서를 『머신러닝 기술을 활용한 국회전자도서관 지식검색 서비스 모델 연구』 과제의 최종보고서로 제출하며, 연구결과 공개에 동의합니다.

2022. 06

■ 연 구 기 관 : 중앙대학교 산학협력단
■ 연구책임자 : 김학래 (중앙대학교 문헌정보학과 교수)
■ 연 구 원 : 김장원 (군산대학교 소프트웨어학과 교수)
■ 연구보조원 : 김현지 (중앙대학교 문헌정보학과 석사과정)
박하람 (중앙대학교 문헌정보학과 석사과정)
선은택 (중앙대학교 문헌정보학과 석사과정)
송채은 (중앙대학교 문헌정보학과 석사과정)

중앙대학교 산학협력단 (인)

[목 차]

제 1 장 서론 ····· **1**
1. 연구 필요성 ····· 1
2. 연구 내용 ····· 2
3. 연구 방법 ····· 3

제 2 장 인공지능 기술 동향 ····· **5**
1. 배경 ····· 5
2. 인공지능 ····· 8
3. 발전 동향 ····· 12

제 3 장 자연어 처리 기술 동향 ····· **16**
1. 자연어 처리 ····· 16
2. 딥러닝 기반 자연어 처리 ····· 17
3. 한국어 자연어 처리 ····· 23
4. 응용서비스 ····· 27

제 4 장 국외 주요도서관 사례 ····· **39**
1. 개요 ····· 39
2. 미국 ····· 42
3. 일본 ····· 44
4. 프랑스 ····· 47
5. 스웨덴 ····· 50
6. 핀란드 ····· 52
7. 네덜란드 ····· 54
8. 노르웨이 ····· 57

제 5 장 국회전자도서관 서비스 현황 분석 ····· **58**
1. 개요 ····· 58
2. 국회 전자도서관 ····· 60
3. 국회법률도서관 ····· 63
4. 국회의원 정책자료 ····· 64
5. 국회 지방의회 의정정보시스템 ····· 65
6. 국회기록보존소 ····· 66

7. 전문가정보 데이터베이스 ····· 66
8. 지능형 의회정보 융합분석 시스템 Argos ····· 68
9. 국회회의록 빅데이터 ····· 74

제 6 장 국회전자도서관 도입을 위한 인공지능 서비스 모델 제안 ····· 80
1. 국외 도서관 서비스 벤치마킹 ····· 80
2. 서비스 도입을 위한 분석 프레임워크 ····· 85
3. 단기 과제 ····· 87
4. 중기 과제 ····· 88
5. 장기 과제 ····· 91
6. 요약 ····· 94

제 7 장 결론 ····· 97
1. 결과 요약 ····· 97
2. 시사점 ····· 99

[참고문헌] ····· 107

[표 목차]

[표 2-1] 주요국 디지털 정책 ····· 7
[표 2-2] 인공지능에 대한 4가지 관점 (Artificial Intelligence, 2020) ····· 9
[표 2-3] 과학기술일자리진흥원(COMPA) 인공지능 기술 분류 ····· 11
[표 3-1] 한국어 자연어 처리 기술 ····· 24
[표 3-2] 자연어 처리 응용서비스 분야 ····· 28
[표 4-1] 국외 주요도서관의 인공지능 서비스 사례 ····· 41
[표 5-1] 국회전자도서관 서비스 현황 ····· 59
[표 6-1] 국외 주요도서관의 서비스 분석 결과 ····· 81
[표 6-2] 국회전자도서관의 서비스 분석 결과 ····· 83
[표 6-3] 도입 가능한 인공지능 기반 서비스 목록 ····· 85
[표 6-4] 단기·중기·장기 과제의 개발 소요 시간과 서비스 모델 ····· 87
[표 6-5] 지능형 검색 서비스 제공을 위한 주요 기술 ····· 96

[그림 목차]

<그림 2-1> 산업혁명 주기 (4차산업혁명위원회, 2018) ····· 5
<그림 2-2> 인공지능 기술 분야 (Future Architecture, 2016) ····· 8
<그림 2-3> 인공지능의 하위 분야 ····· 10
<그림 2-4> 인공지능 트렌드 (LG 경제연구원 이승훈, 2017) ····· 12
<그림 2-5> 2021 Gartner Hype Cycle (Gartner, 2021) ····· 13
<그림 2-6> 인공지능 분야 주요 기술 (NIA, 2021) ····· 14
<그림 3-1> 자연어 처리의 하위 분야 ····· 16
<그림 3-2> 일반적인 CNN 아키텍처 ····· 17
<그림 3-3> LSTM 모델의 주요 계층 구조 ····· 18
<그림 3-4> 인코더와 디코더 아키텍처 ····· 19
<그림 3-5> 학습된 attention의 예시 ····· 19
<그림 3-6> 다층 퍼셉트론을 추가한 트랜스포머 아키텍쳐 ····· 20
<그림 3-7> BERT 파인튜닝 아키텍처 ····· 21
<그림 3-8> GPT-3를 사용한 달리(DALL·E)의 이미지 생성 예시 ····· 22
<그림 3-9> ERNIE 모델의 구조 ····· 23
<그림 3-10> KorBERT와 BERT 모델 비교 ····· 26
<그림 3-11> hyperCLOVA의 텍스트 생성 과정 ····· 27
<그림 3-12> 구글 지식베이스 QA(왼쪽)와 네이버 오픈 도메인 QA(오른쪽) ····· 29
<그림 3-13> 클로바 질의응답 과정 예시 (전동현, 2021) ····· 30
<그림 3-14> DrQA의 질의응답 과정 ····· 31
<그림 3-15> 연합뉴스의 세 줄 요약 예시 ····· 32
<그림 3-16> CLOVA OCR 예시 ····· 34
<그림 3-17> 대화형 인공지능 SIRI(왼쪽) 대화형 인공지능 구글 어시스턴트(오른쪽) · 35
<그림 3-18> Kakao I 적용 예시 ····· 36
<그림 3-19> 파파고의 기계번역 예시 ····· 37
<그림 3-20> 지식그래프 예시 ····· 37
<그림 4-1> Newspaper Navigator의 검색 결과 ····· 42
<그림 4-2> 오디오 결합 인터페이스 ····· 43
<그림 4-3> 키워드 검색 결과로 선택된 이미지 화면 ····· 45
<그림 4-4> 분류 결과와 정확도 ····· 46
<그림 4-5> 도서 검색 시각화 결과 ····· 47
<그림 4-6> GallicaPix 검색 화면 ····· 48
<그림 4-7> GallicaPix 구축 과정 ····· 49

<그림 4-8> GallicaPix 검색 결과 ····· 49
<그림 4-9> 광고 분류기 작동 과정 ····· 51
<그림 4-10> Finto AI 서비스 화면 ····· 53
<그림 4-11> 네덜란드 NTA(저자 이름 시소러스) ····· 54
<그림 4-12> Frame generator 서비스 화면 ····· 55
<그림 4-13> DBNL ngram viewer 검색 결과화면 ····· 56
<그림 4-14> 문서 수(왼쪽)와 DDC 자릿수(오른쪽)에 따른 정확도 차이 ····· 57
<그림 5-1> 국회도서관 소장자료 자동완성 검색 예시 ····· 60
<그림 5-2> 검색 결과 인포그래픽 예시 ····· 61
<그림 5-3> 지능형 질의응답 예시 ····· 62
<그림 5-4> 지능형 질의응답 엔진 흐름도 ····· 63
<그림 5-5> AI 일본법 번역기의 서비스 화면 ····· 64
<그림 5-6> 국회의원 정책자료의 '기후변화' 검색 결과 ····· 65
<그림 5-7> 국회 지방의회 의정정보 시스템의 인공지능 검색 결과 ····· 66
<그림 5-8> 전문가 상세보기에서 제공하는 그래프 데이터베이스 정보 ····· 67
<그림 5-9> 이슈 키워드 예시 ····· 68
<그림 5-10> 연관어 분석 결과 ····· 69
<그림 5-11> 비례대표 검색 결과 일부 ····· 70
<그림 5-12> 국회의원별 광장 예시 ····· 72
<그림 5-13> 비교분석 결과 예시 ····· 73
<그림 5-14> Argos 스토리보드와 연계한 국회회의록 데이터 분석 자료 ····· 74
<그림 5-15> 자동완성 예시 ····· 75
<그림 5-16> 발언자 탭 검색 결과 예시 ····· 75
<그림 5-17> 회의록 탭 검색 결과 예시 ····· 75
<그림 5-18> 국가학술정보 분석서비스와 연계한 데이터 분석 서비스 제공 기능 예시 ··· 76
<그림 6-1> 서비스 모델의 페이오프 매트릭스 현황 ····· 86
<그림 6-2> IBM 왓슨의 동영상 메타데이터 자동 생성 기능 예시 ····· 89
<그림 6-3> Argos 서비스: 초록 자동 추출을 제공하지 않음 ····· 90
<그림 6-4> 강기윤 국회의원 정보의 서로 다른 메타데이터 기술 ····· 91
<그림 6-5> 의미적으로 연계된 의안 정보 지식그래프 ····· 93
<그림 6-6> GovTracks.us의 H.R.2377 의안 처리 과정 ····· 94
<그림 6-7> 인공지능 기술 기반 지능형 검색 시스템을 위한 기능 중심 개요도 ····· 95

요 약

○ 본 연구는 머신러닝과 자연어 처리 기술을 활용한 국회전자도서관의 지식검색 서비스 모델을 제안하고 있음

○ 현재 미국, 중국, 일본, 한국 등 선진 주요국은 4차 산업혁명에 대응하기 위해 디지털 트랜스포메이션 정책을 추진하고, 인공지능 산업 생태계 구축에 투자하고 있음

○ 최근 인공신경망을 활용한 언어모델(예: 트랜스포머, BERT, GPT-3 등)의 비약적인 발전으로 딥러닝 기반의 자연어 처리 기술이 주목받고 있으며, 한국어 자연어 처리를 위한 대규모 언어모델이 개발됨

○ 머신러닝 기반 자연어 처리 기술을 적용한 서비스는 다양한 분야에서 활발히 제공되고 있으나, 도서관은 다른 분야에 비해 인공지능 서비스를 도입하지 못하고 있음. 도서관 서비스를 위한 인공지능과 자연어 처리 기술을 분석하고, 국회전자도서관에 적용 가능한 인공지능 서비스 모델을 구성해야 함

○ 2장과 3장은 인공지능과 자연어 처리 분야의 발전 동향과 최신 핵심 기술을 조사하고, 기술의 개념과 특징, 한계와 성과를 조사함. 또한, 국회전자도서관의 소장자료를 위해 한국어에 특화한 형태소 분석기 4개와 대규모 언어모델 5개를 선정해 분석함

○ 자연어 처리 기술을 적용한 서비스를 정보검색, 질의응답(QA), 문서 요약, 문서 분류, 이미지 인식, 음성 인식, 기계번역, 지식그래프 등 9개 서비스 분야로 나누고, 각 서비스의 제공 사례와 적용 기술을 구체적으로 제시함

○ 4장은 7개의 국외 주요도서관(미국, 일본, 프랑스, 스웨덴, 핀란드, 네덜란드, 노르웨이)을 대상으로 인공지능을 적용한 도서관 서비스 사례 19건을 조사하고, 공식 홈페이지와 논문, 코드 등을 대상으로 적용 기술과 알고리즘을 심층 분석함

- 국외 주요도서관이 머신러닝 기술을 적용해 서비스하는 분야 중 정보검색(7개)과 문서 분류(7개)의 사례가 가장 많음
- 정보검색은 미국, 일본, 프랑스, 네덜란드 도서관에서 7개 서비스를 제공하며, 그중 미국과 일본, 프랑스는 이미지 인식 기술을 활용한 이미지 검색 서비스를 제공함. 일본과 네덜란드는 지식그래프를 적용해 분산된 자료를 통합적으로 검색할 수 있는 서비스를 제공함

- 문서 분류 서비스는 일본, 스웨덴, 핀란드, 네덜란드, 노르웨이 총 5개 도서관에서 7개 서비스를 제공함. 일본과 노르웨이 도서관은 표준십진분류법에 머신러닝 기술을 적용한 자동 분류를 지원하고, 스웨덴은 OCR 기술을 통해 디지털화한 자료의 자동분류를 수행함. 핀란드와 네덜란드는 핀란드 도서관이 개발한 자동 주제 색인과 분류를 위한 오픈소스를 활용함

○ 5장은 국회도서관이 관리하는 8개 사이트(국회전자도서관, 국회법률도서관, 국회의원 정책자료, 국회·지방의회 의정정보시스템, 국회기록보존소, 전문가정보 데이터베이스, 지능형 의회정보 융합 분석 Argos, 국회회의록 빅데이터)를 대상으로 14개의 정보서비스 현황을 분석하여, 국회전자도서관의 서비스 현황과 기술 수준을 파악함

- 국회전자도서관의 8개 사이트에서 총 11개의 정보검색 서비스를 제공하고 있고, 문서 분류(6개), 지식그래프(5개) 서비스를 지원함
- 8개 사이트에서 총 11개의 정보검색 서비스를 제공하고 있으나, 키워드 기반 텍스트 검색에 한정되어 있고, 질의응답(QA)를 제공하는 서비스는 국회전자도서관의 지능형 질의응답이 유일함
- 문서 분류는 3개 사이트에서 6개 서비스에 적용되고 있으며, 보유하고 있는 소셜미디어 데이터를 분류하여 시각화하는 데 적용됨. 4개 서비스가 지식그래프 기반의 검색 기능을 제공하고 있으나, 통합적으로 관리되지 못하고 있음

○ 6장은 국외 주요도서관과 국회전자도서관의 현황을 바탕으로 벤치마킹할 수 있는 서비스 분야를 도출하고, 국회전자도서관에 도입 가능한 차세대 전자도서관 서비스 모델을 단기, 중기, 장기 과제로 구분해 제안함

- 서비스 모델의 도입 가능성과 기술 성숙도를 고려해 단기, 중기, 장기 과제의 단계별 도입 전략을 제안하고, 실제 도입을 위해 서비스의 도입 효과와 관련 기술, 해외 사례를 제시함

○ 7장은 연구 결과를 요약하고, 본 연구의 시사점을 바탕으로 국회전자도서관의 차세대 전자도서관 서비스 개발을 위한 향후 연구를 제안함

- 인공지능을 적용한 지식검색 서비스의 도입과 지속적인 운영을 위해서는 국회전자도서관의 데이터 거버넌스 구축에 관한 연구가 필요함

제 1 장 서론

1. 연구 필요성

□ 배경

○ 자연어 처리 기술은 텍스트로부터 의미를 이해하고 의미있는 정보를 추출하는 것으로, 최근 심층 머신러닝 기술의 발전에 따라 급격히 발전하고 있고 4차 산업혁명시대를 촉진하는 핵심 기술로 인식하고 있음

○ 인공지능 기술의 발전으로 광범위한 디지털 전환(Digital Transformation)이 이루어지고 있고, 미국·영국 등 주요국 정부는 인공지능과 데이터를 국가의 주요 정책으로 본격적으로 대응하고 있음

○ 국내는 디지털댐 프로젝트, 인공지능 학습데이터 구축 등 다양한 인공지능 정책이 추진되고 있으나, 개별 도메인에서 활용할 수 있는 데이터를 확보하고 기술적 완성도를 높이기 위한 노력이 필요함

- 예를 들어, 자연어 처리 분야는 심층 신경망 기반의 자연어 언어모델 기술이 급격하게 발전하고 있으나 한국어 처리를 위한 기술은 많은 연구가 필요함

○ 머신러닝 기반 자연어 질의응답, 가상 비서, 지식검색 시스템은 금융, 통신 분야에서 활발히 적용되고 있지만, 도서관 분야는 상대적으로 차세대 기술을 도입하는 데 한계가 있음

□ 필요성

○ 차세대 전자도서관을 위한 기술과 서비스 개발을 위한 전략 수립을 마련해야 함

- 주요 국가별 의회도서관에 적용된 자연어 질의응답, 본문 요약 등 최신 머신러닝 기술 동향 분석을 바탕으로 국회도서관에 적용 가능한 기술 수준을 검토하고 있음
- 최신 머신러닝, 자연어 처리 분야의 핵심 기술 동향과 특징을 상세히 분석하고, 도서관 분야에 적용 가능한 머신러닝 연구와 적용 방안을 제시함

○ 국회의원, 국민 누구나 쉽고 편하게 사용할 수 있는 차세대 전자도서관 서비스를 도출함

- 머신러닝·자연어 질의응답 기술을 도서관 분야에 적용하기 위한 타당성을 검토함

- 한국어 중심의 자연어 질의응답, 본문 요약 등 머신러닝 기반 알고리즘에 대한 동향 분석을 바탕으로 실용적인 서비스 모델을 구성하고, 성능 향상을 위한 구체적인 방안을 도출함

□ 기대효과

○ 국회도서관의 차세대 국가지식정보서비스와 입법지원서비스의 개발 방향성을 제시함

- 머신러닝 분야의 최신 연구 동향과 국내외 주요 의회도서관의 적용 사례 검토를 통해 차세대 국회전자도서관 서비스 모델을 개발함
- 자연어 질의응답, 본문 요약 등 머신러닝 기술의 적용방안을 수립함으로써 국회전자도서관의 디지털 서비스 역량 강화를 통해 전문성 있는 입법 활동 지원과 국민의 지식역량을 제고함

○ 국회도서관에 적합한 인공지능 기술과 서비스 범위의 이해와 적용을 위한 전략 수립 방안을 제시함

- 국회도서관에 적용할 수 있는 자연어 질의응답, 본문요약 서비스 등 머신러닝을 적용한 전자도서관 서비스 방안과 전략 수립을 위한 기초자료로 활용할 수 있음

2. 연구 내용

□ 인공지능과 자연어 처리 기술 동향 조사

○ 디지털 트랜스포메이션 추세와 자연어 처리, 머신러닝 등 다양한 신기술에 대한 분석과 국회도서관의 서비스 적용방안을 제안함

- 해외 주요국의 디지털 트랜스포메이션 관련 정책과 추세, 인공지능, 자연어 처리 최신 동향을 조사함
- 최신 신경망 언어모델 (BERT, GPT-3 등)에 대한 심층 분석과 한국어 처리를 위한 자연어 처리 분야의 핵심 기술에 대한 심층 분석을 수행함
- 국내외 기업, 도서관 분야의 자연어 처리 기술을 활용한 응용서비스와 구체적인 사례를 조사함

□ 국외 주요도서관 사례조사를 통한 벤치마킹 모델 발굴

○ 국외 주요도서관(미국, 일본 등) 사례조사를 통한 신규 정보기술 활용 도서관 서비스 벤치마킹 모델을 발굴함

- 미국, 일본 등 국외 주요도서관에서 인공지능 기술을 적용한 사례, 관련 프로젝트를 조사함
- 최신 논문, 동향 리포트, 뉴스를 조사하여 자연어 질의응답, 본문요약 기능 등 온라인 도서관 서비스에 적용된 알고리즘을 분석함

○ 현황 분석과 국외 사례 검토를 종합하여 실현 가능성이 높은 신규 전자도서관 서비스 모델을 개발함
- 국회전자도서관 서비스 현황을 분석하고, 적용 기술, 활용 데이터 현황을 검토함
- 해외 주요도서관 벤치마킹 모델을 바탕으로 기존 국회전자도서관 서비스의 문제 정의와 개선방안을 도출함

□ 인공지능 서비스 모델 제안

○ 페이오프 매트릭스(Payoff matrix)를 이용해 국회전자도서관이 도입할 수 있는 인공지능 서비스 모델의 현황을 제공함
- 서비스 도입 시 기대되는 성과와 도입 난이도를 기준으로 인공지능 기반 서비스의 도입 가능성을 제시하고, 단기·중기·장기적 관점의 추진 과제를 도출함
- 국회전자도서관이 보유하고 있는 데이터를 지식검색 서비스에 효과적으로 활용하기 위한 데이터 구축방안을 제안함
- 기계가 판독 가능한 방식으로 데이터를 표현하기 위한 지식그래프 기술, 데이터 상호운용과 재사용을 위한 메타데이터 표준화 방안을 제안함

3. 연구 방법

□ 사례조사

○ **(기술 동향 조사)** 인공지능 분야의 기술 동향을 조사하고, 최근 대두되는 심층머신러닝 기반 자연어 처리 분야의 핵심 기술을 분석함
- 미국, 중국, 일본과 국내 디지털 트랜스포메이션 관련 정책과 인공지능 분야의 핵심 기술을 조사·분석함
- 자연어 이해(Natural Language Understanding)와 자연어 생성(Natural Language Generationg)을 비롯한 자연어 처리 분야의 최신 알고리즘과 한국어 처리를 위한 언어모델에 대한 분석하고 주요 특징을 검토함
- 질의응답/검색, 문서 분류와 요약, 이미지 인식, 음성 인식, 기계번역, 지식그래프 등 자연어 처리 응용서비스 분야와 적용 가능한 기술을 조사함

○ **(국외 주요도서관 사례조사)** 주요 국가의 도서관에 적용된 인공지능 기술의 현황과 관련 프로젝트를 분석함

- 인공지능 기술 도입을 위한 프로젝트를 수행하는 국외 주요도서관 선정함*

* 미국 의회도서관, 일본 국립국회도서관, 프랑스 국립도서관, 노르웨이 국립도서관, 스웨덴 국립도서관, 핀란드 국립도서관, 네덜란드 국립도서관

- 기존 인쇄자료의 디지털화와 디지털 콘텐츠의 수집, 관리, 보존에 대한 사례를 조사함
- 인공지능 기술을 활용한 도서관 서비스 현황과 개발 프로젝트 사례를 조사하고, 실제 적용된 관련 기술과 알고리즘을 분석함
- 개별 도서관의 공식 홈페이지, 보도자료, 보고서 등 공식적인 자료, 인공지능 관련 프로젝트의 주요 깃허브, 학술논문 등 구체적인 기술 내용을 조사함

○ **(국회전자도서관 서비스 현황 조사)** 국회전자도서관이 제공하는 8개 웹사이트를 중심으로 서비스 현황을 분석함

- 국회전자도서관을 통해 서비스하는 국회법률도서관, 국회의원 정책자료, 국회·지방의회 의정정보시스템, 국회기록보존소, 전문가정보 데이터베이스, 지능형 분석(Argos)과 국회회의록 빅데이터 웹사이트의 서비스 현황을 검토함
- 공식 홈페이지, 사용설명서, 관련 보도자료와 국회도서관 연간보고서 바탕으로 관련 서비스 현황과 적용 기술을 조사함

□ 차세대 전자도서관 서비스 모델 개발

○ **(벤치마킹)** 자연어 처리 응용서비스 분야를 기준으로 국외 주요도서관 서비스 사례 벤치마킹을 위한 분석을 수행함

- 자연어 처리 분야를 6개의 응용서비스로 구분하고, 국외 주요도서관 사례조사, 국회전자도서관 서비스 현황에 대한 비교분석을 수행함
- 비교 분석 결과를 바탕으로 기존 국회전자도서관 서비스의 발전방안을 도출하고, 적용 가능한 국외 주요도서관 서비스 사례를 제안함

○ **(신규 서비스 도입을 위한 프레임워크)** 국회전자도서관에 도입 가능한 서비스들을 성과와 도입 난이도 기준으로 분류하여 제안함

- 국외 주요도서관 사례와 국회전자도서관 서비스 현황 분석을 통해 도입 가능한 서비스 모델을 제안함
- 성과와 비용을 기준으로 최선의 의사결정을 위한 페이오프 매트릭스를 적용해 서비스 모델을 분석함
- 제안 서비스는 관련 기술의 시장 상황, 국회도서관 적용 가능성 등을 고려해 단기, 중기, 장기 과제로 구분하고, 각 서비스의 도입 효과와 구현 기술, 해외 사례를 구체적으로 분석함

제 2 장 인공지능 기술 동향

1. 배경

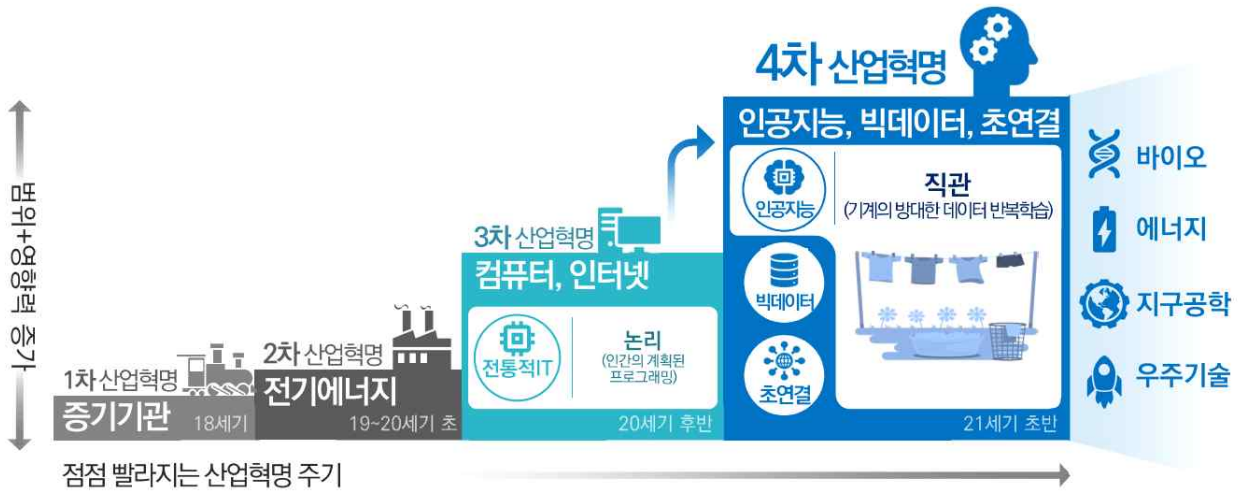
□ 디지털 기술 분야의 발전과 정책 환경 변화

○ **(D.N.A)** 디지털 기술의 발전으로 4차 산업혁명의 주요 요인인 데이터(D), 네트워크(N), 인공지능(A)은 사회 전반에 영향력이 강화되고 있음

- 정보화 시대를 넘어 디지털 환경이 사회 전반의 주류로 변화해 스마트시티, 사물인터넷, 자율주행 등 디지털 기술이 삶 전반에 영향을 미침
- 정부는 디지털 뉴딜, 데이터 댐으로 대표되는 정책을 통해 인공지능 학습에 필요한 데이터를 제공하고 있고, '데이터 3법*'을 통해 데이터 생태계를 변화시키고 있음

* 데이터 3법: 개인정보 보호법, 정보통신망 이용촉진 및 정보보호 등에 관한 법률(정보통신망법), 신용정보의 이용 및 보호에 관한 법률(신용정보법)을 통칭

- 정부는 2019년 '인공지능 국가전략'을 발표하여 경제와 사회문제 해결에 인공지능을 핵심 방안으로 제안하고, 인공지능을 활용한 새로운 부가가치를 창출하기 위한 패러다임의 변화를 이끌고 있음



<그림 2-1> 산업혁명 주기 (4차산업혁명위원회, 2018)

○ **(디지털 트랜스포메이션)** 4차 산업혁명의 영향으로 디지털 신기술을 도입하고 활용하는 디지털 트랜스포메이션(Digital Transformation) 추진하고 있음

- 산업 전반의 디지털 전환 과정을 통해 생산성과 경쟁력을 향상시키고 일자리를 창출하는 등 사회적인 문제 해결에 도움이 될 것으로 예상됨
- 디지털 트랜스포메이션의 주요 동인은 빅데이터, 인공지능, 모빌리티, 클라우드 컴퓨팅, 사물인터넷이며, 빅데이터와 인공지능 관련 기술은 최근 10년간 급속히 성장하고 있음

- 기업과 산업계를 중심으로 시작된 디지털 트랜스포메이션 기술은 공공 분야에 빠르게 적용되어 사회적 문제 해결을 위해 적극적으로 도입되고 있음

□ 선진 주요국의 디지털 트랜스포메이션 정책

○ 선진 주요국은 국가 경쟁력을 강화하기 위해 디지털 트랜스포메이션 정책을 국가 차원에서 적극적으로 추진하고 있음

- 2010년대 이후 국가별 기술력, 산업구조, 정책 등 환경에 따라 차별화된 전략을 추진하면서 디지털 정책에 따라 국가 경쟁력 순위에 영향을 미침
- 대부분의 OECD 국가는 디지털 전환에 적합한 혁신과 성장을 국가 차원에서 주도하고 있고, 정책적 의사결정 과정에서 디지털 정책의 중요성을 강조함
- 특히, 디지털 트랜스포메이션의 주요 요소인 데이터와 인공지능의 사회적·경제적 성장을 주도하기 위해 국가별로 인공지능 전략을 수립하고 적극적으로 추진하고 있음

○ **(미국)** 디지털 트랜스포메이션을 위한 인공지능 산업 육성에 집중하여 민간의 산업 혁신과 정부의 기초연구 투자로 생태계를 조성함

- 오바마 정부부터 시작된 '디지털 정부' 혁신 정책은 국가 차원에서 인공지능이 초래할 미래상을 조망하고 산업계와 정부에서 대응해야 할 전략 방향을 제시함
- 바이든 정부는 '인공지능 국가 이니셔티브(National Artificial Intelligence Initiative, NAII)' 정책을 내세워 미국의 인공지능 기술력과 시장을 세계적으로 확대하고 있음
- 정부는 대규모의 비용과 장시간이 필요한 원천기술 개발과 인프라 조성에 초점을 둔 정책 추진하여 기초연구에 투자함. 민간의 산업 혁신을 촉진하기 위해 AI 기술 개발을 지원하고, 산업 적용을 저해하는 규제를 완화함
- 공공분야에 클라우드, 빅데이터 등 신기술 도입을 적극적으로 추진하여 선순환 생태계 조성을 위한 산·학·관 사이의 협력을 중시함

○ **(중국)** 인공지능 산업 생태계 전환을 위한 응용연구에 초점을 두고 데이터 개방·공유 등 기업 친화적 정책을 추진하고 있음

- 중국은 양적 성장에서 질적 성장으로 발전하기 위해 학제간 연구 지원, 인공지능 연구를 위한 인프라를 개선하기 위해 노력함
- 2017년 'A Next Generation AI Develpment Plan' 인공지능 전략을 통해 국가 중점 분야에서 인공지능을 적극적으로 활용하고 디지털 전환을 위한 정책을 강력하게 추진하고 있음
- 2020년 코로나19로 침체된 경기를 회복을 위해 '신(新) 인프라' 분야에 해당하는 5G 통신망, 인공지능, 사물인터넷 등에 투자함

○ **(일본)** 사회적 현안을 해결하기 위해 디지털 트랜스포메이션 기술을 도입해 전 분야에 할 수 있는 인프라를 조성함

- 2017년 'AI Technology Strategy' 정책을 발표해 고령화, 인구부족 등의 사회문제를 해결하기 위한 포용적 디지털 전환 정책을 발표함
- 지속적인 디지털 산업 생태계 조성을 위한 구체적인 미래상을 제시하며 기반기술과 인재양성에 우선 투자하는 단계별 이니셔티브를 추진하고 있음
- 인공지능 기술을 도입하여 다양성과 포용성, 지속가능성을 위한 사회로 전환하기 위한 디지털 포용 부문에 집중하여 각 산업의 영역을 복합적으로 연계하기 위해 노력함

○ **(한국)** 국가 경쟁력 향상을 위한 인재 양성과 양질의 일자리 창출을 위한 디지털 전환 정책을 추진함

- OECD 국가를 대상으로 공공데이터를 평가하는 'OUR data Index'에서 3회 연속 1위를 기록하며, 국가적 차원에서 데이터 부문에 적극적으로 투자하고 있음
- 문재인 정부에서 디지털 역량을 국가 경쟁력의 핵심 요소로 주목해 경제 활성화와 일자리 창출을 위한 '디지털 뉴딜' 사업을 추진하고, D.N.A(데이터, 네트워크, 인공지능) 생태계를 강화하고, 빅데이터와 인공지능 활성화를 위한 '데이터 댐'을 구축함
- 과학기술정보통신부는 2018년 인공지능 기술, 인재, 인프라 분야의 현황을 분석하고 경쟁력을 확보하기 위한 'AI R&D Strategy' 전략을 추진하고, 2020년에 관계부처 합동으로 '인공지능 국가전략'을 발표하여 경제 활력 제고와 사회문제 해결을 위한 국가적 전략 마련함

[표 2-1] 주요국 디지털 정책

국가	주요 정책
미국	• 2019.12 미국 AI 이니셔티브(Amerian AI Initiative) • 2020.01 AI 어플리케이션 규제에 관한 가이드 • 2020.01 국가 AI 이니셔티브(National AI Initiative) • 2021.06 국가 인공지능 연구자원 태스크 포스 출범
중국	• 2017.07 차세대 인공지능 발전계획 • 2017.12 차세대 AI 산업발전 3개년 행동계획 • 2020.03 신(新)인프라 건설 정책
일본	• 2017.02 AI 산업화 로드맵 • 2017.05 신산업구조 비전 • 2019.06 AI 전략 2019
한국	• 2018.05 인공지능 R&D 전략 • 2019.01 데이터·AI 경제 활성화 계획('19~'23) • 2019.12 인공지능 국가전략

2. 인공지능

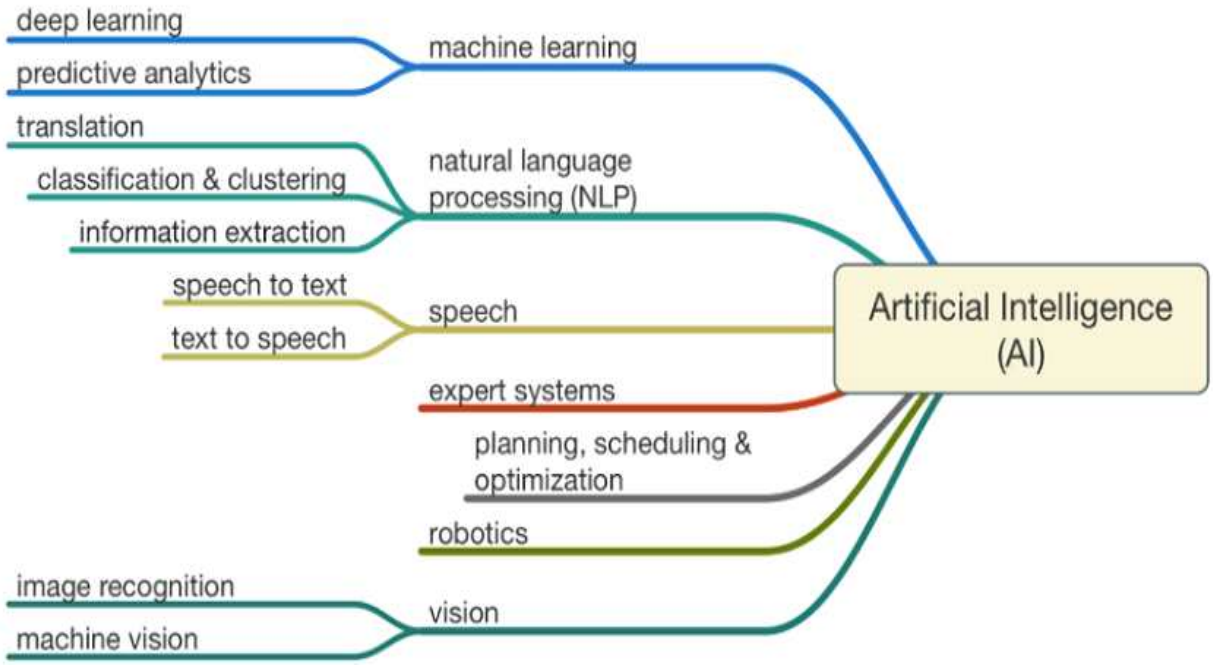
□ 정의

○ 인공지능(AI, Artificial Intelligence)은 인간의 개입과 조작이 통제된 상태에서도 인간의 사고와 유사한 결과를 도출할 수 있는 ICT(Information & Communication Technology) 기술

- 존 매카시(John McCarthy)는 인공지능을 '고도의 지능을 가진 컴퓨터 디바이스를 만드는 과학과 공학' 또는 '사람이 그렇게 행동했다면 지능적이라고 말할 수 있는 행동을 하는 기계'로 정의함
- 인공지능은 인간의 인지능력, 학습능력, 추론능력, 이해능력 등과 같이 인간의 고차원적인 정보처리 능력을 구현한 것으로, 인간 사고와 지능적인 행위의 매커니즘을 이해하고, 기계에 구현하는 기술

○ 인공지능은 머신러닝(머신러닝), 딥러닝, 자연어 처리, 음성 인식, 시각인식 등 첨단 기술을 개발하며, 다양한 응용 분야에 특화한 소프트웨어와 머신러닝, 클라우드, 고성능 컴퓨터 기술 등 포함

- 발달한 반도체 기술을 바탕으로 한 고사양 하드웨어를 바탕으로 패턴인식, 머신러닝, 전문가시스템, 인공신경망, 자연어 처리 등 다양한 분야의 소프트웨어 기술과 융합하며 발전하고 있음



<그림 2-2> 인공지능 기술 분야 (Future Architecture, 2016)

□ 유형

○ 인공지능의 결과가 인간의 통제범위 내에서 발생한 것인지에 따라 ‘약한 인공지능(Weak AI)’와 ‘강한 인공지능(Strong AI)’로 구분

- **(약한 인공지능)** 인간의 지시와 통제범위 내에 존재하는 인공지능으로, 제한적이고 특정한 문제를 해결해 생산성을 증진하기 위한 도구적 측면으로 사용하는 기술
- **(강한 인공지능)** 인간과 동등한 지능을 가지고 사람처럼 느끼는 인공지능을 의미하며, 기능적으로 인간과 대등하거나 경쟁할 수 있는 수준

○ 스튜어트 러셀(Stuart Russell)과 피터 노빅(Peter Norvig)은 인공지능이 목적을 달성하는 방식에 대해 사람처럼 또는 이성적으로 생각하고, 행동하는가에 따라 4가지 관점으로 구분함

[표 2-2] 인공지능에 대한 4가지 관점 (Artificial Intelligence, 2020)

	Human	Ideal
Thought	Thinking humanly	Thinking rationally
behavior	Acting humanly	Acting rationally

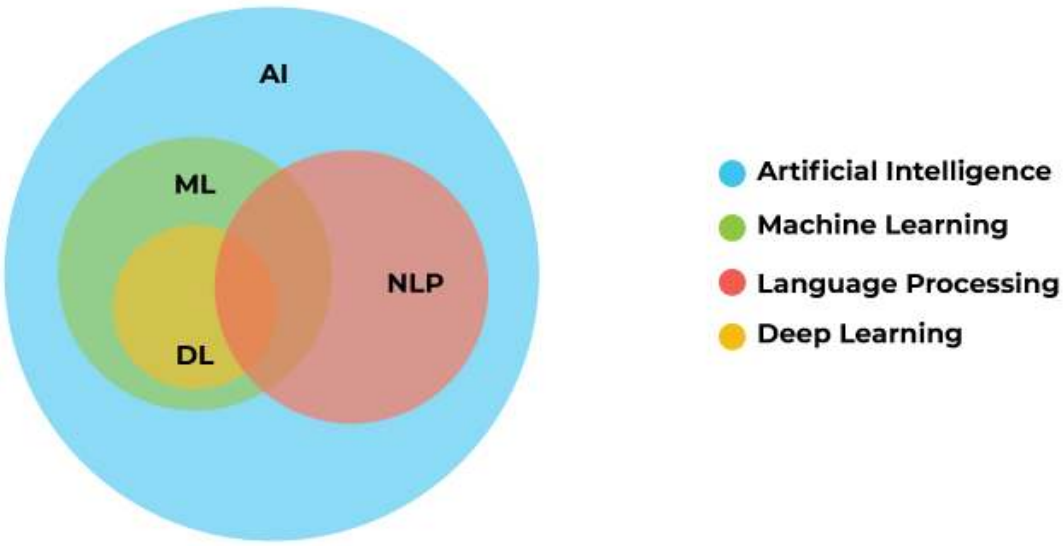
- **(Thinking humanly)** 사람이 생각하는 방식을 모방하여 사람 뇌의 동작 방식과 동일하게 동작하는 인공지능을 의미하며, 주로 인지 과학 분야에서 연구됨
- **(Thinking rationally)** 논리적으로 생각할 수 있다면 인공지능이라는 뜻으로, 논리학, 논리적 추론의 관점에서 인공지능을 정의하는 관점
- **(Acting humanly)** 사람같이 행동 또는 응답하는 것을 인공지능으로 정의하며 튜링 테스트(Turing test)*의 근거가 되는 관점

* 튜링 테스트: 1950년 앨런 튜링은 기계가 사람인지 기계인지 구별할 수 없을 정도로 대화를 잘 이끌어간다면, 기계가 ‘생각’하고 있음을 판단할 수 있다는 이론

- **(Acting rationally)** 기대 효용을 최대화하는 관점으로, 프로세스의 최적화 또는 효율화의 관점에서 인공지능을 정의함

□ 하위 분야

○ 인공지능은 인간의 지적 능력을 컴퓨터를 통해 구현하는 기술로, 머신러닝과 딥러닝, 자연어 처리 세 분야를 포함하고, 자연어 처리는 머신러닝과 딥러닝을 활용함



<그림 2-3> 인공지능의 하위 분야

○ **(머신러닝)** 입력된 데이터로부터 학습을 통해 규칙을 학습하고, 결과를 예측하거나 판단을 제공하는 기술

- 지능을 구현하기 위한 소프트웨어이며, 주로 통계학에서 유래된 알고리즘을 활용하지만, 최근 수학이론, 통계학보다 경험을 바탕으로 하는 이론이 주목받고 있음
- 학습데이터의 제공방식에 따라 지도학습(Supervised Learning), 비지도학습(Unsupervised Learning)으로 구분할 수 있음. 지도학습은 정답이 있는 데이터(labeled data)를 활용해 모델을 학습시키는 방법이고, 비지도학습은 정답이 없는 데이터로 컴퓨터를 학습시키는 방법으로, 데이터의 숨겨진 특징이나 구조를 발견하는 데 사용함
- 강화학습(Reinforcement Learning)은 컴퓨터 에이전트가 주어진 상태를 인식하고, 주어진 선택지 중 보상을 최대화하는 행동을 선택하면서 학습. 강화학습은 동적인 상태(dynamic environment)에서 데이터를 수집하는 과정이 포함된 학습 방법이며, 최근 신경망 기반의 강화학습이 주목받고 있음

○ **(딥러닝)** 여러 층을 가진 인공신경망을 사용해 머신러닝을 수행하는 기술로, 기계가 자동으로 학습데이터에서 특징을 추출하여 모델을 학습함

- 딥러닝 이론은 1980년대에 처음 등장했으나, 느린 속도로 인해 일반적으로 사용되지 못함. 2000년대 이후 강력한 GPU를 갖춘 고성능 컴퓨터의 등장으로 속도 문제가 해결되면서 딥러닝 기술이 다시 주목받고 있음
- 텍스트, 오디오, 이미지, 비디오 등 비정형 데이터에 적용할 수 있어 다양한 분야에서 활용되고 있으며, 특히 자동 음성 인식(ASR, Automatic Speech Recognition)과 컴퓨터 비전 분야에서 탁월한 성능을 보임

○ **(자연어 처리)** 인간의 언어를 기계가 이해하고, 해석하고, 조작할 수 있도록 하는 기술로, 인공지능의 주요 분야

- 자연어 처리는 기계가 읽고, 이해하는 것을 의미하는 자연어 이해(Natural Language Understanding)과 데이터를 사람의 단어로 변환할 수 있는 자연어 생성(Natural Language Generating)으로 구분함
- 정보검색, QA 시스템, 문서 자동분류, 기계번역, 대화형 에이전트 등 다양한 서비스에 응용될 수 있는 기술
- 최근 음성, 영상 등 구조화되지 않은 데이터에서 텍스트 데이터를 효과적으로 수집하여 자연어 처리 기술을 적용하는 딥러닝 기술이 발전함

□ 기술 분류

○ 인공지능 기술은 지식을 확보하는 알고리즘을 연구하는 학습지능, 시각, 청각, 언어 등 한 종류 입력을 바탕으로 지식을 확보하는 단일지능, 여러 형태의 입력을 통합해 이해하고 판단하는 복합지능 기술로 분류함

- 최근 주목받고 있는 복합지능은 강한 인공지능의 한 부분이며, 딥러닝을 기반으로 각기 다른 센서로부터 습득한 정보를 통합해 활용하고, 스스로 학습·판단·진화한다는 특징을 가짐

[표 2-3] 과학기술일자리진흥원(COMPA) 인공지능 기술 분류

소분류	세분류	요소 기술
학습지능	머신러닝	베이지안 학습, 인공신경망, 딥러닝, 강화학습, 앙상블러닝, 판단근거 설명
	추론·지식표현	추론, 지식표현 및 온톨로지, 지식 처리
단일지능	언어지능	언어분석, 언어의미 이해, 대화 이해 및 생성, 자동 통역·번역, 질의응답(QA), 텍스트 요약·생성
	시각지능	영상 처리 및 패턴인식, 객체 인식, 객체 탐지, 행동 이해, 장소·장면 이해, 비디오 분석 및 예측, 시공간 영상 이해, 비디오 요약
	청각지능	음성분석, 음성 인식, 화자인식·적응, 음성합성, 오디오 색인 및 검색, 잡음처리 및 음원분리, 음향인식
복합지능	행동·소셜지능	공간 지능, 운동 지능, 소셜 지능, 협업 지능
	상황·감정이해	감정 이해, 사용자 의도 이해, 뇌신호 인지, 센서 데이터 이해, 오감인지, 다중 상황 판단
	지능형 에이전트	에이전트 플랫폼, 에이전트 기술, 게임 지능, 모방창작 지능
	범용 인공지능(AGI)	상식 학습, 범용 문제 해결, 평생 학습, 도덕·윤리·법 지능

3. 발전 동향

□ 국제 동향

○ 국내외 정부와 IT 기업은 인공지능 관련 정책을 수립하고, 기술 개발에 적극적으로 투자하고 있음

- 4차 산업혁명은 사람과 사물, 공간을 초연결되는 사회를 의미하며, 핵심 요소로 인공지능 기술이 강조됨
- 미국, 일본, 유럽, 중국 등 주요 선진국은 인공지능 관련 정책과 장기적인 로드맵을 수립하고, 핵심 기술 개발과 인재 양성에 주력하고 있음
- 구글, 마이크로소프트, 페이스북, 지멘스, 아마존 등의 선진 기업은 딥러닝 기술을 사용하여 인터넷 서비스, 생산공정 자동화, 의료, 보안, 광고 등 다양한 분야에 적용해 수익을 창출함

○ 최근 인공지능 기술은 단순한 인지능력에서 벗어나 학습 기반의 추론과 예측, 판단을 제공할 수 있는 머신러닝, 딥러닝 기술로 발전하고 있으며, 향후 스스로 문제를 발견하고 해결하는 행동 단계에 이를 수 있도록 연구와 투자가 이뤄지고 있음

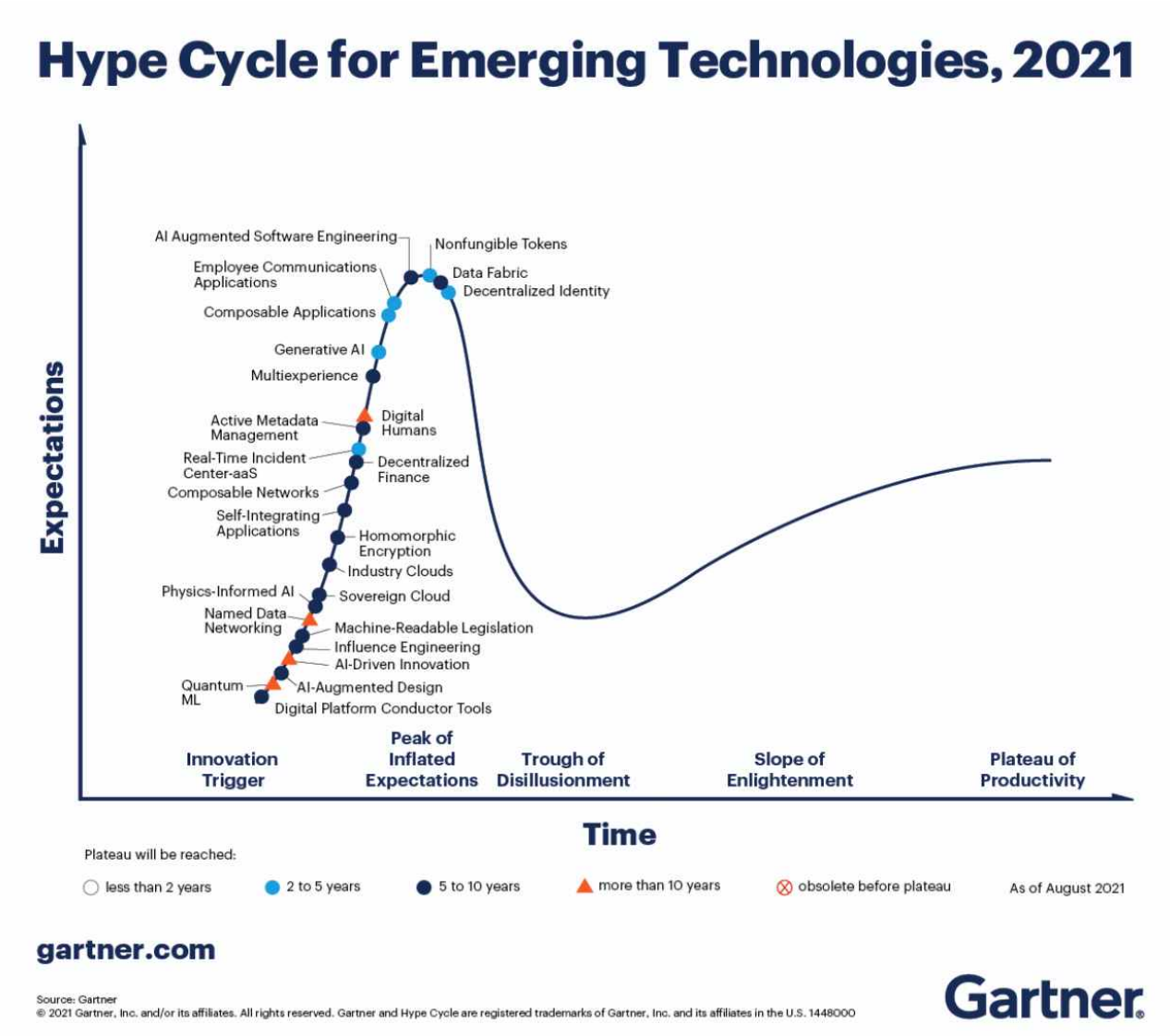
- 최근 자율적인 상황 판단과 능동적 행동을 기반으로 하는 인간의 지능 수준에 도달하기 위해 수학·과학 분야의 딥러닝 연구가 이뤄지고 있음
- 기존 인공신경망 구현 방식을 하드웨어, 소프트웨어적으로 고도화하기 위해 딥러닝 기술에 신경과학, 뇌과학 분야의 연구를 접목하고자 하는 시도가 이뤄짐



<그림 2-4> 인공지능 트렌드 (LG 경제연구원 이승훈, 2017)

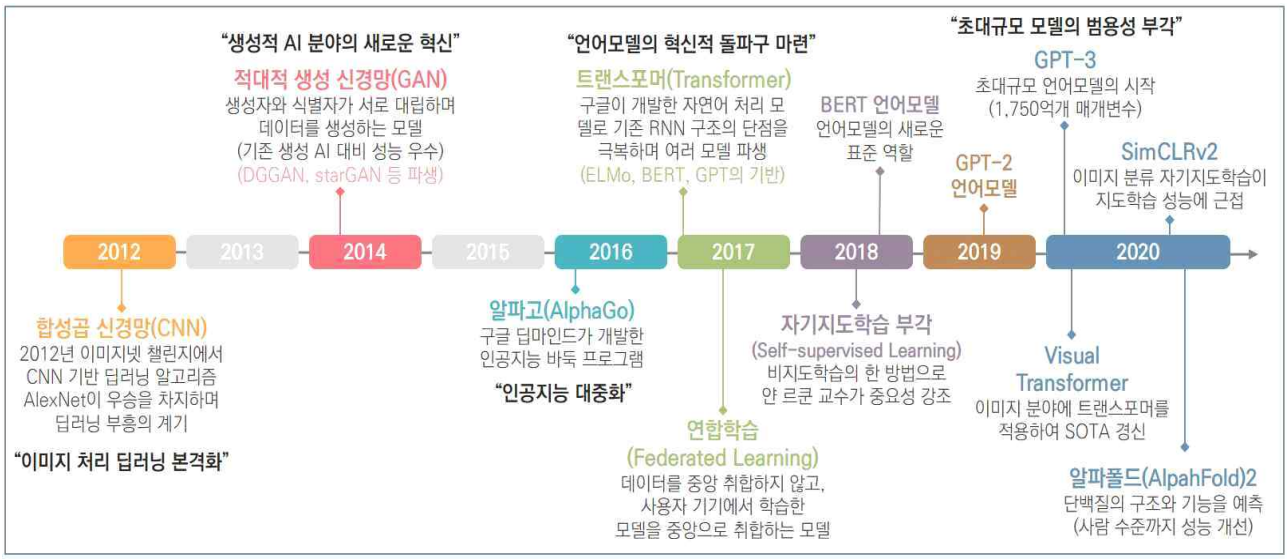
○ 가트너(Gartner)는 향후 2~5년의 핵심 기술로 컴퓨터 비전, 생성적 인공지능, 딥러닝, 머신러닝 등을 선정했고, 자연어 처리와 트랜스포머 기술은 향후 5~10년 이후 핵심 기술로 선정됨

- 미국 정보기술 연구·자문 회사인 Gartner는 매년 핵심 신기술을 식별하고, 새로운 기술에 대한 기대와 수용의 변화를 단계별로 나타내는 Hype Cycle 발표함
- 2021년 8월 발표된 Gartner Hype Cycle은 향후 2~5년 동안 혁신적인 영향을 미칠 핵심 신기술로 복합 인공지능, 컴퓨터 비전, 의사결정 지능, 딥러닝, 생성적 인공지능, 머신러닝 등을 선정함
- 자연어 처리와 트랜스포머 기술은 향후 5~10년 이후의 핵심 기술로 선정됨. 현재 자연어 처리와 트랜스포머 기술은 특정 언어에 최적화되어 있어 기술의 편향성을 극복하기 위해 소수 언어에 대한 기술 개발이 필요함



<그림 2-5> 2021 Gartner Hype Cycle (Gartner, 2021)

□ 핵심 기술 동향



<그림 2-6> 인공지능 분야 주요 기술 (NIA, 2021)

○ 합성곱 신경망(CNN) 알고리즘이 등장하면서 딥러닝 기반 서비스가 발전함

- 2012년 이미지넷(ImageNet) 이미지 인식 대회에서 인간의 뇌 구조를 반영한 인공신경망 모델인 합성곱 신경망을 사용하여 심층 신경망 구현함
- 기존 CPU를 활용한 연산과 달리 여러 명령어를 동시에 처리할 수 있는 병렬 처리방식인 GPU를 활용하여 연산의 속도를 개선함

○ 생성적 AI(Generative AI)의 핵심 기술인 GAN(Generative Adversarial Networks) 모델은 이미지 생성과 같은 창조적인 기술에 딥러닝 활용함

- 적대적 생성 신경망이라고도 하는 GAN 모델은 서로 다른 두 개의 네트워크를 적대적으로 학습시켜 새로운 데이터를 생성하고 있음
- 진짜와 유사한 가짜 이미지나 음성과 같은 콘텐츠를 만들어내는 기술로 악용될 가능성이 있기 때문에 주의가 필요함

○ 2016년 알파고의 등장 이후 사회 전반에 인공지능을 활용한 서비스 도입이 활성화되고 있음

- 각국의 핵심 전략으로 인공지능이 대두되었고, 국내외 IT 대기업은 대화형 인공지능(Siri, Google Assistant 등)과 같은 자연어 처리 기반 인공지능 서비스를 구축함

○ 구글이 개발한 트랜스포머(Transformer) 모델은 순환 신경망(RNN)의 단점을 극복하여 자연어 처리 발전의 기반을 마련함

- 트랜스포머 이전은 RNN 모델과 그 파생 구조인 LSTM(Long Short Term Memory),

GRU(Gated Recurrent Unit)을 주로 사용했으나, 긴 문장을 처리할 때 앞쪽의 단어에 가중치가 높게 측정되는 한계가 있음

- RNN 모델의 단점을 극복한 트랜스포머는 2018년 OpenAI의 GPT(Generative Pre-trained Transformer), 구글의 BERT(Bidirectional Encoder Representations from Transformers)의 기반으로 활용되고 있으며, 기계번역, 텍스트 요약, 텍스트 생성 등 다양한 분야에 영향을 미침

○ 2020년, OpenAI가 공개한 GPT-3는 GPT의 세 번째 언어모델이며, 기존 언어모델에 비해 월등한 성능을 보임

- GPT-3는 약 1,750억 개의 매개변수를 가진 초대규모 언어모델로, 여러 활용 사례에 응용되고 있음
- 대규모 데이터셋을 사용하여 학습하기 때문에 언어모델 학습을 위한 데이터셋 구축의 중요성을 강조하고 있음

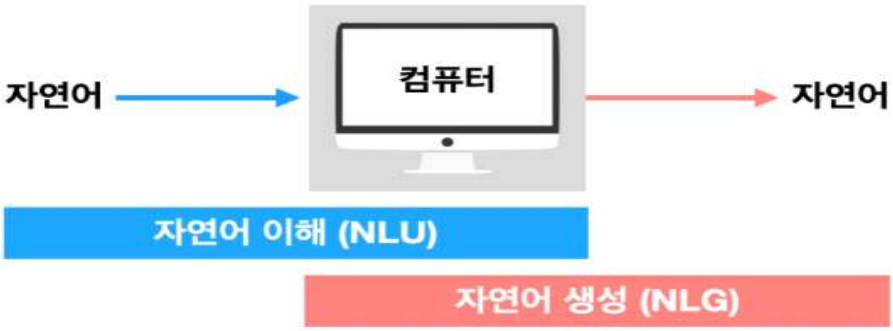
제 3 장 자연어 처리 기술 동향

1. 자연어 처리

□ 자연어 처리

○ 자연어 처리는 컴퓨터가 자연어를 읽고, 해독하고, 이해하기 위한 기술로, 인공지능 기반의 서비스가 인간과 상호작용할 수 있도록 하는 핵심 인공지능 기술임

- 자연어 처리는 기계가 자연어를 이해할 수 있도록 변환하는 기술인 '자연어 이해'와 결과값을 사람이 이해할 수 있도록 변환하는 '자연어 생성'이 포함됨



<그림 3-1> 자연어 처리의 하위 분야

○ **(자연어 이해)** 비구조화된 데이터를 기계가 이해할 수 있는 구조화된 데이터로 변환하여 기계가 단어의 의미를 인식하도록 하는 기술

- 기계가 빠르고 정확하게 사용자의 의도를 이해하고, 유용한 정보를 추출하기 위한 핵심 기술이며, 의도에 따른 분류, 질문에 대한 답변 추출, 추론 등에 적용함
- 과거 사람이 직접 추출한 특징(feature)에 의존하는 규칙 기반의 자연어 이해에서 기계가 자동으로 모델을 학습하는 머신러닝, 기계가 자동으로 특징을 추출하고 학습하는 딥러닝 기반의 알고리즘으로 변화하면서 인간의 개입이 감소함
- 자연어 이해의 대표적인 기술은 단어나 형태소를 벡터화하는 워드 임베딩 기술, MRC 기술*, 대화 모델**이 해당함

* MRC 기술: 질의응답 기술로, 자용자의 질문에 대해 기계가 학습한 내용 중 적절한 답변을 찾아 제공하는 기술

** 대화 모델: 기계가 사람과 같이 자연스럽게 대화할 수 있도록 맥락을 이해하고, 지능적인 반응을 제공하는 AI 모델

○ **(자연어 생성)** 구조화된 데이터를 자연어 형식의 의미 있는 구문으로 변환하는 기술

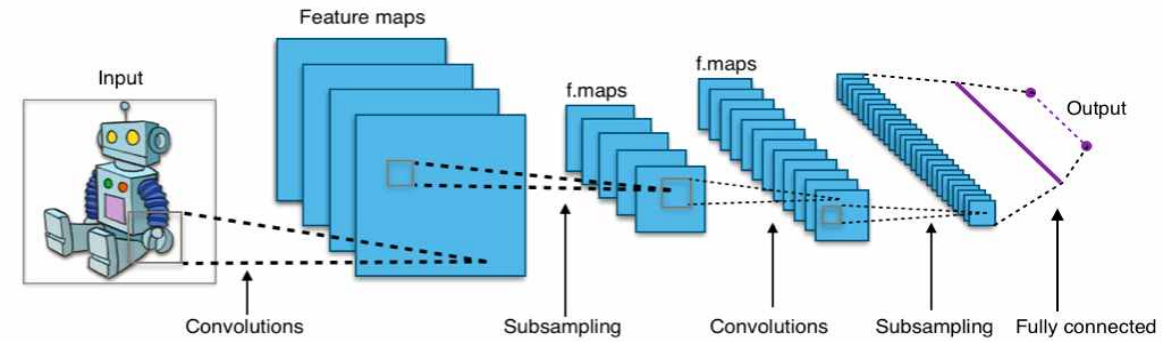
- 말뭉치 데이터를 통해 학습하여 사람이 하는 말의 맥락을 파악해 텍스트를 생성하고, 챗봇, 요약, 인공지능 비서(AI 스피커)에 주로 사용함
- 문장과 단어의 연결성을 파악하기 위해 RNN(Recurrent Neural Network), LSTM(Long Short-Term Memory) 알고리즘을 주로 사용하며, 최근에는 오픈AI(openAI)의 GPT-3 언어모델이 자연스러운 문장 생성에 높은 성능을 보임

2. 딥러닝 기반 자연어 처리

□ CNN(Convolutional Neural Network)

○ CNN은 합성곱(Convolution) 전처리 작업이 들어가는 신경망(Neural Network) 모델로 DNN(Deep Neural Network)의 단점을 보완하기 위해 개발됨

- 기존 DNN은 일반적으로 1차원 형태의 데이터를 사용하고 다차원의 데이터를 입력하는 경우, 1차원으로 변환하여 처리하기 때문에 이 과정에서 공간/지역적 정보의 손실이 발생함
- CNN은 정보의 손실 없이 공간/지역적 정보를 유지한 채 행렬 형태의 데이터를 입력받기 때문에 사람의 신경망과 유사한 방식으로 학습하여, 이미지 처리에 대한 성능이 높아 자율주행, 얼굴, 개체 인식 등에 사용됨
- 합성곱(Convolution)이란 데이터에서 특징을 파악 후 추출하는 작업(예 : 새 → 부리, 눈, 깃털모양 등)으로 정사각형 모양의 필터로 추출하여 위치마다 포개지는 숫자를 곱해서 모두 더하는 방법으로 계산함



<그림 3-2> 일반적인 CNN 아키텍처

□ LSTM(Long Short-Term Memory)

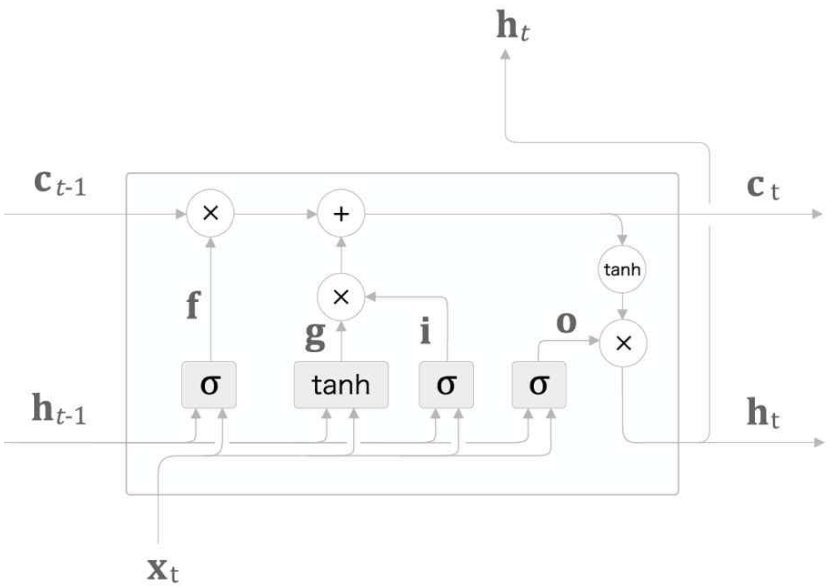
○ LSTM은 출력과 먼 위치에 있는 정보를 기억할 수 없다는 RNN의 단점을 보완하기 위해 장단기 기억이 가능하도록 설계한 신경망 구조

- RNN은 순환신경망으로 입력과 출력을 시퀀스* 단위로 처리하여 연속형 데이터에 적합한 모델이지만, 시퀀스가 길어지면 데이터의 장기 의존 관계에 대한 학습능력이 낮아지는 문제가 발생함

* 시퀀스(Seqeunce): 연관된 연속의 데이터로, 문장의 단어 나열, 음성, 시계열 데이터 등을 나타냄

○ 장기 의존성 문제를 해결하기 위해 연산과정에 입력(Input)·망각(Forget)·출력(Output) 게이트를 추가함

- 셀 상태(Cell state)라는 구조를 만들어 각 상태 값을 메모리 공간 셀에 저장하고, 게이트의 결과에 따라 불필요한 연산을 줄여 장기 의존성 문제의 일부분을 해결함
- 입력(Input) 게이트는 기억 셀의 각 원소가 새로 추가되는 정보의 가치를 판단하여 새 정보를 무비판적으로 수용하는 것이 아니라 적절히 선택하여 저장함
- 망각(Forget) 게이트는 RNN의 장기 메모리 희소성 문제를 해결하기 위해 과거 정보의 삭제 여부를 결정함
- 출력(Output) 게이트는 추가된 값이 다음 계층에 얼마나 영향을 미치도록 할지 시그모이드(Sigmoid) 함수를 통해 결정함



<그림 3-3> LSTM 모델의 주요 계층 구조

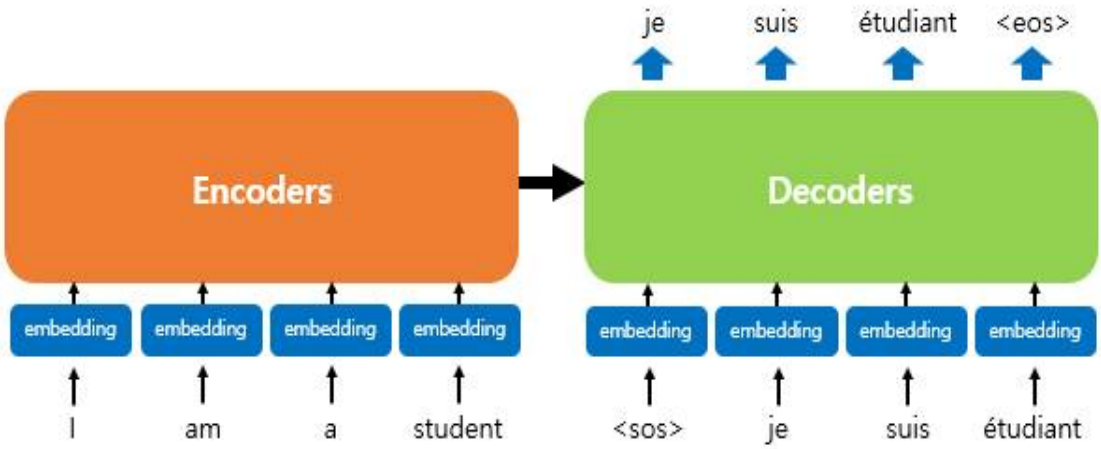
□ 트랜스포머(Transformer)

○ 2017년 구글이 발표한 논문인 “Attention is all you need”의 Seq2Seq 모델의 구조를 따르지만 어텐션(Attention)만으로 구현함

- Seq2Seq 모델은 단어같은 특정 속성을 지닌 입력값을 다른 속성의 결괏값으로 변환하는 작업을 의미하며, 이 과정에서 입력값을 처리하는 인코더(Encoder)와 결괏값을 생성하는

디코더(Decoder)로 구성됨

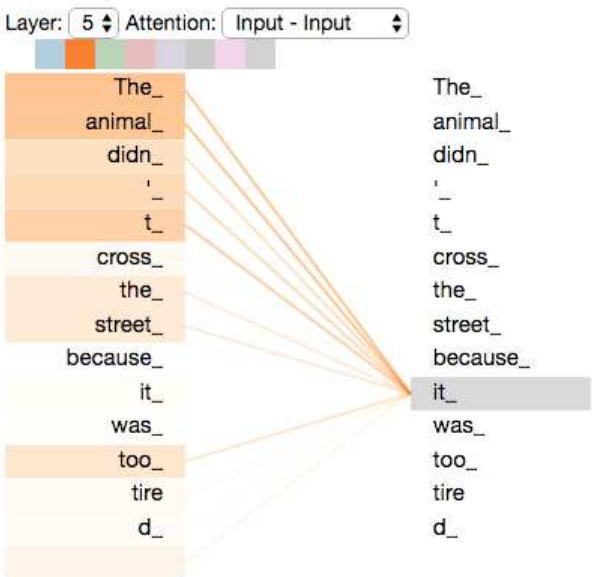
- 기계번역에 대표적으로 사용되는 모델로, 인코더에서 입력값의 모든 단어를 벡터화한 컨텍스트 벡터를 디코더에서 전송받아 순차적으로 번역하여 출력함



<그림 3-4> 인코더와 디코더 아키텍처

○ 트랜스포머는 기존의 Seq2Seq의 단점을 보완하기 위해, 디코더에서 결괏값을 예측할 때마다 인코더의 입력값을 다시 한번 참고하는 어텐션(attention) 기법을 적용함

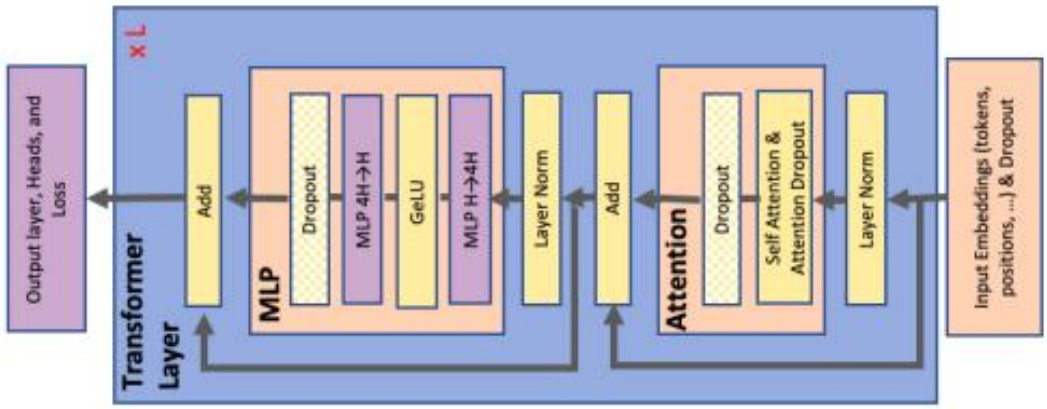
- 어텐션 기법의 기본적인 아이디어는 단어를 예측하는 시점마다 전체 입력 문장의 단어들의 간의 관계를 참고하여 각 시점에서 단어와 가장 연관된 단어들에 '집중'해서 반영함
- "The animal didn't cross the street because it was too tired." 문장을 보면 it이 의미하는 것이 'animal'인지 'street'인지 사람은 구별하지만 기계(컴퓨터)는 구별하지 못함
- 컴퓨터가 it의 의미를 이해하기 위해 셀프 어텐션을 통해 문장 내 단어와 자신의 연관성을 계산해 어텐션 값을 계산하여 자신과 연관성이 높은 단어를 찾음



<그림 3-5> 학습된 attention의 예시

○ 자연어 처리의 다양한 문제를 해결하기 위해 사용되며 대표적으로 언어모델링, 문장 예측, 질의응답(QA), 자연어 이해(NLU), 텍스트 요약 등에 사용함

- 트랜스포머 모델을 사용한 BERT와 RoBerta, SpanBERT 등과 같은 BERTology(BERT 계열 언어 모델) 와 같은 다양한 파생 모델들이 등장하였으며, 사전학습 모델(Pre-Trained Model, PTM)의 시대의 도래에 가장 큰 영향을 미침



<그림 3-6> 다층 퍼셉트론을 추가한 트랜스포머 아키텍쳐

□ BERT(Bidirectional Encoder Representations from Transformers)

○ 2018년 구글이 개발한 자연어 처리 태스크를 위한 트랜스포머 기반의 사전학습 모델로, 100여 개가 넘는 언어학습을 지원함

- 방대한 양의 데이터(약 33억 개 단어)로 먼저 학습되어 있고 본인의 사용 목적에 따라 본인 데이터를 추가로 학습시키는 파인튜닝*이 가능함

* 파인튜닝(Fine-Tunning): 특정 태스크에 맞추기 위해 모델의 매개변수를 매우 정확하게 조정해야 하는 프로세스

- 대량의 텍스트 데이터와 다양한 언어를 적용할 수 있어 자연어 처리 분야에서 범용적으로 좋은 성능을 나타내는 언어모델임

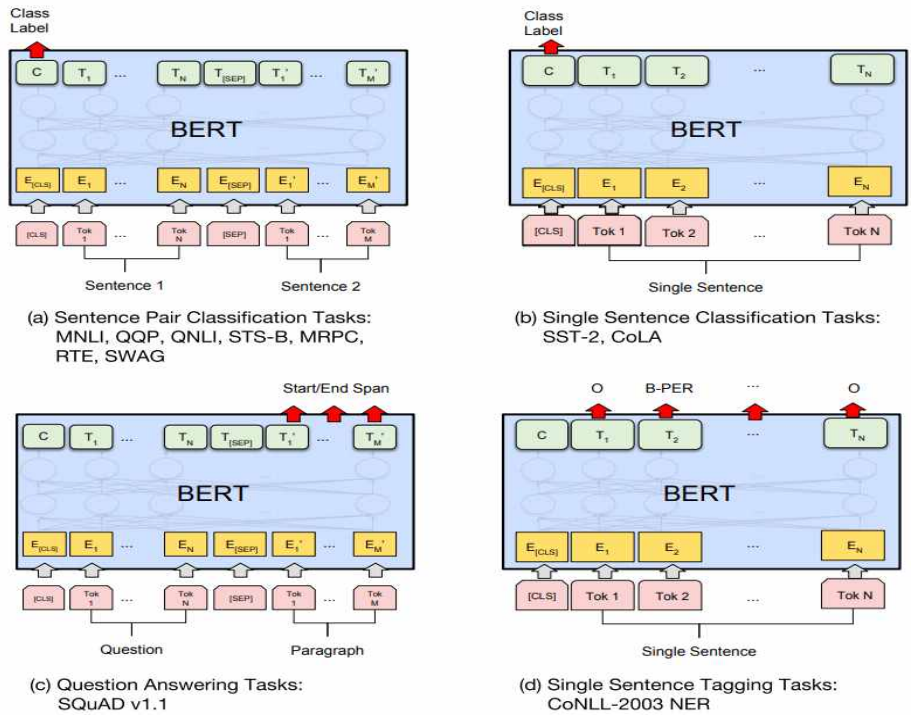
○ BERT는 특정 모델을 시작하기 전 사전 훈련 임베딩을 통해 특정 모델의 성능을 높일 수 있음

- BERT 등장 이전의 전처리 임베딩 기법인 Word2Vec, Glove, FastText 방식은 문맥 정보를 반영하기 어렵다는 단점이 존재함

- 반면, BERT 모델은 기본적으로 대용량의 단어 임베딩 등에 대해 사전학습이 되어 있는 모델을 제공하기 때문에 상대적으로 적은 자원만으로 다양한 태스크에 적용 가능함

○ 일반 NLP 모델에서 잘 작동하지만, Bio, Science, Finance 등 특정 분야의 전문 어휘가 등장하는 경우 언어모델을 적용하기 어려움

- 특정 도메인에 대해 BERT를 적용하려면 특정 분야의 특성을 수집할 수 있는 도메인 데이터로 추가적인 언어모델 학습이 필요함



<그림 3-7> BERT 파인튜닝 아키텍처

□ GPT(Generative Pre-trained Transformer)

○ OpenAI에서 개발한 언어모델로, 트랜스포머의 디코더 부분을 사용하여 이전 단어들을 가지고 다음 단어를 단방향으로 계산함

- 2018년에 발표한 GPT-1은 트랜스포머의 디코더 부분을 사용한 구조로, 원시 텍스트의 장기 의존성(Long-term dependencies)에 높은 성능을 보여주고 기존 RNN 등에 비해 구조화된 메모리를 사용할 수 있음
- GPT-2는 단방향 언어모델로 양방향 언어모델인 BERT에 비해 단점이 더 크다는 지적이 있었으나, 텍스트 생성 작업에는 더 적합하다고 평가됨

○ 2020년 OpenAI는 초대규모 매개변수를 가진 GPT-3을 공개하여 최신 언어모델 중 가장 주목받는 언어모델로 자리 잡음

- 약 1,750억 개의 매개변수를 가진 초대규모의 언어모델로, GPT-3를 활용하여 새로운 과제 도입 시 소수의 데이터로 해결할 수 있는 퓨샷러닝*의 성능이 우수함

* 퓨샷러닝(few-shot learning): 소수의 데이터만 사용하고, 모델 전체를 몇 건의 데이터에 맞게 업데이트 후 태스크 수행

- GPT-3은 AR(Auto-Regressive) 방법으로 학습하는 모델로 모델 구조와 데이터, 학습 과정을 포함한 전체 사전학습 과정은 GPT-2와 유사하지만, 모델의 크기, 데이터 세트의 크기, 다양성, 학습 횟수를 전반적으로 늘린 모델임

○ OpenAI는 GPT-3을 이용한 API를 제공하여 개발자들이 다양한 서비스를 창출함

- 2021년에 공개한 달리(DALL·E)는 텍스트로 설명한 입력값을 이미지로 생성하는 모델로, 동물과 사물을 의인화한 이미지, 관련 없는 개념을 결합한 이미지, 기존 이미지를 변형한 이미지를 생성하는 등 다양한 기능을 제공함
- 외국어 학습 서비스인 듀오링고(duolingo)는 프랑스어 문법 수정 기능에 GPT-3를 도입해 성능을 개선함



<그림 3-8> GPT-3를 사용한 달리(DALL·E)의 이미지 생성 예시

□ ERNIE

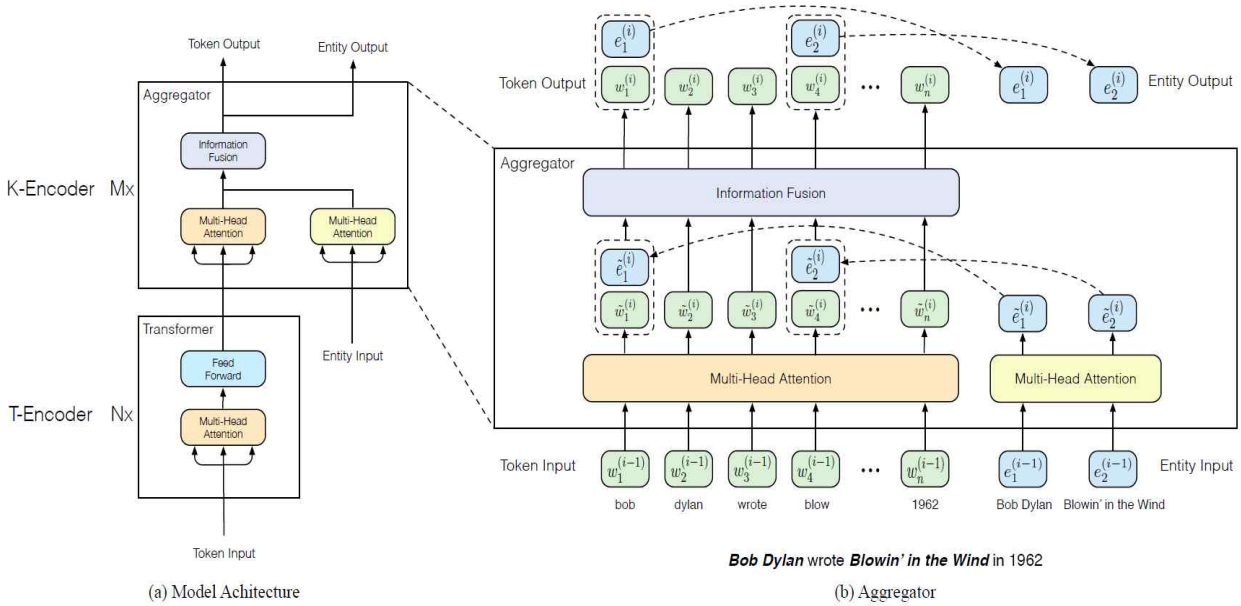
○ ERNIE는 대규모 텍스트 말뭉치와 지식그래프(KG)를 사용하여 어휘적, 의미적, 지식 정보를 동시에 학습한 언어 표현 모델임

- 지식정보를 추출하고 인코딩하기 위해 텍스트에서 명명된 엔티티(Named Entity)를 구분하고 지식그래프 안에 해당하는 엔티티와 연계함
- 토큰(Token)과 개체(Entity)를 모두 예측하기 위해 원시 텍스트와 문맥 정보를 담고 있는 지식 정보(Triple)를 함께 사용하여 임베딩을 수행함

○ ERNIE 모델의 구조는 ERNIE는 K-Encoder, T-Encoder 2개의 모듈로 구성됨

- T-Encoder에서는 주어진 토큰에서 기본적인 어휘적 문맥적 정보를 얻으며 K-Encoder는 토큰에서 유래한 추가적인 지식정보와 원시 텍스트 정보를 통합함

- 토큰과 엔티티라는 다차원적 정보를 통합된 특징(feature) 공간에 표현 가능하고, 각 시퀀스에 대해 특징을 계산하면 K-Encoder를 사용하여 언어 표현(representation) 벡터에 지식 정보를 반영함



<그림 3-9> ERNIE 모델의 구조

3. 한국어 자연어 처리

□ 한국어 자연어 처리 특징

○ 한국어는 교착어* 특징을 가진 고맥락 언어로 문맥 정보의 이해를 위한 한국어 자연어 처리의 난이도가 높고, 도메인 분야별 학습데이터가 부족함

* 교착어: 어근과 접사에 의해 단어의 기능이 결정되는 언어의 형태

- 한국의 문화적 맥락에 대한 이해가 바탕이 되는 대화가 많고, 신조어가 빠르게 등장하고 사라지기 때문에 탄력적인 대응이 필요함
- 최근 국립국어원에서 대규모 말뭉치를 제공하고 있지만, 다른 언어문화권에 비해 한국어로 이루어진 다양한 도메인을 고려한 데이터는 부족함

○ 자연어 처리 성능 향상을 위해 한국어 전처리를 위한 형태소 분석기와 언어모델이 다양하게 개발됨

- KoNLPy, Soynlp, Khaiii, Nori 등은 한국어 문장의 토큰화(Tokenization)를 위해 개발된 형태소 분석기로 처리 속도, 정확도 측면에서 성능이 향상되고 있음
- KoBERT, KoGPT-2, koGPT, hyperCLOVA는 한국어 자연어 처리를 위한 대규모 언어모델

로 텍스트 분류, 개체명 인식, 관계 예측, 텍스트 생성 등 여러 분야에서 활용되고 있음

[표 3-1] 한국어 자연어 처리 기술

기술	내용
KoNLPy	한국어 형태소 분석기를 통합하여 제공하는 파이썬 라이브러리
Soynlp	비지도 학습을 활용한 한국어 형태소 분석기
Khaiii	카카오에서 개발한 딥러닝 기반의 형태소 분석기
Nori	Elasticsearch 6.6 버전부터 한글 형태소 분석기를 Elastic사에서 공식적으로 개발해서 지원 (처리 속도 향상)
KoBERT	SKTBrain에서 BERT기법을 활용하여 개발한 한국어 언어모델
KorBERT	한국전자통신연구원(ETRI)에서 형태소분석 기반과 어절 기반의 2가지 언어모델 제공
KoGPT-2	OpenAI의 GPT-2 기반으로 만든 한국어 언어모델로 SKT 챗봇에 활용
KoGPT	카카오브레인(kakaobrain)에서 공개한 GPT-3 기반 한국어 언어모델
hyperCLOVA	네이버에서 개발한 GPT-3 기반 언어모델로 가장 많은 한국어 토큰으로 학습된 언어모델

□ 한국어 형태소 분석기

○ **(KoNLPy)** 한국어 자연어 처리를 위한 형태소 분석, 태깅을 위한 파이썬 라이브러리

- 국내에 이미 만들어져 사용되고 있는 총 5개의 형태소 분석기 Hannanum(한나눔), Kkma(꼬꼬마), Komoran(코모란), Mecab(메캅), Open Korea Text를 통합하여 제공함
- 사용자 정의 사전 생성(Meca)을 할 수 있으며 자소 분리에 우수한 성능(Komoran)을 보이는 Java 기반의 형태소 분석기로 한국어 형태소 분석 및 데이터 전처리에서 널리 사용됨

○ **(Soynlp)** 학습데이터를 이용하지 않고 텍스트에 존재하는 단어를 찾거나 단어열로 분해할 수 있는 비지도 학습 방식을 사용함

- 통계적 패턴을 이용하여 단어를 추출하기 때문에 하나의 문장보다는 동일한 집단의 문서에서 효율적으로 작동함
- 세 가지 버전의 명사 추출기를 제공하며 그 중 NewsNounExtractor에 더 많은 기능이 포함됨

○ **(Khaiii)** 카카오에서 개발한 딥러닝 기반의 형태소 분석기

- 전처리 과정의 시간을 단축시키기 위해 신경망 알고리즘 중 CNN을 사용하였고, 약 85만

문장, 1000만 코퍼스를 학습함

- 신경망 알고리즘의 결과를 유추하는 과정을 파악하기 어려운 단점을 개선하기 위해 기분석 사전과 오분석 패치라는 사용자 사전 장치를 제공함

○ **(Nori)** 엘라스틱서치(Elasticsearch) 6.6 버전부터 공식적으로 공개된 한국어 형태소 분석기

- 한국어 형태소 분석기의 하나인 Mecab의 mecab-ko-dic 사용자 사전을 추가할 수 있음
- 형태소 분석 속도 향상을 위해 분산 환경을 고려한 형태소 분석 기술을 적용함
- 엘라스틱서치를 이용한 검색엔진 구현 및 문서 색인 시 Full text search를 위한 한국어 형태소 분석에 적용할 수 있음

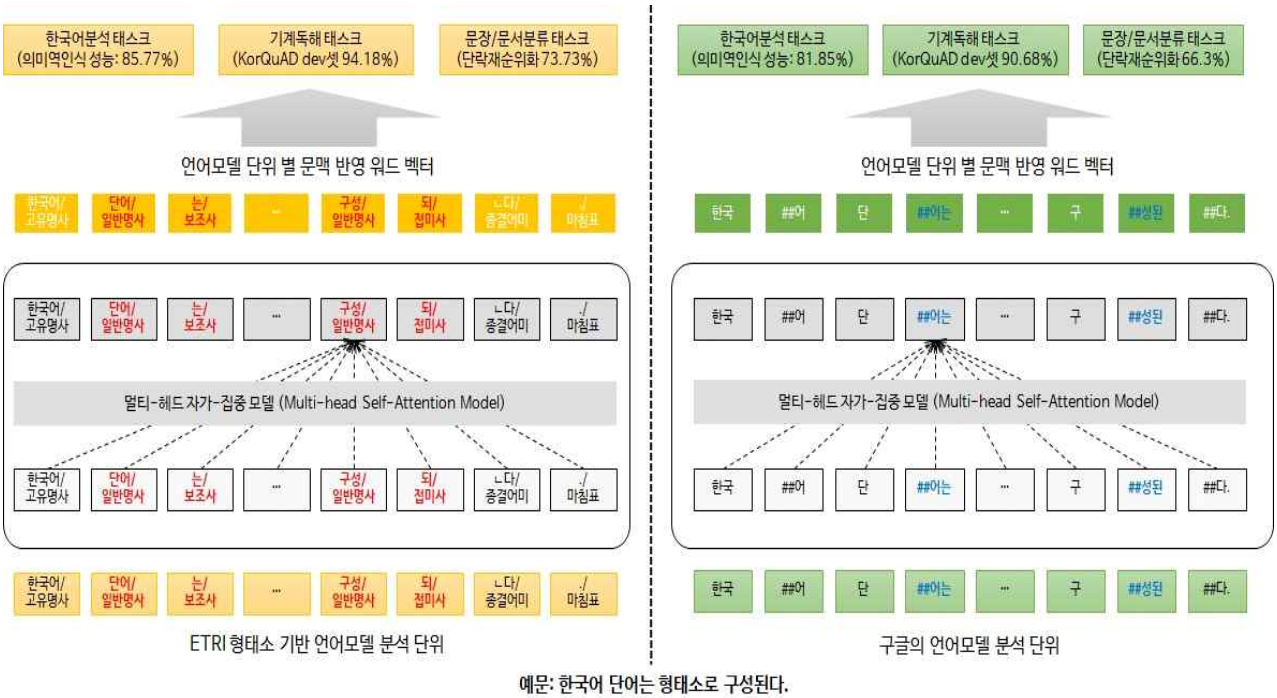
□ 한국어 언어모델

○ **(KoBERT)** SKTBrain은 2019년에 기존 구글의 BERT의 한국어 성능 한계를 극복하기 위한 모델인 KoBERT를 발표함

- 수백만 개의 한국어 단어들을 학습시키고 한국어의 불규칙한 언어 변화의 특성을 반영하기 위해 데이터 기반 토큰화 기법을 적용함
- 대량의 데이터를 빠른 시간에 학습하기 위해 링 리듀스(ring-reduce) 기반 분산 학습 기술을 사용하여, 십억 개 이상의 문장을 다수의 머신에서 빠르게 학습함
- 파이토치(PyTorch), 텐서플로(TensorFlow), ONNX, MXNet을 포함한 다양한 딥러닝 API를 지원하여 다양한 분야의 자연어 이해 서비스 확산에 기여함

○ **(KorBERT)** 한국전자통신연구원(ETRI)이 과학기술정보통신부와 정보통신기획평가원(IITP)의 프로젝트인 엑소브레인 사업에 필요한 KorBERT 언어모델을 개발함

- 신문기사와 백과사전 등 23GB의 대용량 텍스트를 대상으로 약 47억 개의 형태소를 사용하여 학습한 2가지 모델을 공개함
- 형태소 분석 기반의 언어모델은 한국어의 특성을 반영하여 어절 기반 언어모델보다 우수한 성능을 보이고, 어절 기반 언어모델은 구글과 유사한 방식으로 구성하여 형태소 분석을 수행하지 않아도 된다는 장점이 있음
- 한국어 분석·기계독해·문서 분류 등 다양한 태스크에 활용 가능하고, 구글이 배포한 한국어 언어모델과 비교한 결과 평균 4.5% 우수한 것으로 평가됨



<그림 3-10> KorBERT와 BERT 모델 비교

○ **(KoGPT-2)** SKT-AI는 2020년에 AWS와 협력해 openAI의 GPT-2 기반을 둔 최초의 한국어 언어모델인 KoGPT-2를 발표함

- 소규모 데이터만으로도 대규모 언어모델 학습 같은 자연어 처리 업무를 빠르게 처리할 수 있으며 실제 SKT 챗봇의 어투를 자연스럽게 만드는 데 활용되었고, 2021년에 성능을 개선하여 KoGPT-2 모델 2.0을 발표함

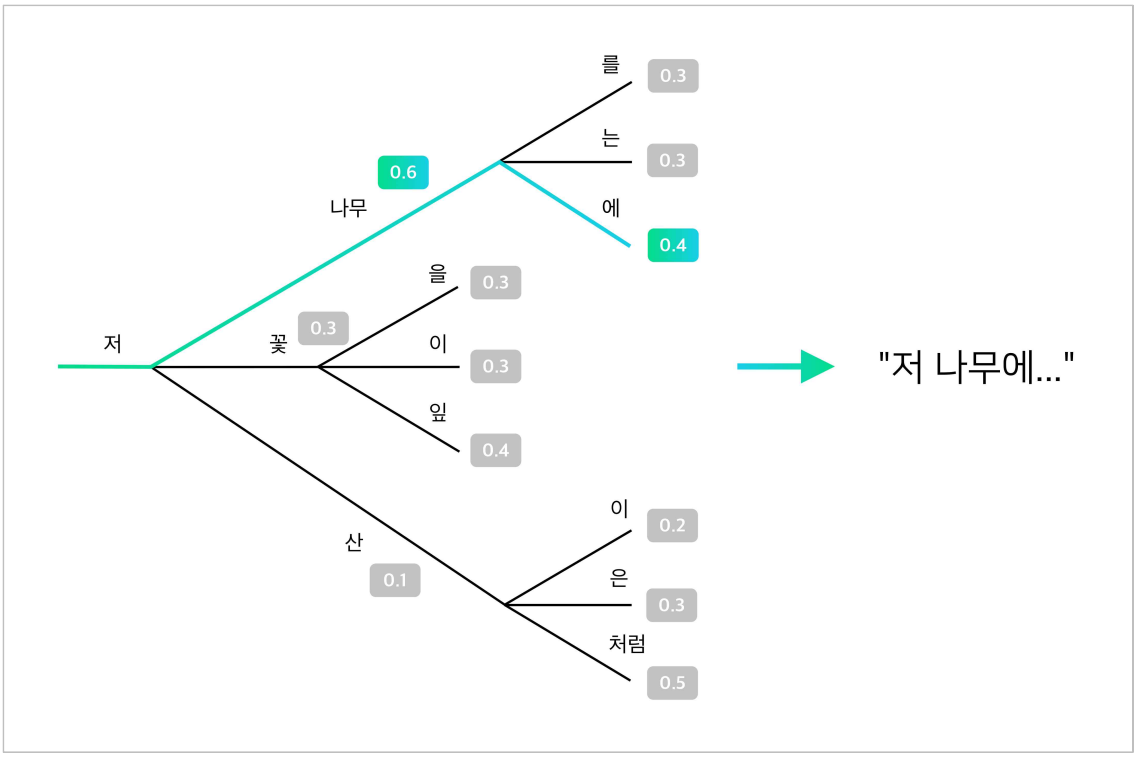
○ **(KoGPT)** 2021년 카카오브레인(kakaobrain)에서 공개한 GPT-3 모델 기반 한국어 처리 언어모델

- 60억 개의 매개변수와 2,000억 개 토큰의 한국어 데이터로 구축되었으며, 현재는 AI 커뮤니티를 위한 연구용으로 전처리를 하지 않은 원시 데이터로 학습됨
- 한국어를 사전적, 문맥적으로 이해하여 주어진 문장의 긍정과 부정을 판단하거나 긴 문장을 한 줄 요약하는 기능, 맥락에 따라 자동으로 텍스트 생성하는 기능 등 다양한 자연어 처리 기술을 제공함

○ **(HyperCLOVA)** 2021년 네이버에서 개발한 한국어 기반의 GPT-3 모델로 대규모 한국어 언어모델

- 네이버에서 서비스하는 블로그, 카페, 뉴스 기사와 댓글, 지식IN 데이터와 국립국어원 모두의 말뭉치, 위키백과 데이터를 전처리한 대규모 한국어 데이터를 구축함
- 한국어 발음 체계를 고려하여 형태소를 기반으로 구축한 데이터를 토큰화하여 총 5,618억 개의 토큰 사용함

- 네이버의 검색, 쇼핑 등 내부의 다양한 서비스에 활용하거나, 스타트업, 크리에이터 등을 지원하기 위한 기술도 제공함



<그림 3-11> hyperCLOVA의 텍스트 생성 과정

4. 응용서비스

□ 개요

○ 최근 자연어 처리에 대한 딥러닝 기술의 활용이 급증하면서, 다양한 분야에서 자연어 처리 기술이 활용되고 있음

- 자연어 처리는 학문 분야의 구분 없이 대부분의 학문 분야, 서비스 영역에서 사용될 수 있어 활용도가 높고, 뉴스, 보고서, SNS 등 비정형 텍스트 데이터를 활용한 빅데이터 분석에 활용함

○ 자연어 처리 기술은 정보검색, 질의응답(QA), 문서 요약, 문서 분류, 이미지 인식, 음성 인식, 기계번역, 지식그래프 등에 적용되고 있음

* 자연어 처리 기술의 응용서비스 분야에 대한 설명은 [표 3-2]를 참고

[표 3-2] 자연어 처리 응용서비스 분야

응용서비스 분야	설명
정보검색	웹 문서, 이미지, 동영상 데이터로부터 입력된 질의와 관련된 정보를 찾는 기술
질의응답(QA)	컴퓨터가 자연어 질문의 맥락을 파악하여 자동으로 응답을 제공하는 기술
문서 요약	주어진 문서에서 중요한 내용을 추출하거나, 새로운 문장을 생성해 요약문을 만드는 기술
문서 분류	텍스트의 내용을 기준으로 주어진 범주(Category)에 따라 나누어 할당하는 기술
이미지 인식	이미지로부터 인물, 객체, 텍스트 등을 인식하는 기술 자연어 처리 분야에서는 디지털화한 문서에서 텍스트 데이터를 인식하는 데 주로 사용됨
음성 인식	사람이 말하는 음성 언어를 컴퓨터가 해석하여 텍스트 데이터로 전환하는 기술
기계번역	컴퓨터를 이용하여 특정 언어의 단어, 문장, 문서를 다른 언어로 번역하는 기술
지식그래프	객체 사이의 관계를 의미적으로 표현한 그래프 형태의 데이터

□ 정보검색

○ 정보검색은 구조화되지 않은 문서나 텍스트에서 사용자의 질의에 맞는 정보를 추출하기 위한 기술을 의미함

- 정보검색 모델은 사용자가 검색한 키워드에 적합한 문서를 선택하고 문서의 우선순위 판단
- 효율적인 정보검색을 위해 문서 인덱싱, 페이지 랭킹 알고리즘, 문서 요약 알고리즘과 자연어 검색에 대응하기 위한 자연어 처리 알고리즘 활용됨

○ 정보검색 시스템은 데이터 수집, 데이터 저장, 검색 알고리즘 3가지로 구성됨

- **(데이터 수집)** 대량의 데이터를 수집하고 가공하는 과정으로, 인터넷상의 웹사이트를 크롤링* 또는 스크래핑**하는 작업이며, 검색 결과의 품질을 유지하기 위해 지속적으로 수집해야 함

* 웹 크롤링: 웹 페이지와 페이지 내 다른 웹 페이지의 주소를 통해 연결된 문서를 수집하는 기술

** 웹 스크래핑: 웹 페이지의 내용을 긁어서(scrap) 저장하는 기술

- **(데이터 저장)** 수많은 문서를 효율적으로 저장하고 검색하기 위해 역 색인(inverted index)을 수행함. 역 색인을 통해 문서 단위가 아닌 단어와 단어가 출현한 문서를 저장하여 키워드에 맞는 데이터를 빠르게 검색할 수 있음
- **(검색 알고리즘)** 사용자가 원하는 정보를 얻을 수 있도록 질의를 분석해 키워드를 추출하고, 역 색인한 데이터에서 키워드가 포함된 문서를 반환하는 알고리즘을 구축해 사용함

○ 구글, 네이버 등 국내외 포털 사이트는 은 웹 페이지, 이미지, 비디오, 기사 등 다양한 유형의 웹 자원 검색을 위한 검색 엔진(Search engine)을 구축함

- 포털 검색 엔진이 요구하는 여러 조건을 최적화된 값으로 설정하기 위한 검색 엔진 최적화(Search Engine Optimization)가 필요함
- 도서관은 정보검색 기술을 초기에 도입하여 저자명, 제목, 키워드 등 서지 정보를 활용한 도서 검색 시스템을 구축, 점차 디지털 도서관으로 발전하여 검색이 가능한 범위를 확장하고 있음

□ 질의응답(QA)

○ 질의응답은 컴퓨터가 사용자로부터 받은 질문의 맥락을 파악하여 자연어로 대답을 제공하는 시스템이며, 질문에 대한 응답을 제공하는 방식에 따라 QA 방식을 세분화할 수 있음

- 오픈 도메인(Open-Domain) QA는 위키피디아와 같이 특정 도메인에 국한되지 않고 다양한 주제의 질의에 응답을 제공함
- 커뮤니티(Community) QA는 Q&A 포럼 또는 게시판을 활용해 응답을 제공하는 방식으로, 스택 오버플로우(Stack Overflow)나 쿼라(Quora)가 대표적인 사례임
- 지식베이스(Knowledge Base) QA는 체계화된 지식 구조인 지식그래프(Knowledge Graph)를 기반으로 사용자의 질문에 대해 응답을 제공함



<그림 3-12> 구글 지식베이스 QA(왼쪽)와 네이버 오픈 도메인 QA(오른쪽)

○ 질의응답 시스템은 질의문장을 분석해 유용한 쿼리를 생성하고, 쿼리를 통해 지식베이스에서 문장을 추출하여 가장 적합한 문장을 응답으로 선택함

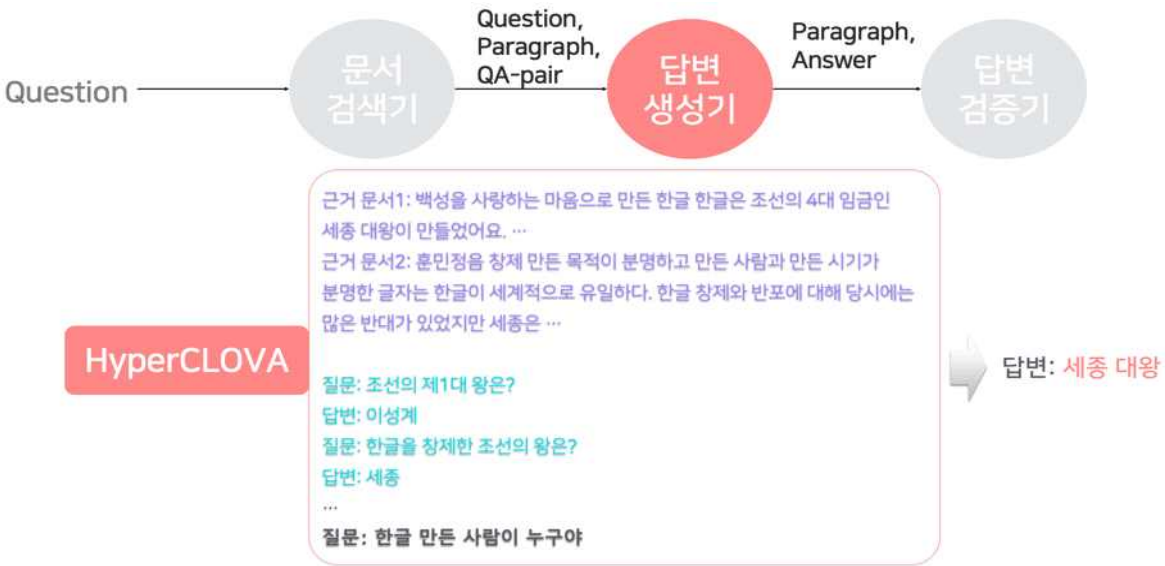
- 질의응답 시스템은 지식베이스 또는 데이터베이스에서 문장을 추출하거나, 요약하여 질의에 대한 응답을 생성하기 때문에 데이터의 품질과 정확성이 중요함
- 추출 방식을 통한 질의응답은 주어진 문맥에서 응답 문장을 추출하여 제공하고, 요약을 통한 질의응답은 주어진 문맥을 기반으로 텍스트를 생성하여 응답함

○ 최근 BERT 또는 GPT와 같은 트랜스포머 기반 언어모델의 발전에 따라 질의응답의 성능이 비약적으로 상승함

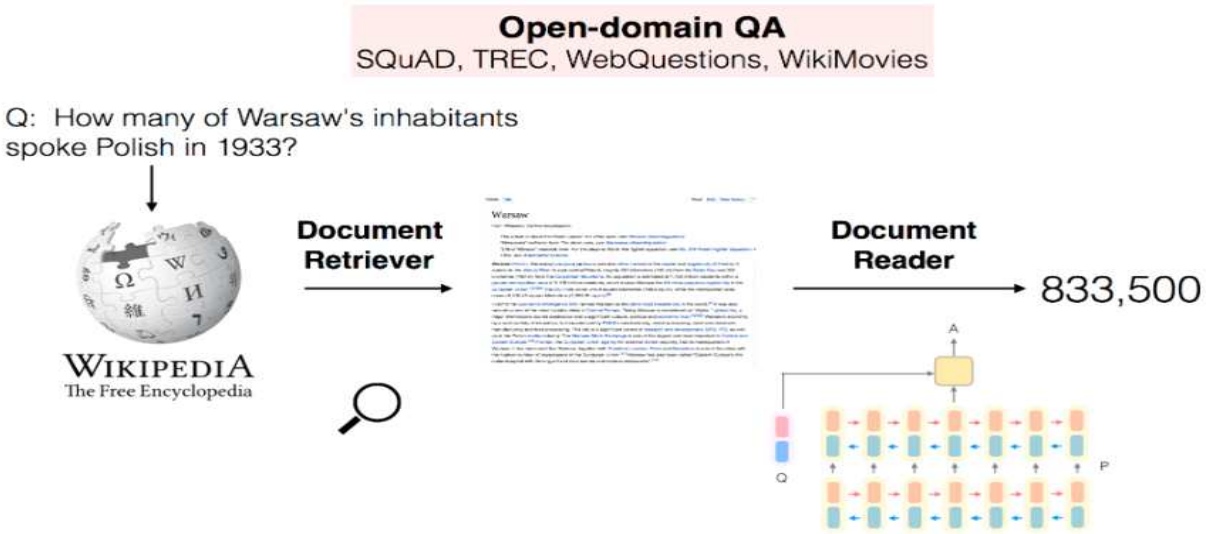
- 딥러닝을 이용하여 학습된 문서로부터 응답 텍스트를 추출하거나, 생성하여 질문에 대한 응답을 제공하고, RNN, CNN, Transformer 기반 모델 등 다양한 모델이 사용됨

○ 네이버의 인공지능 비서 '클로바(CLOVA)'와 페이스북 '오픈도메인 QA'는 질의응답 서비스를 위해 자체 언어모델을 개발해 적용함

- 클로바(CLOVA)는 네이버가 개발한 인공지능 언어모델인 '하이퍼클로바(HyperCLOVA)'를 적용하여 사용자의 질문에 전문(expert) 지식으로 응답하는 전문 질의응답 서비스를 제공함
- 페이스북이 개발한 'DrQA'는 오픈도메인 QA에 적용되는 시스템으로, 위키피디아를 지식베이스로 사용함



<그림 3-13> 클로바 질의응답 과정 예시 (전동현, 2021)



<그림 3-14> DrQA의 질의응답 과정

□ 문서 요약

○ 문서 요약은 원본 문서의 핵심적인 내용을 자동으로 요약하여 제공하는 기술으로, 문서의 길이, 문체, 구문론 등의 변수를 고려함

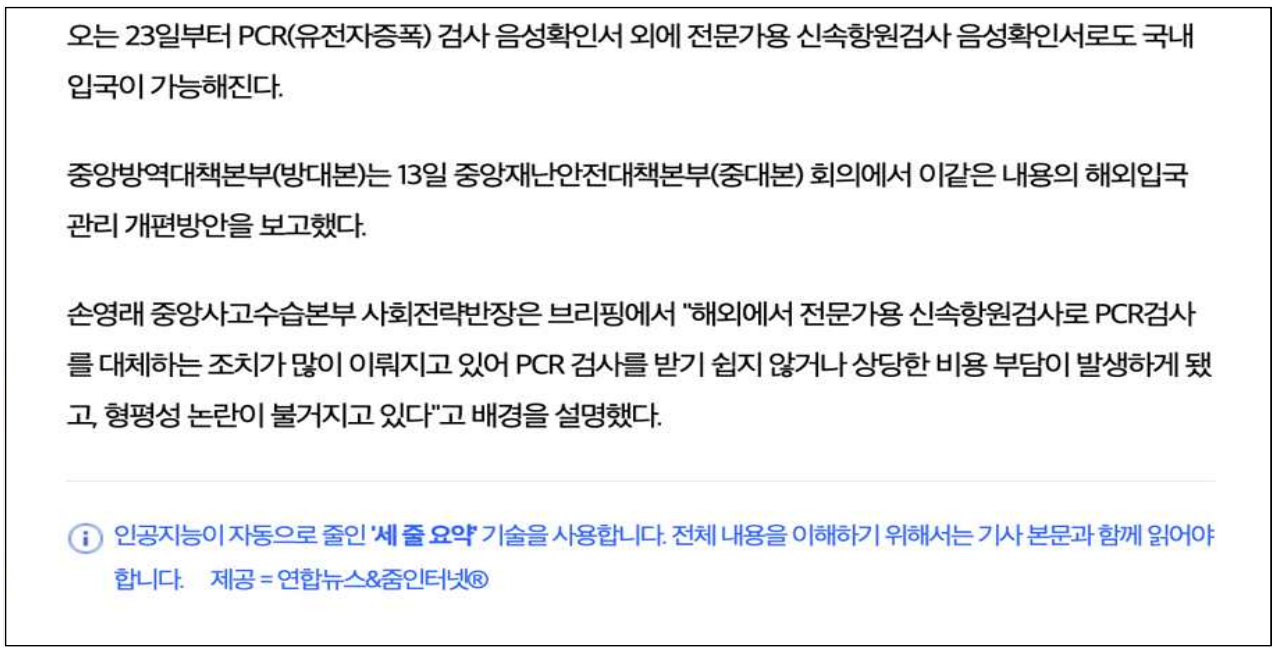
- 텍스트 요약 기술은 텍스트에서 핵심 문장을 추출, 적절한 순서로 정렬하여 요약문을 완성하는 '추출 요약 방식(Extractive summarization)'*과 텍스트의 문맥을 파악하여 요약문을 직접 생성하는 '생성 요약 방식(Abstractive summarization)'**으로 구분함

* 추출 요약: 원문에서 중요한 핵심 문장 또는 단어구를 추출해 요약문을 생성하는 기술로, 대표적인 딥러닝 알고리즘으로 BertSUM, KoBertSUM이 모델이 있음

** 생성 요약: 토큰 간의 관계를 계산해 핵심 문맥을 반영한 새로운 문장을 생성해 요약하는 기술이며, 딥러닝 알고리즘으로 BARTSUM, KoBARTSUM 모델이 있음

○ 인터넷을 통해 길이가 긴 문서를 제공하는 언론사, 뉴스 포털은 이용자가 원문의 핵심 내용을 빠르게 파악할 수 있도록 문서 자동 요약을 서비스하고, 관련 API를 제공

- 국내 줌인터넷과 연합뉴스가 공동 개발한 '뉴스 세 줄 요약'과 문서 요약기인 'WIGO Recap', 네이버 뉴스 요약봇은 'CLOVA Summary' 텍스트 요약 서비스 제공
- Microsoft Azure Cognitive Service LUIS는 사용자의 대화 또는 자연어 텍스트에서 전체적인 의미를 예측하고, 자세한 정보를 추출하는 클라우드 기반 대화형 API
- AYLIEN Text API는 반환되는 문장의 수를 조절 가능한 옵션 제공
- MeaningCloud는 Summarization를 활용하여 텍스트에서 가장 관련성이 높은 문구를 찾고, 요약 결과에 포함할 구문의 수 지정할 수 있음


오는 23일부터 PCR(유전자증폭) 검사 음성확인서 외에 전문가용 신속항원검사 음성확인서로도 국내 입국이 가능해진다.

중앙방역대책본부(방대본)는 13일 중앙재난안전대책본부(중대본) 회의에서 이같은 내용의 해외입국 관리 개편방안을 보고했다.

손영래 중앙사고수습본부 사회전략반장은 브리핑에서 "해외에서 전문가용 신속항원검사로 PCR검사를 대체하는 조치가 많이 이뤄지고 있어 PCR 검사를 받기 쉽지 않거나 상당한 비용 부담이 발생하게 됐고, 형평성 논란이 불거지고 있다"고 배경을 설명했다.

인공지능이 자동으로 줄인 '세 줄 요약' 기술을 사용합니다. 전체 내용을 이해하기 위해서는 기사 본문과 함께 읽어야 합니다. 제공 = 연합뉴스&줌인터넷®

<그림 3-15> 연합뉴스의 세 줄 요약 예시

□ 문서 분류

○ 문서 분류는 텍스트 데이터(문장, 문단, 문서 등의 단위)를 사전에 정의한 주제, 장르, 감성, 용도, 의도 등의 분류 기준에 맞게 자동으로 분류하는 기능

- 주제 분류뿐 아니라 감성(sentiment), 장르(genre), 개인화(personalization) 기반 분류체계를 활용 가능하며 기술적으로는 실용화 단계이며, 딥러닝 기술의 적용으로 성능이 훨씬 증가할 것으로 주목받고 있음
- 분류를 위해 텍스트 데이터를 기계가 읽을 수 있도록 숫자로 표현하는 Bag-of-Words(BoW)*와 Word2Vec**과 같은 워드 임베딩 기법을 사용함

* Bag-of-Words(BoW): 텍스트(문단, 문장)에서 순서와 상관없이 고유한 단어의 빈도수를 벡터로 표현하는 방법

** Word2Vec: 인공신경망을 사용한 추론 기반 임베딩 알고리즘

○ Amazon, IBM과 국내 Kakao, 네이버 등은 감성 분류, 주제 분류, 메타데이터 추출 등, 텍스트 분류를 쉽게 수행할 수 있는 다양한 분류 API를 제공함

- CLOVA Sentiment API는 네이버의 빅데이터와 형태소 분석기를 활용하여 한국어에 최적화된 사전 훈련을 바탕으로 감성분석을 수행할 수 있음
- Google Cloud Natural Language API는 사전 정의된 700개 이상의 주제로 자동 분류, 감성분석 서비스를 제공함
- Amazon Comprehend는 문서 내에서 인사이트를 추출할 수 있는 기술이며 UTF-8 형식의 텍스트 파일과 PDF, Word 문서와 같은 반구조화된 문서를 처리함

- IBM Watson Natural Language Understanding을 이용하여 문서 내에서 개념, 엔티티, 키워드, 카테고리, 감성, 감정, 관계 등 문서와 관련된 메타데이터를 추출할 수 있음

□ 이미지 인식

○ 이미지 인식은 이미지 데이터로부터 인물, 객체, 텍스트를 인식하는 기술이며, 이미지 내의 특정 사물을 분류하거나, 추출하거나, 식별하여 분할할 수 있음

- 최근 딥러닝 기반 이미지 인식 기술이 주목받고 있으며, 학습데이터의 절감을 위해 자기지도학습 또는 액티브 러닝(Active learning)*을 이미지 인식에 적용하는 연구가 이뤄짐

* 액티브 러닝: 무작위로 많은 데이터를 어노테이션한 후 학습하는 것이 아닌, 학습에 크게 기여할 수 있는 데이터를 선별해 어노테이션하고, 점진적으로 정확도를 높이기 위해 데이터를 늘려가는 방식

○ 자연어 처리 분야에서 주로 사용되는 이미지 인식 기술은 광학 문자 인식(Optional Character Recognition, OCR)이며, 사람이 쓰거나 이미지에 인쇄된 문자를 스캔하여 기계가 읽을 수 있는 문자로 변환하는 데 사용됨

- 머신러닝 기술을 활용해 이미지에서 문자 영역을 찾고, 문자 영역에서 텍스트를 인식하여 기계가 처리할 수 있는 형태로 변환하는 기술
- 도서관, 박물관 등이 소장한 자료를 디지털화하는 과정에서 사용, 최근 이미지 자료의 실시간 번역, 영수증 내용 추출 등 활용 분야를 확장할 수 있음

○ OCR은 이미지 획득, 전처리, 패턴 매칭 방법 또는 특징 추출 방법을 기반으로 하는 텍스트 인식, 후처리의 순서로 작동함

- **(이미지 획득)** 스캔한 문서를 이진데이터로 변환하고, 분석을 통해 밝은 부분을 배경, 어두운 부분을 텍스트로 분류함
- **(전처리)** 파일의 기울기를 보정하고, 텍스트가 반듯하게 보일 수 있도록 보정함
- **(텍스트 인식)** 패턴 매칭 방법은 문자 이미지를 따로 추출하여 저장된 데이터와 비교해 인식하는 방법으로 저장된 데이터와 폰트, 크기가 비슷한 경우에만 작동, 특징 추출 방법은 문자 이미지를 선, 방향, 교차 등과 같은 특징으로 분해하여 일치하는 데이터를 찾아내는 방법임
- **(후처리)** 추출한 텍스트 데이터를 컴퓨터가 읽을 수 있도록 변환하여 저장함

○ OCR 기술은 문서 뷰어를 제공하는 IT 기업, 정보검색을 위한 포털 사이트, 오픈소스 라이브러리를 통해 손쉽게 이용할 수 있음

- 데이터 분석 오픈소스인 텐서플로우(Tensorflow)는 손글씨 데이터세트를 제공하여 직접 OCR 프로그램을 만들 수 있는 환경을 제공함

- 어도비(Adobe)가 개발한 어도비 아크로뱃(Adobe Acrobat)은 PDF 파일에서 글자를 인식하는 OCR 서비스를 제공하고 있음
- 네이버의 CLOVA OCR은 영수증, 신용카드, 명함 등 사용자의 이미지로부터 필요한 정보만 추출하고, 추출 결과를 JSON 형태로 제공함



<그림 3-16> CLOVA OCR 예시

□ 음성 인식

○ 음성 인식(Speech-to-Text, STT)은 사람의 음성을 컴퓨터가 해석해 텍스트 데이터로 전환하는 기술이며, 구어체 형식의 음성을 텍스트 형식으로 변환함

- 목소리 인식(Voice recognition)은 음성 인식과 유사하지만, 목소리 인식은 개별 사용자의 목소리를 식별하는 데 사용된다는 차이점이 있음
- 대표적인 응용서비스 사례는 음성 어시스턴트와 모바일 앱의 음성-텍스트 변환 기능으로, 마이크로소프트 Cortana, 애플 Siri, 구글 Assistant, 아마존 Alexa 등 다양한 인공지능 비서에 적용됨

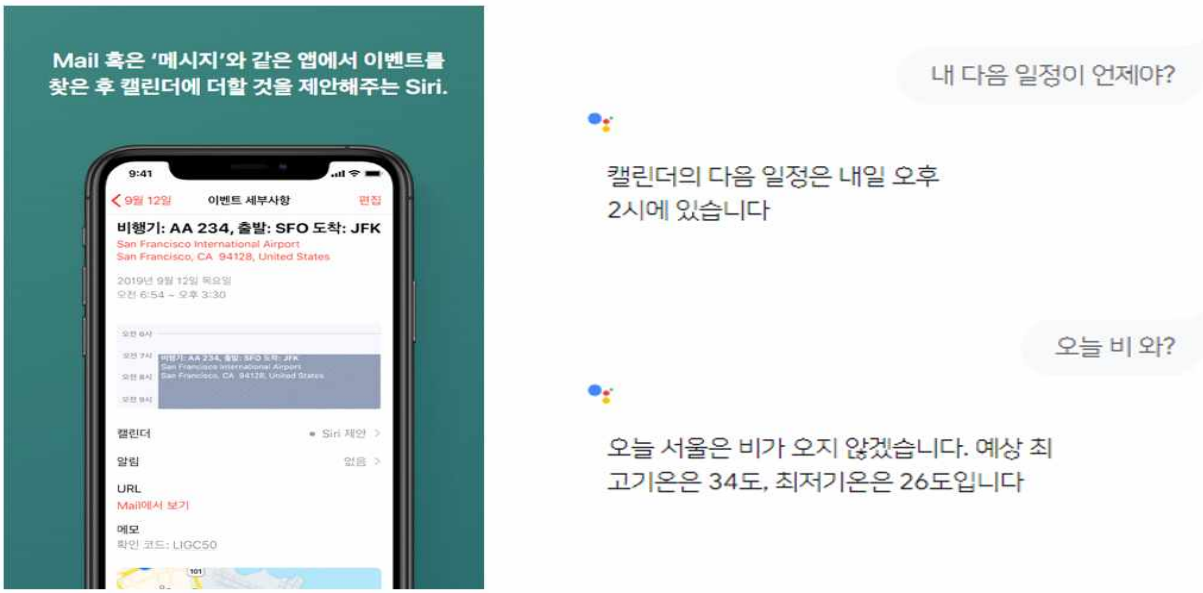
○ STT 대표 알고리즘인 HMM(Hidden Markov Model)은 다양한 음성을 통계적으로 모델링하여 음향 모델을 구축하고, 말뭉치 수집을 통해 언어모델을 구성함

- 한 상태의 확률은 직전 상태에 의해서만 결정된다는 Markov 체인 모델을 기반으로 구축됨
- 품사 태깅(part-of-speech tag) 같은 순차적인 데이터를 다루는 데 유용하고, 음성 인식 내의 시퀀스 모델로 활용할 수 있음

- 시퀀스 모델을 통해 단어, 음절, 문장 등의 각 유닛에 레이블을 지정하고, 레이블은 주어진 입력으로부터 최적의 레이블 시퀀스를 판별함

○ Google, IBM, Amazon은 오디오 데이터를 텍스트로 변환하는 API, CLI/SDK 기반 인터페이스와 서비스를 제공하고 있음

- Google Cloud Speech-to-Text API는 12개 이상의 언어와 방언을 지원하여 오디오 파일이나 스트리밍 오디오를 텍스트로 변환함
- IBM Watson의 Speech to Text는 다양한 형식의 오디오 파일을 텍스트로 변환하는 기능 지원함
- Amazon Transcribe는 일반 음성 데이터뿐만 아니라 저음질의 전화 음성 데이터 모두 지원하며 콘솔 기반과 CLI/SDK 기반의 인터페이스 제공함



<그림 3-17> 대화형 인공지능 SIRI(왼쪽) 대화형 인공지능 구글 어시스턴트(오른쪽)

○ Kakao는 사용자가 직접 단어와 패턴을 추가할 수 있는 커스텀 API를 제공하고, 타 기업의 제품에 음성 인식 기술을 적용하고 있음

- Kakao 커스텀 STT API는 각 기업이 직접 클라우드에 접속해 고객과 산업 특성에 맞는 모델을 구축하기 위해 직접 단어와 패턴을 추가하는 기능을 제공함
- Kakao STT API는 각 기업이 현대자동차는 Kakao mini에 탑재된 Kakao I 음성 인식 기술을 적용한 차량을 확대할 예정임



<그림 3-18> Kakao I 적용 예시

□ 기계번역

○ 기계번역은 인간이 사용하는 자연언어를 다른 언어로 자동 번역하는 기술이며, 자동 번역과 번역 지원 두 가지 유형으로 구분됨

- **(자동번역)** 인간의 개입이 최소화된 기계적인 번역을 수행하며, 대표 사례로 인터넷을 통해 사용되는 번역기가 있음
- **(번역 지원)** 전문 번역가가 번역 작업을 효율적으로 수행할 수 있도록 도움을 주는 전문 소프트웨어

○ 최근 BERT와 같은 인코더-디코더(encoder-decoder) 어텐션(attention) 기반 아키텍처의 발전으로 기계번역의 성능이 크게 향상됨

- 기계번역 분야에서 주로 사용되는 언어모델은 특정 언어의 단어 시퀀스(sequence)를 다른 언어의 단어 시퀀스로 변환하는 시퀀스-투-시퀀스(sequence-to-sequence) 과제를 수행함
- 시퀀스-투-시퀀스 과제를 수행하는 모델은 대부분 번역하고자 하는 언어의 단어 시퀀스를 압축하는 인코딩(encoding)과 인코더가 보낸 단어 시퀀스 정보를 받아 다른 언어의 단어 시퀀스를 생성하는 디코딩(decoding)으로 구성됨

○ 네이버의 파파고 번역은 인공신경망 기반의 NMT(Neural Machine Translation)모델을 적용한 번역 서비스를 제공함

- 입력된 텍스트의 언어를 감지하는 언어 감지 API, 한글로 된 이름을 로마자 표기로 변환하는 한글 인명-로마자 변환 API 제공하고 있음
- 국회도서관은 네이버 파파고와 공동 개발한 AI 일본법률 자동번역 서비스를 바탕으로 일본

법률 체계와 법률용어에 특화된 서비스 제공함

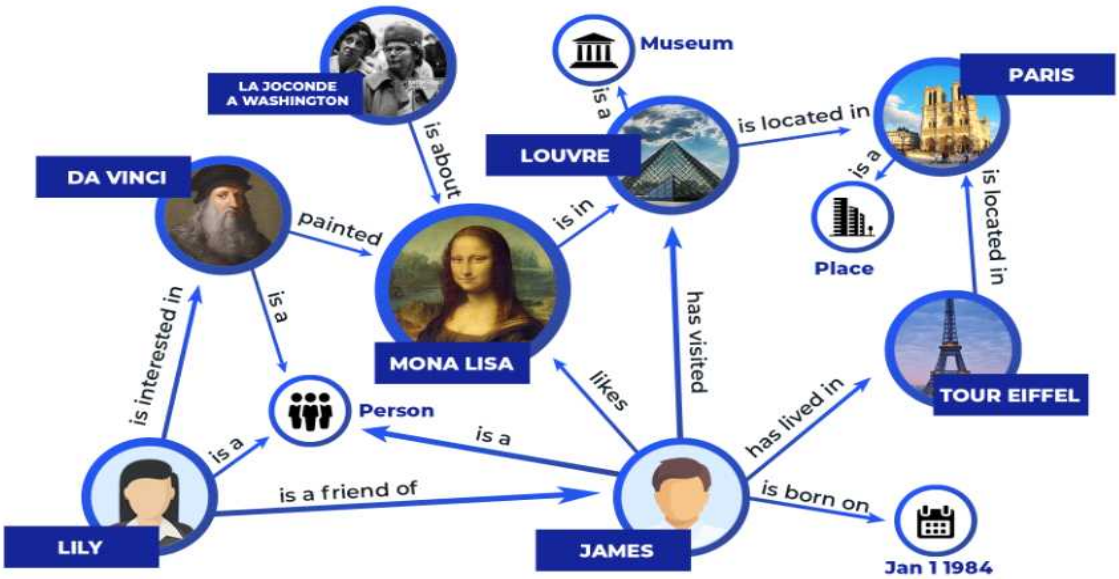


<그림 3-19> 파파고의 기계번역 예시

□ 지식그래프

○ 지식그래프는 실제 세계에 존재하는 객체(object), 이벤트, 상황 또는 추상적 개념 사이에 존재하는 연관성을 보여주는 표현 기법임

- 각 노드는 실제 세계의 객체(entity)를 의미, 노드 사이의 연관성은 방향성이 있는 선(Edge)으로 표현됨



<그림 3-20> 지식그래프 예시

○ 웹 자원을 기계가 이해할 수 있도록 시맨틱 웹과 링크드 데이터 기술을 사용하여 검색 결과를 향상하는 데 사용됨

- 가트너에서 발표한 2022년 IT 전략 기술 트렌드 중 'DataFabric'*의 핵심 요소로 지식그래프, 능동적 메타데이터 분석, 데이터 카탈로그를 언급함

* DataFabric: 이기종 시스템에 분산된 데이터를 원활하게 공유하고 관리하기 위한 체계

- 인공지능과 결합한 지식그래프 기술은 데이터 사이의 의미적인 관계를 바탕으로 기계가 개체 사이의 맥락을 이해함으로써 자연어 처리, 자율주행, 의사결정 등 다양한 분야에 활용됨

○ 디비피디아(DBPedia)는 위키백과의 대규모 데이터에서 구조화된 메타데이터를 추출하여 각 자원의 관계를 의미적으로 정의한 지식베이스를 구축함

- 2012년 구글은 최초로 디비피디아와 Freebase 기반의 지식그래프를 도입하고, 웹 자원을 기계가 이해할 수 있도록 시맨틱 웹과 링크드 데이터 기술을 사용해 검색 결과를 향상함
- 각국의 주요 도서관은 서지 메타데이터를 활용하여 지식그래프를 구축하고, 이용자가 모든 정보에 접근 가능한 엔드포인트(Endpoint) 제공하고 있음

제 4 장 국외 주요도서관 사례

1. 개요

□ 조사 대상

○ 인공지능을 적용한 도서관 서비스를 개발하거나, 관련 로드맵 또는 비전을 발표한 국외도서관 7개를 조사함

- 미국 의회도서관, 일본 국립국회도서관, 스웨덴 국립도서관, 네덜란드 국립도서관은 도서관 자체 연구소에서 도서관에 적용할 수 있는 머신러닝 기술을 개발하고, 관련 보고서와 소스 코드를 개방하고 있음
- 프랑스 국립도서관, 노르웨이 국립도서관, 핀란드 국립도서관은 풍부한 디지털 자료를 기반으로 디지털 컬렉션에 대한 접근성을 높이기 위한 프로젝트를 진행함

※ 7개 도서관의 인공지능 적용 사례는 [표 4-1]를 참고

□ 국외 주요도서관

○ **(미국)** 미국 의회도서관은 세계 최대 규모의 도서관으로, 미국의회 의원 활동과 연구 지원을 제공하는 입법보조기관임

- 미국 의회도서관 온라인 카탈로그(Library of Congress Online Catalog)와 다양한 유형의 자료를 디지털 컬렉션으로 구축해 온라인으로 검색, 탐색할 수 있도록 서비스하고, 특정 주제에 대한 디지털화 프로젝트, 디지털 아카이브 구축 등 다양한 온라인 서비스를 위한 프로젝트를 수행함
- 미국 의회도서관 연구소(LC Labs)는 머신러닝, AI를 기반으로 도서관 소장자료에 대한 접근, 검색, 참여를 향상하기 위한 프로젝트 진행과 보고서를 발간해왔음

○ **(일본)** 일본 국립국회도서관은 일본 국회의 입법 활동을 보좌하기 위한 목적으로 설립된 기관이며, 법정 납본 도서관으로 일본 유일의 국립도서관으로서 일반 국민에 대한 서비스를 제공하고 있음

- 2021년, 향후 5개년 도서관의 모든 분야에서 지적 활동과 정보 자원의 연결성을 높이기 위한 '디지털 시프트(Digital Shift)' 발표. 광범위한 정보 자원의 사용을 위한 보편적 접근성 제고와 영구적인 국가 디지털 정보 인프라 개발을 위한 프로젝트를 수행함
- 일본 국립도서관 차세대 시스템 개발 연구실(NDL Labs)은 도서관이 보유하고 있는 디지털화된 자료, 서지 데이터를 제공하고, 도서관 데이터를 활용하기 위한 기술과 소프트웨어를 개발하고 있음[1)]

1) https://lab.ndl.go.jp/about/aboutus_en/

○ **(프랑스)** 프랑스 국립도서관은 미테랑 도서관, 리슐리외 도서관, 아스널 도서관, 오페라 도서관과 박물관을 통칭하는 기관이며, 국가 기록 유산을 수집, 목록화, 보존해 풍부한 정보와 함께 전달하는 것이 목적임

- 1997년 온라인 사용자를 위한 디지털 도서관 Gallica가 구축되어 1,200권의 도서, 500여 개의 라디오 파일, 8만 건의 이미지 자료 검색이 가능함
- 2020년 발표한 디지털 다이어그램은 디지털 컬렉션 구축과 배포, 데이터의 역할, 인공지능 적용을 위한 프랑스 국립도서관의 비전으로, 디지털 컬렉션, 데이터, 인프라, 컬렉션 처리, 가능성과 예측, 기술, 자원 탐색과 공유, 보고와 카탈로그 영역으로 구성됨

○ **(스웨덴)** 스웨덴 국립도서관은 유럽에서 가장 빠르게 디지털화 작업을 수행함

- 국립도서관 디지털 컬렉션과 인쇄자료의 온라인 목록을 구축해 검색을 지원하고 있음
- 스웨덴 내 대학교, 연구도서관, 공공도서관을 비롯한 500개 스웨덴 도서관이 소장하고 있는 자료를 통합적으로 검색하기 위한 온라인 종합 목록(LIBRIS)과 다양한 데이터베이스에서 제공하는 자료를 검색할 수 있는 통합 데이터베이스 검색 시스템 구축하는 등 다양한 출처의 데이터에 대한 접근을 지원하고 있음
- 디지털 인문학과 사회과학 연구를 위한 연구소(KBLab)[2]를 통해 국립도서관의 컬렉션을 구조화되고 정량적인 형태로 접근할 수 있도록 지원하고, 도서관의 텍스트 데이터와 시청각 자료의 분석을 위한 머신러닝 기술, AI 모델 등을 테스트하고 연구 협력 체계를 구축함

○ **(핀란드)** 핀란드 국립도서관은 공공교육의 중심이며, 각 도서관에 대해 국가적 예산과 기술 적용을 지원하는 주체임

- 연구 커뮤니티를 지원하고, 연구출판물, 디지털 인문학을 위한 오픈 사이언스 생태계 구축을 위한 수집 서비스를 개발. 특히, 도서관 부문을 위한 연구 프로젝트에 적극적으로 참여하며 생산된 데이터는 FAIR 데이터 원칙을 기반으로 관리하고 있음
- 국내외 문화유산, 도서관 기관의 협력과 서비스 호환성을 위해 머신러닝, AI를 활용한 디지털 솔루션을 개발하고, 오픈소스 코드, 오픈 인터페이스 기반으로 기술 개발과 공유함

○ **(네덜란드)** 네덜란드 국립도서관은 디지털 전환 사업과 디지털 아카이빙을 적극적으로 수행하며, 유로피아나 프로젝트(Europeana)의 주도적인 역할을 수행함

- 장기적으로 모든 장서를 디지털로 전환하기 위한 작업을 진행하고, 유럽의 여러 국가와 연합하여 온라인 환경에서 도서관 자료를 공유하기 위한 유로피아나 프로젝트에 참여함
- 자료의 디지털화와 영구 보존, 디지털 정보 검색과 접근 방법을 개발하기 위해 Google, ProQuest 등 민간 기업과의 제휴 협력을 통한 기술 개발과 연구를 수행함
- 연구부서의 KB Lab[3]은 디지털 컬렉션의 접근과 사용을 촉진하기 위한 도구를 개발하고,

2) https://kb-labb.github.io/
3) https://lab.kb.nl/

관련 데이터와 소스 코드 등을 공개 라이선스로 게시함

○ **(노르웨이)** 노르웨이 국립도서관은 '국가기억장치'로서 자료의 보관과 유통을 위해 디지털 국립도서관을 구축하고, 중세시대부터 현재에 이르는 다양한 자료를 디지털화해 제공하고 있음

- 디지털 출판물의 메타데이터는 출판업자로부터 구축되어 고품질의 디지털 콘텐츠를 확보할 수 있고, 이를 바탕으로 컬렉션의 확장과 검색 성능을 향상함
- 문화유산, 역사적 자료의 디지털화와 소장자료의 전체 디지털화 프로젝트가 수행되었으며, 국립도서관의 디지털 인문학 연구소(Digital Humanities laboratory)[4]에서 데이터 마이닝, 디지털 코퍼스와 사전 구축, 음악 분석 AI 시스템 등에 관한 연구를 수행함

[표 4-1] 국외 주요도서관의 인공지능 서비스 사례

국가	설명
미국	- **Newspaper Navigator**: 신문 데이터베이스에서 시각 자료를 찾아주는 서비스 - **Citizen DJ**: 소장하고 있는 오디오 자료를 활용해서 음악을 만들 수 있는 서비스
일본	- **차세대 디지털 도서관**: 텍스트와 이미지로 디지털 자료를 검색하는 서비스 - **NDC Predictor**: 일본의 십진분류법(NDC9판)을 자동으로 생성하는 서비스 - **MIMA 검색**: 문서 사이의 의미적 관계에 기반한 도서 검색과 시각화 서비스 - **재팬 서치**: 디지털 아카이브를 연계한 통합 일본 디지털 자원 열람 서비스
프랑스	- **GallicaPix**: 다양한 주제의 시각 자료를 탐색할 수 있는 서비스
스웨덴	- **Swedish BERT model**: 도서관의 텍스트 자료 기반으로 스웨덴어 말뭉치 구축 - **Swedish spaCy**: 대규모 텍스트 분석을 위한 스웨덴어 자연어 처리 모델 구축 - **Advertisement Classification**: 디지털 아카이브 자료를 이용해 일반 기사문과 광고문을 분류하는 분류기 구축 - **스웨덴-노르웨이 연합 언어모델 학습**: 연합 기계 학습을 통해 스웨덴-노르웨이 연합 언어모델 학습 구축
핀란드	- **Annif**: 자동 주제 색인과 분류를 위한 오픈소스 도구 - **Finto**: 국가 온톨로지 인프라와 온톨로지 및 어휘 활용을 위한 소프트웨어 - **FintoAI**: Finto를 기반으로 하는 인공지능 서비스
네덜란드	- **Demosaurus**: 텍스트 인식하여 자동으로 메타데이터 추천하는 시스템 - **Frame generator**: 토픽 모델링 기반 ㄴ키워드 자동 생성기 - **DBNL ngramviewer**: ngram 알고리즘을 사용해 키워드 등장 빈도 시각화 도구 - **PoliMedia**: 시맨틱 웹 기술을 활용한 국회회의록과 언론 텍스트 데이터 분석
노르웨이	- **머신러닝을 이용한 듀이 십진분류법 생성**: 디지털 도서관 환경에서 DDC 기반으로 문서 자동분류를 연구하는 프로젝트

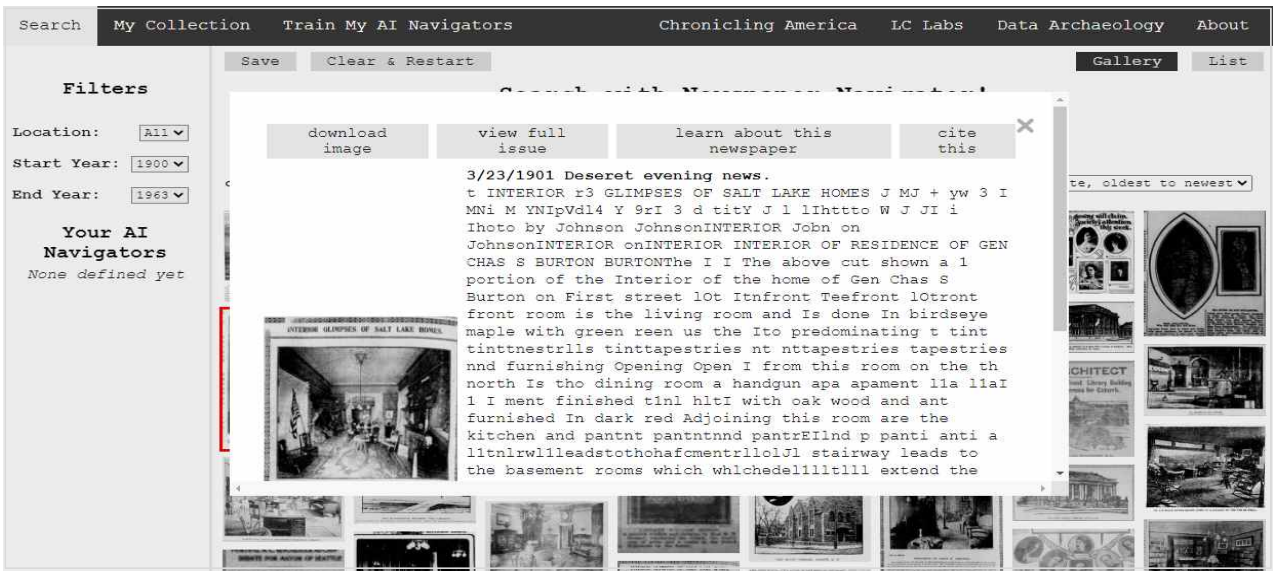
4) https://www.nb.no/en/forskning/digital-humanities-laboratory/

2. 미국

□ Newspaper Navigator[5)]

○ 미국의회도서관 LC labs가 2020년 9월부터 2021년 6월까지 진행한 머신러닝 기반의 신문 데이터베이스 시각 자료 재현 프로젝트의 결과물로, Chronicling America 신문 데이터베이스에서 시각 자료를 찾아주는 AI 서비스를 제공함

- 1789~1963년에 발행된 총 1,600만 부의 Chronicling America 신문 컬렉션 대상·신문 1600만 부에서 헤드라인, 사진, 일러스트, 만화, 지도, 광고 등 150만 장 이상을 반복 학습 훈련하고, OCR을 이용해 각 기사에서 헤드라인과 주석을 추출하여 데이터를 수집함
- 시각 자료 다운로드, 관련 전체 이슈를 열람할 수 있는 탐색적 검색 응용 프로그램을 구축함



<그림 4-1> Newspaper Navigator의 검색 결과

○ 검색하고 싶은 단어와 신문이 간행되는 미국 주, 시작 연도와 끝 연도를 지정하면 조건에 맞는 시각 자료를 반환함

- 시각 자료를 클릭하면 발행 신문사와 날짜, 텍스트 정보와 사진 다운로드, 해당 사진이 발행된 신문의 전체 페이지를 열람할 수 있는 링크, 해당 신문사의 정보를 얻을 수 있는 링크와 인용 문구 정보를 제공함

○ 나중에 다시 열람하고자 하는 시각 자료는 클릭하여 My collection에 저장하여 원할 때 확인이 가능함

- Train My AI Navigator 기능을 사용하여 선택한 사진들과 시각적으로 유사한 다른 사진들

5) https://news-navigator.labs.loc.gov/search

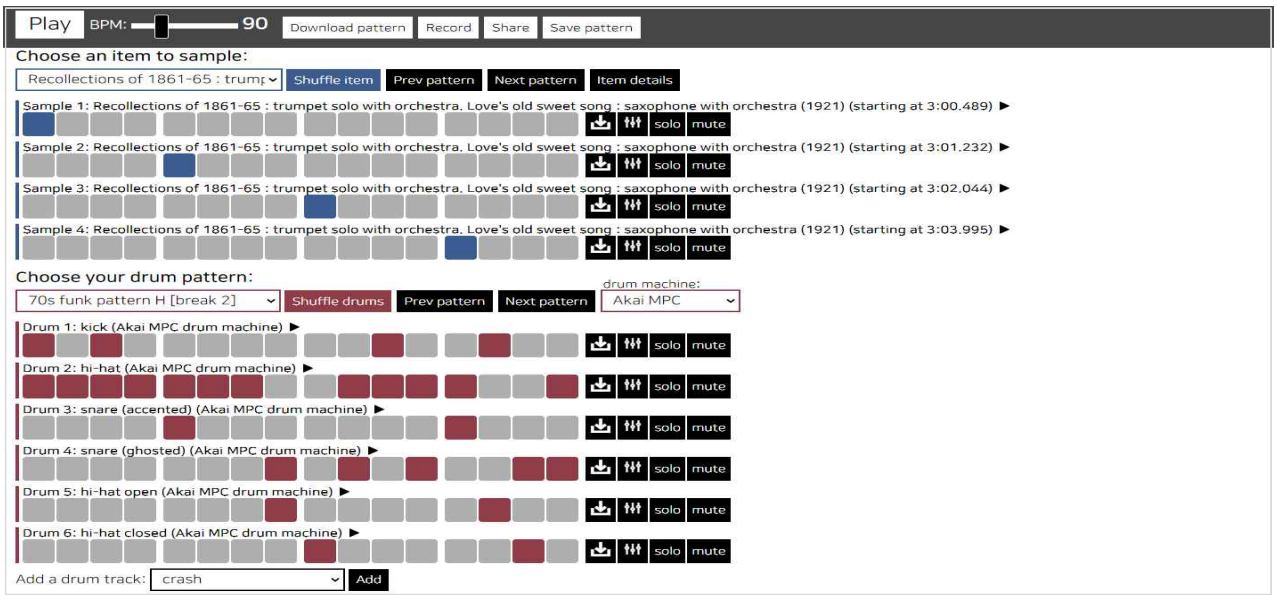
탐색할 수 있음

- 여러 개의 Navigator를 이용자가 원하는 이름으로 지정하여 저장할 수 있음
- 유사한 시각 자료에 대한 검색 및 추천 결과를 빠르게 생성하기 위해 미리 계산된 이미지 임베딩을 활용하고 있음

□ Citizen DJ[6)]

○ 이용자들이 도서관이 소장하고 있는 오디오와 동영상 컬렉션을 사용하여 힙합 음악을 만들 수 있는 기능을 제공함

- 미국 의회도서관에서 무료로 사용 가능한 사운드를 탐색하고, 웹 페이지에서 바로 비트와 결합하여 선호하는 소프트웨어로 파일 다운로드가 가능함
- 오디오 자료 처리를 위해 scikit-learn 파이썬 라이브러리 활용하여 수백 시간 분량의 사운드 자료에서 다양한 샘플을 자동으로 생성함



<그림 4-2> 오디오 결합 인터페이스

○ 미국 의회도서관이 소장하고 있는 3백만 개 이상의 오디오와 비디오 자료들을 사용하여 새로운 음악을 생성할 수 있음

- 문화적 가치가 있고 무료로 사용할 수 있는 오디오와 비디오 컬렉션을 일반 이용자들이 활용할 수 있는 방안을 제공하기 위해 소장 사운드 컬렉션을 지속적으로 공개하고, 소리를 직접 들어보고 다른 비트와 결합하고 파일을 내려받을 수 있는 인터페이스 서비스를 제공함

6) https://citizen-dj.labs.loc.gov/

3. 일본

□ 차세대 디지털 도서관(Next Digital Library)[7]

○ 일본 국립국회도서관 차세대 시스템 개발 연구실(NDL Labs)가 개발한 검색 서비스로, 국립국회도서관의 디지털 컬렉션에 공개된 약 336,000건의 자료를 검색할 수 있는 텍스트 검색(keyword search)과 이미지 검색(illustration search)을 제공하고 있음

- 전문 텍스트 검색이나 머신러닝을 이용한 자동 처리(예: 목차 자동생성), 국제 이미지 호환 프레임워크(International Image Interoperability Framework; IIIF) API 등의 기술적 효용성을 검증하는 것이 목적임
- 머신러닝이 활용한 학습용 데이터셋은 국립국회도서관의 GitHub에서 공개하고 있음

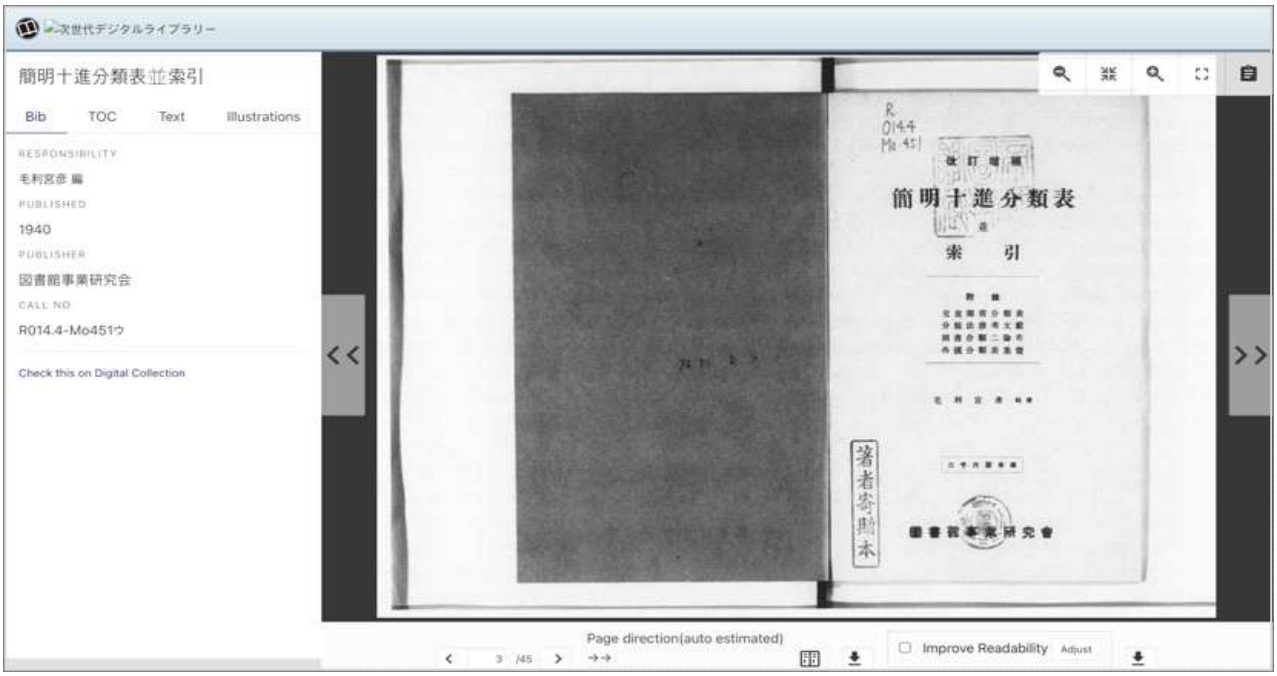
○ 키워드 기반으로 텍스트 전문을 검색하고, OCR로 생성된 텍스트에서 원하는 키워드 검색 시 키워드가 포함된 책과 구체적인 페이지를 탐색할 수 있음

- 텍스트 검색은 오픈소스 분산 검색엔진인 Elasticsearch를 사용하고 있음
- 디지털화된 도서의 본문과 함께 서지정보(Bib), 목차(TOC), 검색 키워드 위치(Text), 추출된 이미지(Illustration) 정보를 제공함

○ 도서에 포함된 이미지를 자동 추출하여 키워드 기반으로 이미지를 검색할 수 있고, 사용자가 업로드한 이미지와 유사한 이미지가 포함된 도서를 검색할 수 있음

- 이미지 검색을 위해 ANN(approximate nearest neighbor) 기반의 벡터 검색엔진 Vald를 사용해 이미지 인식과 추천, 탐지 등의 처리를 수행함
- 머신러닝을 통해 추출된 이미지는 시맨틱 태그(semantic tag)가 자동으로 추가되어 이미지 검색에 활용되고, 고해상도 이미지를 온라인에서 대규모로 제공하기 위한 공개 표준 IIIF를 활용한 이미지 뷰어를 제공함
- 노화로 인해 변색된 이미지를 탐지하고, 배경을 백색화하는 서비스도 제공하고 있음

7) https://lab.ndl.go.jp/index.html



<그림 4-3> 키워드 검색 결과로 선택된 이미지 화면

□ NDC Predictor[8)]

○ NDL Labs가 개발한 자동분류 서비스로, 머신러닝을 사용하여 도서의 서지정보로부터 일본의 십진분류법(NDC9판)을 자동으로 분류하는 서비스를 제공함

- 서지정보 입력 시 해당 도서의 NDC 분류 후보 3가지와 함께 각 후보에 따른 정확도를 제공함
- 머신러닝 모델은 Facebook AI Research가 개발한 fastText를 사용하고, 학습된 모델은 일본 국립국회도서관의 Github에서 결과를 공개하고 있음. 일본어 형태소 분석은 일본어를 위한 형태소 분석기 MeCab과 검색용 일본어 형태소 분석기 Kumoroji를 사용함
- 분류 예측을 위한 학습데이터는 일본 국립국회도서관이 소장하는 서지 데이터(2017년 7월 기준) 중 NDC가 존재하는 자료 4,256,238건의 제목, 출판사 등의 서지 정보를 사용하고 있음

○ NDC의 분류 데이터는 일본 도서관 협회가 링크드 데이터 형태로 제공하는 NDC 9판을 사용함

- NDC 데이터는 NDC의 계층 구조를 표현하는 RDF 형식으로 구축되었고, 분류 항목은 SKOS를 광범위하게 재사용하여 분류체계의 계층 관계를 표현하였음
- 링크드 데이터로 구축된 NDC 8판과 9판의 데이터는 Turtle 형식으로 한 번에 데이터를 다운로드할 수 있는 서비스를 제공하고 있음

8) https://lab.ndl.go.jp/ndc/



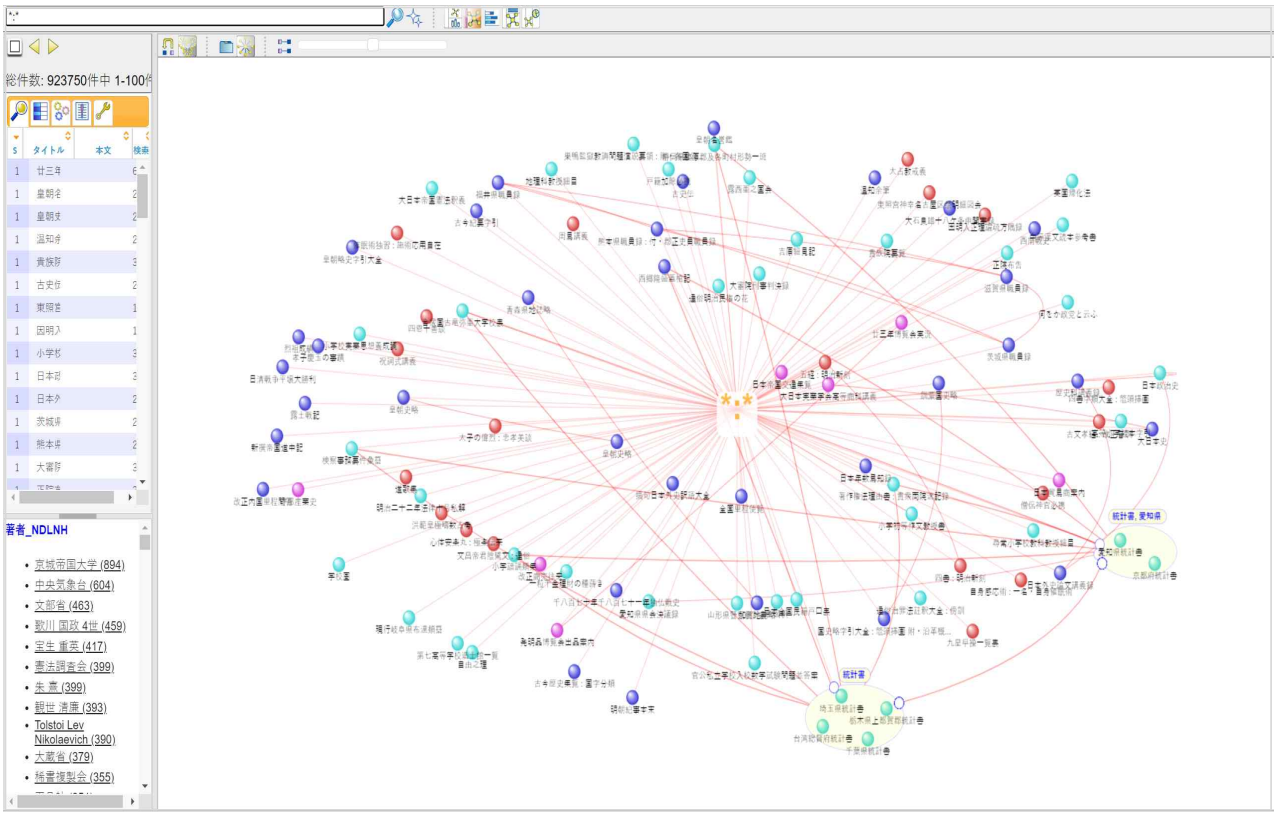
<그림 4-4> 분류 결과와 정확도

□ MIMA 검색[9)]

○ 자연어 처리 기술을 사용하여 문서 사이의 의미적인 관계에 기반한 도서 검색과 시각화 서비스를 제공하고 있음

- 해당 서비스는 일본 국립국회도서관이 디지털화한 도서의 서지 정보 약 92만 건과 목차 정보 약 71만 건을 대상으로 검색과 시각화 서비스를 제공함
- 검색된 문서는 점으로 표현되고, 문서와 관계있는 문서는 선으로 연결된 점으로 표현됨
- 문서 사이의 연관성이 높을수록 문서를 표현하는 점은 가깝게 시각화되고, 내용적으로 가까운 문서의 집합이나 연관 구조를 한눈에 파악할 수 있음

9) https://lab.ndl.go.jp/ut/



<그림 4-5> 도서 검색 시각화 결과

□ 재팬 서치(Japan Search)[10]

○ NDL Labs가 개발한 다양한 분야(도서/간행물, 문화재, 미디어 예술 등)의 디지털 아카이브와 연계해 일본이 보유한 다양한 디지털 자원의 검색과 열람, 활용 서비스를 제공하는 국가 플랫폼임

- 다양한 기관에서 구축한 디지털 아카이브를 연계하기 위해 링크드 오픈 데이터(linked open data)를 구축하였고, 이를 Web API와 SPARQL API로 제공하고 있음
- 83개 기관의 디지털 아카이브를 연계하기 위해 디지털 자원의 메타데이터 기술을 위한 RDF 기반의 JSP-RDF schema 설계하고, 약 10억 8천만 건의 트리플데이터(triple data) 제공

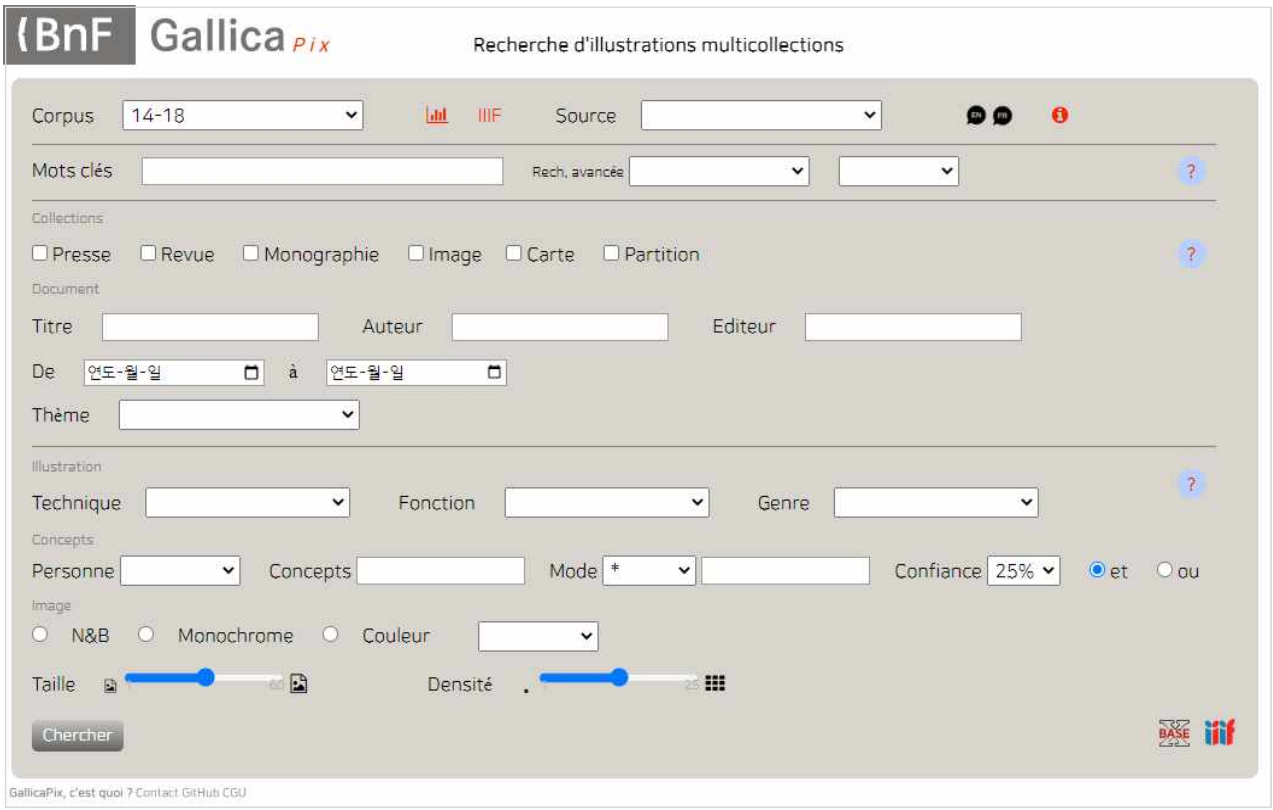
4. 프랑스

□ GallicaPix[11]

10) https://jpsearch.go.jp/
11) https://gallicapix.bnf.fr/rest?run=findIllustrations-form.xq

○ 프랑스 국립도서관이 개발한 검색 서비스로, 딥러닝을 사용하여 프랑스 국립도서관이 소장하고 있는 시각 자료를 탐색할 수 있는 검색 서비스를 제공하고 있음

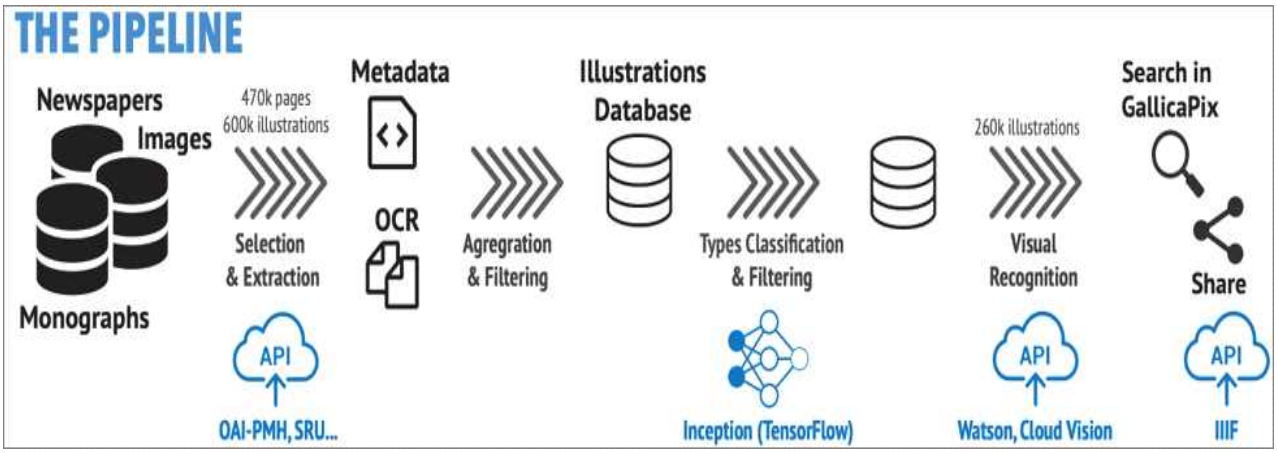
- 검색하고 싶은 주제와 단어와 그 외 자료의 형태를 선택할 수 있고, 자료의 유형(문서/삽화/개념/이미지 색상)에 따라 조건을 지정할 수 있는 선택지를 제공하고 있음
- 1차 세계 대전, 광고(1910-1920), 영국 국립 기록 보관소와 공동 작업한 결과물, 패션 잡지 보그 프랑스판 삽화(1920-1940), 동물(물고기, 포유류, 곤충, 식물, 조개, 파충류 등) 등을 검색을 제공함



<그림 4-6> GallicaPix 검색 화면

○ OCR과 딥러닝을 사용하여 시각 자료의 메타데이터를 강화하고 있음

- OCR과 주석을 사용해 문서 구조를 광학적으로 인식하여 기존 메타데이터 요소 외 시각 자료에 대해 추가적인 데이터 수집이 가능함
- 텐서플로우(딥러닝 프레임워크)를 사용하여 삽화의 유형, 특성과 기능 구별하여 서비스를 구축하였음
- 시각 자료의 인식을 위해 인공지능 Watson의 Visual Recognition의 API와 머신러닝 모델을 사용하는 Google Cloud Vision의 API를 활용하였음



<그림 4-7> GallicaPix 구축 과정

○ 검색 결과, 조건에 맞는 시각 자료를 반환하고, 추가로 시각 자료의 종류와 시각 자료에 포함된 개념, 태그 정보를 필터링할 수 있는 기능을 제공함

- 시각 자료가 발행된 날짜와 주석 정보 확인 가능하며, JSON 형식의 파일을 다운로드할 수 있음
- 사진을 소장하고 있는 프랑스 국립도서관 홈페이지 바로가기 링크와 SNS로 바로 공유 가능한 링크를 제공함



<그림 4-8> GallicaPix 검색 결과

5. 스웨덴

□ Swedish Bert (KB-BERT)

○ 스웨덴 국립도서관 연구소(KB Lab)는 스웨덴어 최적화 BERT 모델을 개발함

- 구글이 개발한 사전 훈련 언어모델인 BERT을 사용했고, 스웨덴 도서관 아카이브의 모든 디지털 텍스트에서 개체를 식별하고, 자동으로 주석을 입력하는 데 활용되고 있음

○ 도서관의 풍부한 텍스트 자료를 기반으로 학습용 스웨덴어 말뭉치를 구축함

- 1940년~2019년 수집 자료 중 디지털화한 신문, 스웨덴 정부의 공식 보고서, 출판물, 소셜 미디어와 스웨덴 위키피디아에서 단어를 추출하고, 추출한 단어를 스웨덴어의 문법적, 매체 특성을 고려하여 보정하였음

□ Swedish spaCy

○ 스웨덴 국립도서관 연구소는 대규모 텍스트 분석을 위한 스웨덴어 자연어 처리 모델 Swedish spaCy를 개발함

- 파이썬 기반 오픈소스 라이브러리인 spaCy는 품사 태깅(POS), 개체명인식(NER), 의존 구분 분석(Dependency parser)을 위한 언어모델을 제공하고, 현재 spaCy v3.3을 통해 스웨덴어 spaCy 모델을 사용할 수 있음

○ 개별 모델의 특성에 맞는 스웨덴어 말뭉치를 수집 또는 구축하여 학습에 사용함

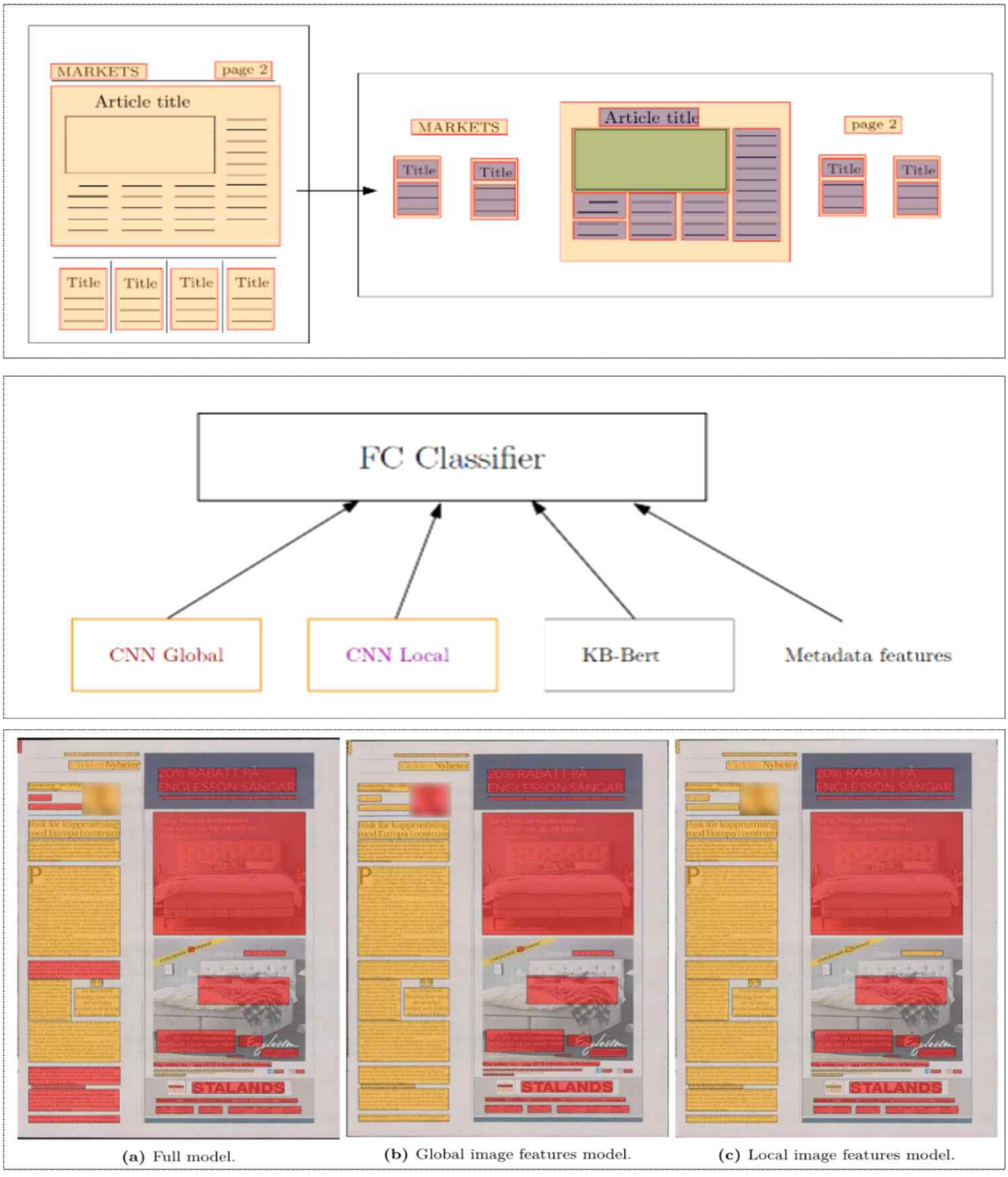
- 학습용 스웨덴어 말뭉치는 웁살라 대학의 Swedish-Talbanken Treebank, 영문 말뭉치인 English_LineES의 스웨덴어 번역인 Swedish_LineES, 스톡홀름 대학과 우메오 대학의 협력으로 구축된 SUC 3.0을 수집함
- SUC 3.0 말뭉치의 주석을 생성하기 위해 스웨덴어 텍스트에 대한 자동 주석 처리 도구인 Sparv을 사용함

□ Advertisement Classification

○ 디지털 아카이브의 디지털화된 신문의 이미지, 텍스트, 메타데이터를 활용해 일반 기사문과 광고문을 분류하는 분류기를 개발함

- OCR 과정에서 신문 페이지의 콘텐츠를 분할하고, 위치와 높이 등 분할된 콘텐츠에 대한 메타데이터를 생성하고, 분할된 콘텐츠(텍스트 또는 이미지)에 수동으로 주석(editorial, ad, mixed)에 추가할 수 있음

- 이미지 데이터의 처리에 사용되는 신경망 모델인 CNN과 스웨덴어에 최적화한 BERT 모델인 KB-BERT을 연결하여 FC 분류기를 구축함
- CNN 모델은 전체 신문 페이지의 이미지를 학습하는 CNN Global과 분할된 이미지를 학습하는 CNN Local를 구성함
- KB-BERT는 분할된 텍스트에 대한 특징(feature)을 추출하고 토큰 임베딩을 수행함



<그림 4-9> 광고 분류기 작동 과정

□ 스웨덴-노르웨이 연합 언어모델 학습

○ 스웨덴 국립도서관, AI Sweden, Scaleout System이 2020년 1분기 동안 개발한 스웨덴-노르웨이 연합 언어모델로, 데이터를 직접 공유하거나 공개하지 않고 모델을 교육할 수 있는 연합 기계 학습(Federate Learning)을 통한 스웨덴-노르웨이 연합 언어모델을 개발함

- 학습용 데이터는 다국어 말뭉치 OSCAR와 위키피디아를 통해 스웨덴어, 노르웨이어 데이터를 수집하고, 연합 머신러닝 프레임워크인 FEDn을 기반으로 구글의 마스킹 언어모델 ELELCTRA 연합 학습을 수행하였음
- 대규모 인공신경망 모델을 연합 학습하는 최초 사례로, 스칸디나비아 언어모델 구축에 활용할 수 있고, 프로젝트의 확장에 따라 스웨덴, 노르웨이, 덴마크, 핀란드 등의 국립도서관 컬렉션을 활용할 계획임

6. 핀란드

□ Annif[12)]

○ 핀란드 국립도서관이 개발한 자동 주제 색인과 분류를 위한 오픈소스 도구를 제공함

- 언어 독립적이며 주어진 언어의 학습이 가능한 상황일 경우, 모든 색인 어휘와 함께 사용할 수 있음
- 소스코드는 공유 플랫폼이자 협업 도구인 깃허브에 공개하고, REST API를 제공하고, 텍스트에서 자동 색인하고 분류한 주제 키워드를 추천하는 기능을 제공함

○ TensorFlow, Omikuji, fastText, Gensim과 같이 기존 머신러닝과 자연어 처리 기술을 조합한 알고리즘을 사용함

- 페이스북에서 개발한 자연어 처리 알고리즘인 FastText와 자연어를 벡터 처리하여 처리하는 Gensim을 사용함

□ Finto[13)]

○ 핀란드 국립도서관이 개발한 다양한 정보 분야를 연결하는 국가 온톨로지 인프라와 온톨로지 및 어휘 활용을 위한 소프트웨어

- Finto의 동의어 사전과 온톨로지 어휘는 도서관, 박물관 데이터베이스에 사용됨

12) https://www.kansalliskirjasto.fi/en/services/annif
13) https://www.kansalliskirjasto.fi/en/services/finto

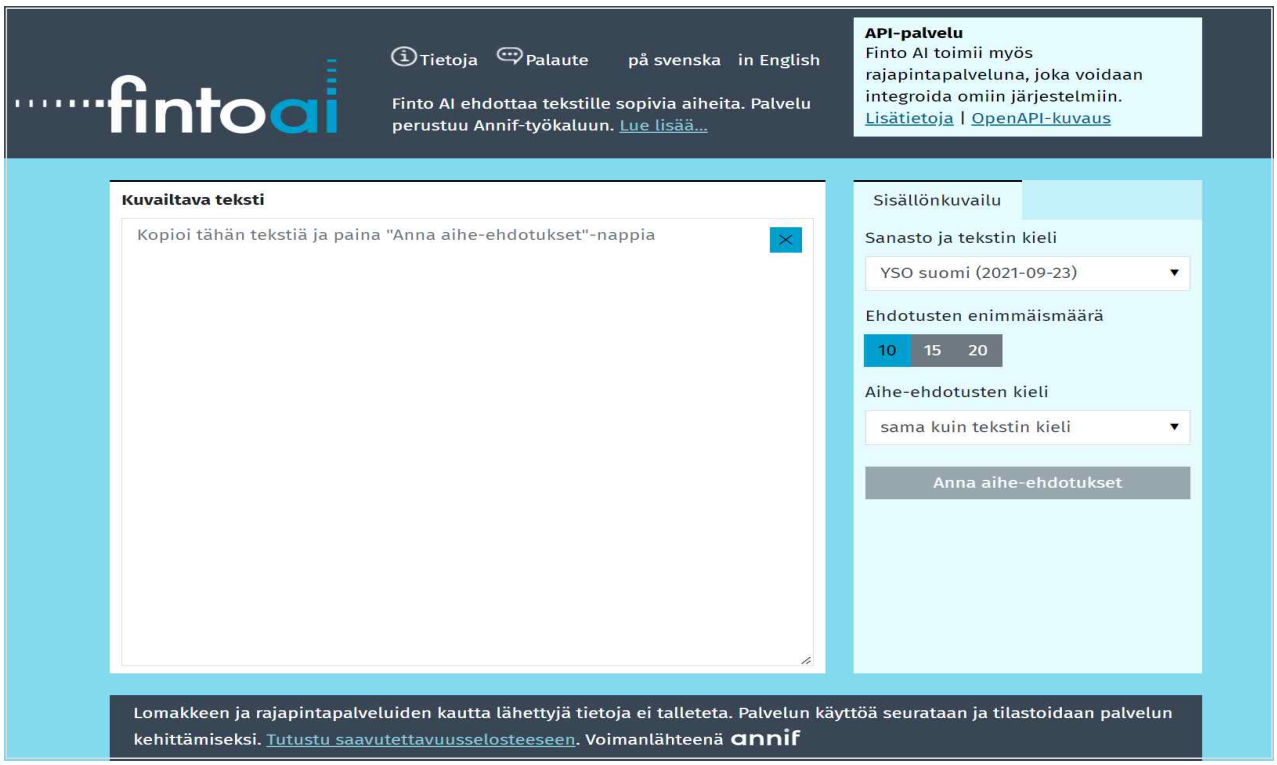
○ 국립도서관의 온톨로지 개발과 네트워크 지원 서비스

- 기계가 이해할 수 있는 형식으로 일부 주제 영역에서 개념과 관계를 설명하고 용어 대신 식별자(URI)로 개념의 계층 구조를 표현하고 있음
- 핀란드 문화권 내의 일반적인 개념으로 구성된 온톨로지와 주제와 지식체계를 표현하는 온톨로지 어휘인 SKOS를 활용하여 문서를 설명하는 색인과 키워드를 추출함

□ Finto AI[14)]

○ 핀란드 국립도서관에서 개발한 자동 주제 인덱싱 서비스

- 핀란드어 온톨로지(YSO)를 기반으로 제안한 웹 서비스로, Annif(자동 주제 색인 도구)와 텐서플로 기반 딥러닝 알고리즘 기반으로 구축됨
- 웹 인터페이스를 통해 입력한 텍스트에 대해 주제를 제안받을 수 있고, 주제의 최대 수량을 정할 수 있으며, REST API로 데이터를 개방함



<그림 4-10> Finto AI 서비스 화면

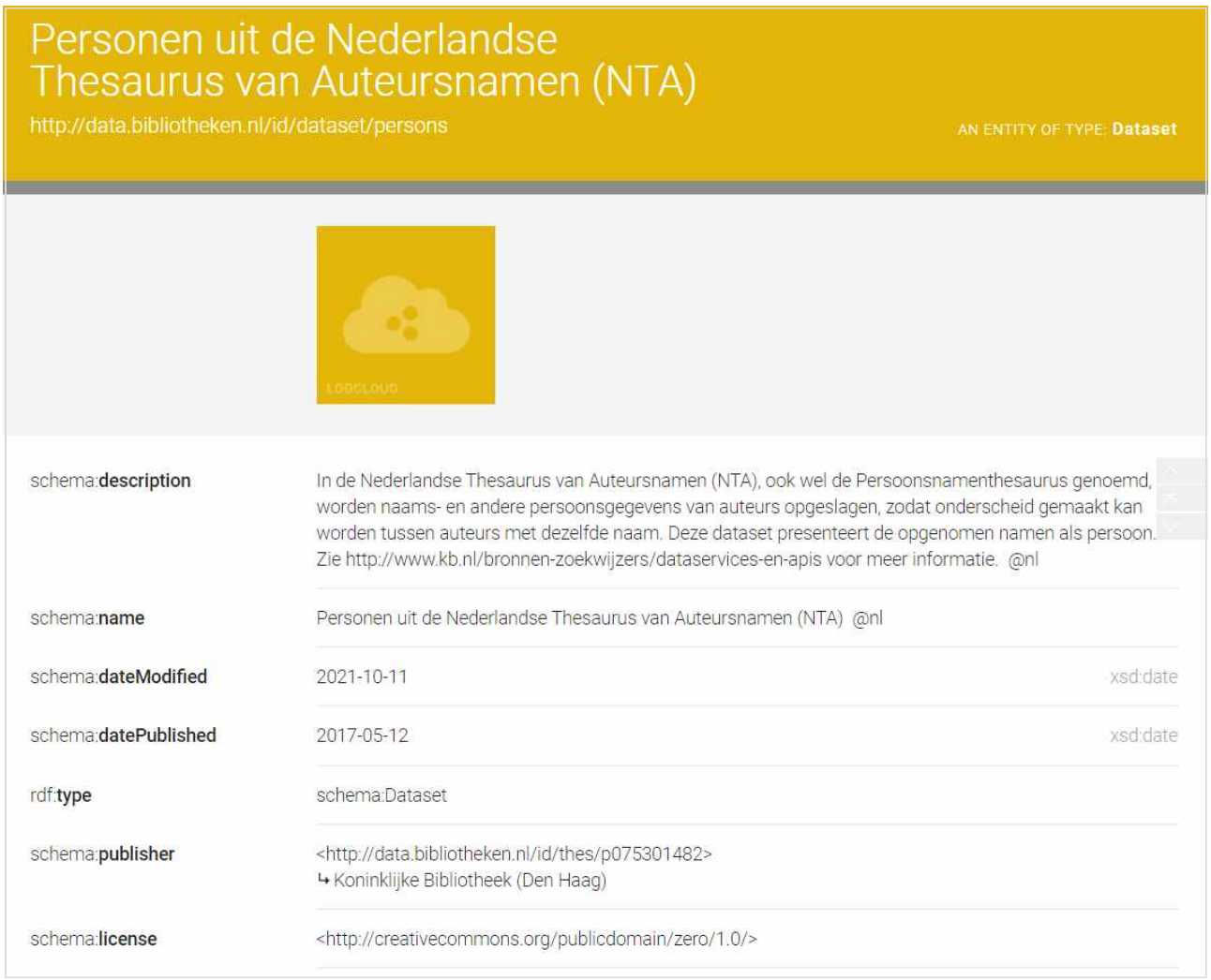
14) https://ai.finto.fi/

7. 네덜란드

□ Demosaurus[15)]

○ 네덜란드 국립도서관 연구소에서 개발한 인공지능 기반 자동 메타데이터 할당 도구

- 각 작품의 저자, 출판 시기, 출판사, 형식 등의 메타데이터를 자동으로 추가할 수 있는 지원 도구를 제공함
- 네덜란드 저자 이름 시소러스(NTA), 자동 주제 색인 도구(Annif) 기반으로 구축됨
- 제목 입력 시, 장르나 주제에 대한 적절한 키워드 목록을 제공하고, 네덜란드 NTA(저자 이름 시소러스)와 연결하여 저자의 이름, 직업, 생년월일을 제공함
- 머신러닝을 활용해 해당 작품의 저자일 가능성이 큰 인물을 추천하고, 도구 사용자가 최종 선택할 수 있음



<그림 4-11> 네덜란드 NTA(저자 이름 시소러스)

15) https://github.com/KBNLresearch/Demosaurus

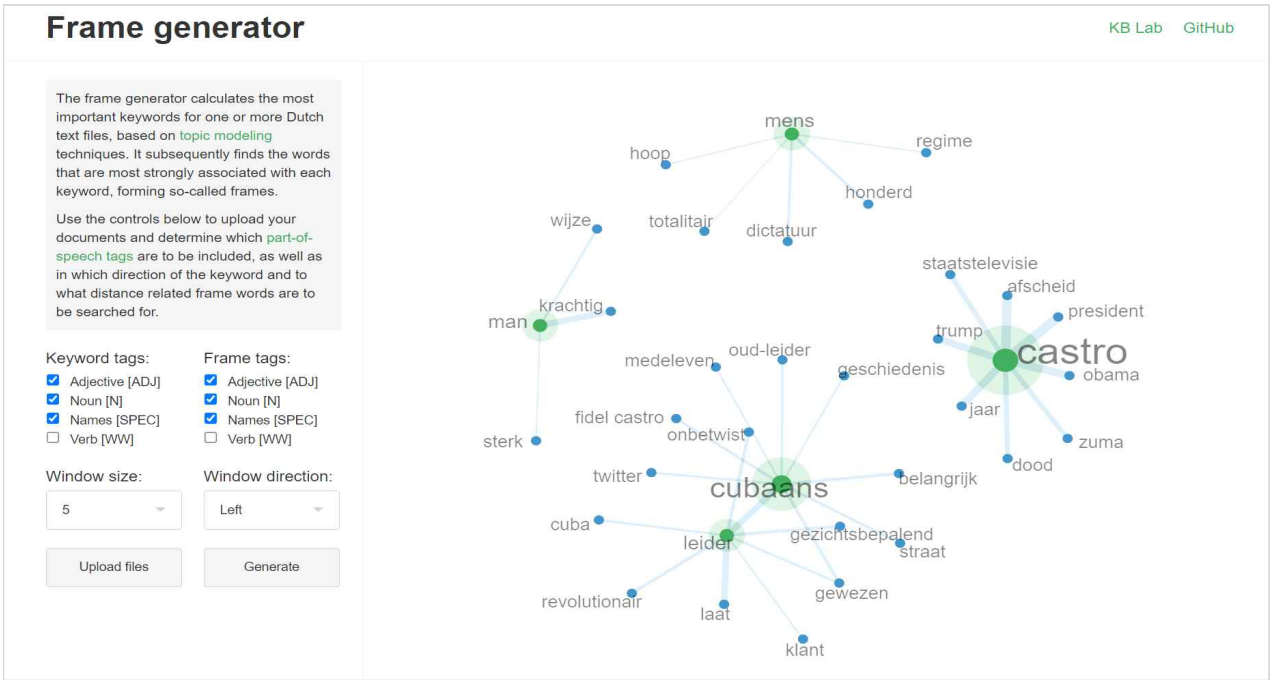
□ Frame generator[16)]

○ 네덜란드 국립도서관 연구소에서 개발한 토픽 모델링 기반 중요 키워드 추출과 시각화 도구

- 네덜란드 텍스트 파일을 업로드하면 단어나 문서의 집합에서 주제(topic)를 찾고, 연결된 키워드를 시각화함
- 파이썬 라이브러리나 웹 어플리케이션 형태로 제공하고 있음

○ Mallet 또는 Gensim 토픽 모델링 알고리즘으로 연관된 주제(topic) 생성

- Java 기반 자연어 처리, 문서 분류, 클러스터링 등을 지원하는 Mallet과 Gensim 알고리즘을 사용함
- TF-IDF 점수를 기반으로 순위가 매겨진 단일 키워드 목록을 추출하고, 키워드에 대한 동시 발생 패턴을 탐색함
- 문서의 집합에서 어떤 단어가 특정 문서 내에서 얼마나 중요한 것인지를 나타낸 값을 추출하여 계산함
- 선택적으로 NLP 제품군인 Frog를 사용하여 입력 텍스트에 표제어와 키워드를 지정하고 특정 품사 태그에 할당



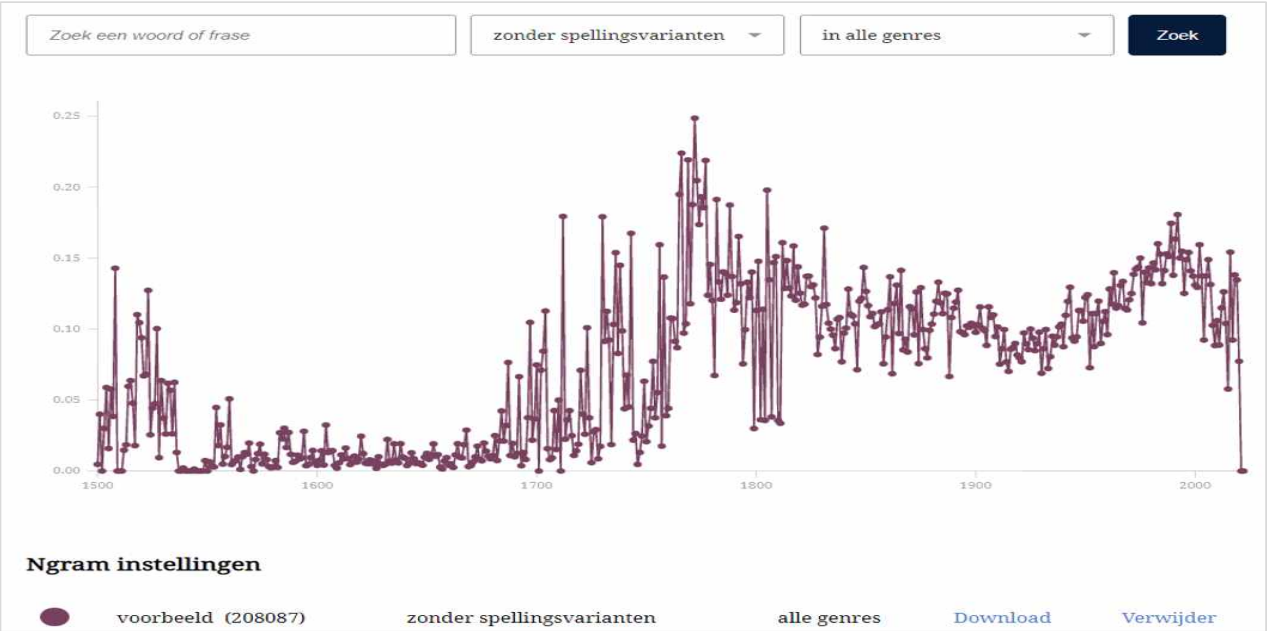
<그림 4-12> Frame generator 서비스 화면

16) https://kbresearch.nl/frames/

□ DBLN ngram viewer[17)]

○ 네덜란드 국립도서관 연구소와 네덜란드 문학 디지털 도서관(DBNL)이 개발한 DBNL 디지털화 텍스트 검색 서비스로, 기간 내 특정 키워드가 언급된 빈도를 시각화함

- 10개 이내의 키워드를 검색할 수 있고, 연도별로 단어가 언급된 횟수를 표현하고, 그래프를 구성하는 꼭짓점을 클릭하면, 해당 연도에 언급된 텍스트를 탐색할 수 있음
- 철자가 변형된 단어를 포함하여 검색할 수 있고, 특정 단어가 출현 빈도와 사용된 맥락에 대한 역사적인 과정을 탐색할 수 있음



<그림 4-13> DBNL ngram viewer 검색 결과화면

□ PoliMedia

○ 네덜란드 언론이 정치적 논쟁을 다루는 방식을 분석하기 위한 네덜란드 국립 도서관 연구소의 프로젝트

- 네덜란드 의회 토론 회의록을 역사 신문, ANP 라디오 게시판의 데이터베이스에 연결하여 보도 범위에 대한 교차 미디어 분석을 진행함
- 시맨틱 웹 기술을 활용하여 토론의 연사, 날짜, 내용에서 관련 정보를 추출하고, 신문, 라디오, 텔레비전 아카이브 데이터베이스에 검색할 수 있는 질의를 생성함
- 질의에 해당하는 미디어 항목과 토론 사이의 연결 링크를 만들어 의회 회의록과 미디어 자료를 의미적으로 탐색할 수 있는 서비스를 제공하고 있음

17) https://dbnl.org/ngram-viewer/

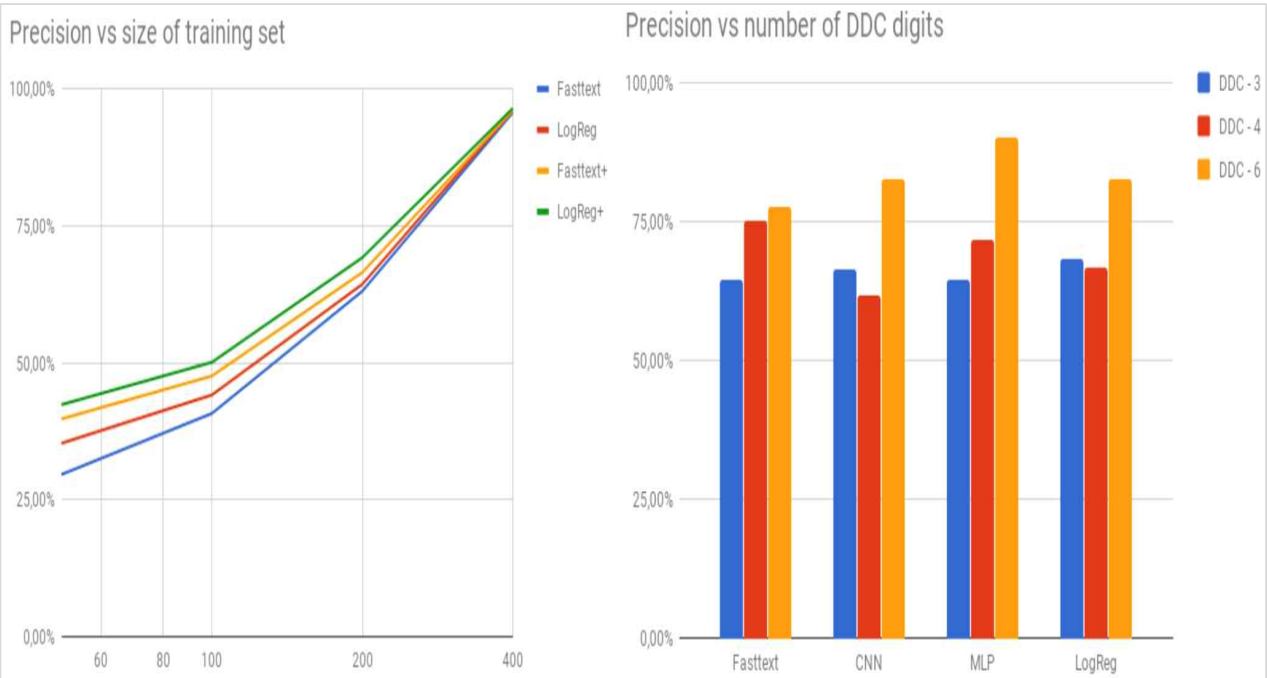
8. 노르웨이

□ 머신러닝을 활용한 듀이 십진분류법 생성

◦ 노르웨이 국립도서관이 개발한 문서 자동분류 프로젝트로, 2018년 8월 26일 IFLA WLIC 2018 컨퍼런스에서 발표되었음*

* Machine learning for production of Dewey Decimal (2017)

- 디지털 도서관 환경에서 DDC를 기반으로 문서 자동분류를 연구하는 프로젝트로, 노르웨이 국립도서관이 보유하고 있는 기존 디지털 문서와 부족한 문서를 보완하기 위해 기존 문서에서 단어를 교체하여 인위적인 문서를 생성하여 실험에 활용함
- 노르웨이 국립도서관이 보유하고 있는 600,000개 이상의 메타데이터 중 36,000개 이상이 디지털 형식으로 존재함. 실험에 사용된 자료는 노르웨이와 북유럽 과학 저널에 발표된 문헌이며, 일반적으로 디지털 형태로 PDF나 HTML을 통해 이용 가능함
- 다층 퍼셉트론(MLP), 합성곱신경망(CNN), FastText, 로지스틱 회귀분석(LR) 등이 프로젝트에 사용되었음
- 연구 결과, 머신러닝이 DDC의 반자동 또는 완전 자동화 생산에 적합한 환경이며, 분류의 정밀도는 학습데이터로 활용된 문서의 수와 생성되는 DDC의 자릿수(3/4/6자리)의 증가에 따라 분류 성능이 향상되었음



<그림 4-14> 문서 수(왼쪽)와 DDC 자릿수(오른쪽)에 따른 정확도 차이

제 5 장 국회전자도서관 서비스 현황 분석

1. 개요

□ 조사 대상

○ 국회도서관에서 관리하는 8개 웹사이트를 대상으로 서비스 현황을 조사함

- **(국회전자도서관)** 국회도서관 소장자료, 테마컬렉션, 지능형 질의응답 서비스를 제공함
- **(국회법률도서관)** 국내외의 법률정보를 종합적으로 제공하기 위해 법률정보회답 제공, 법률 관련 데이터베이스 구축과 검색 서비스를 제공함
- **(국회의원 정책자료)** 국회의원이 정책과 입법의 개발을 위해 주최하는 세미나 정보와 정책 자료집, 의정 보고서 등을 제공함
- **(국회·지방의회 의정정보시스템)** 전국 지방의회의 의정 자료를 국회도서관의 입법정보와 통합하여 제공함
- **(국회기록보존소)** 국회의 업무와 관련하여 생산된 모든 형태의 국회기록물을 관리하는 영구 기록물 관리기관으로, 이용자가 국회기록물에 접근할 수 있도록 다양한 콘텐츠를 제공함
- **(전문가정보 데이터베이스)** 국회의원과 입법 활동에 필요한 전문가의 인적자원 정보를 구축하여 맞춤형 전문가 정보를 제공함
- **(지능형 의회정보 융합 분석 Argos)** 뉴스와 소셜미디어 데이터에서 입법과제와 정책 현안을 수집하고 빅데이터 분석서비스를 제공함
- **(국회회의록 빅데이터 조사)** 국회회의록을 발언자, 키워드 등으로 검색할 수 있고, 발언 내용을 데이터셋으로 제공하여 다양한 형태로 활용할 수 있음

※ 웹사이트별 서비스 현황은 [표 5-1] 참고

○ 웹사이트 대부분은 검색 서비스를 기본으로 제공하며, 검색어에 대한 분석 결과를 시각화하거나 해당 사이트의 독자적인 서비스(예: 기계번역, 데이터 분석)를 제공함

- 국회전자도서관은 국회도서관 소장자료에 대한 통합 검색과 지능형 질의응답 서비스를 제공하고 있음
- 국회법률도서관은 법률정보 통합 검색을 기본으로 제공하고, 네이버와 공동 개발한 AI 일본법 번역기를 제공함
- 국회의원 정책자료, 국회·지방의회 의정정보시스템, 국회기록보존소, 전문가정보 데이터베이스, 국회회의록 빅데이터 5개의 웹사이트는 자료에 대한 검색 서비스만 제공하고, 전문가정보 데이터베이스, 국회회의록 빅데이터는 검색어에 대한 간단한 분석 결과만 구현함
- 지능형 분석 Argos는 국회 내·외부 데이터를 바탕으로 한 데이터 분석 서비스인 키워드 기반의 이슈 분석, 의정활동 분석, 국회의원별 광장, 빅데이터 분석, Argos 스토리보드 총 5종류의 데이터 분석 결과를 제공함

[표 5-1] 국회전자도서관 서비스 현황

사이트명	서비스	설명
국회전자도서관	국회도서관 소장자료 검색	국회도서관 소장자료와 DB에 대한 통합 검색 기능
	지능형 질의응답	국회도서관 보유 원문정보와 위키백과 자료를 바탕으로 기계독해를 적용한 질의응답 시스템
국회법률도서관	법률정보 통합 검색	국내외 법률 현황과 입법사례, 국내 입법 배경과 쟁점 등을 수집, 분석서비스
	AI 일본법 번역기	국회도서관과 네이버가 공동으로 개발한 일본법 자동번역 서비스
국회의원 정책자료	국회의원 정책자료 검색	키워드 검색 기반의 국회의원이 생산하는 정책자료 제공 서비스
국회·지방의회 의정정보시스템	국회·지방의회 의정정보 통합 검색	전국 지방의회 의정자료와 국회 입법정보 통합 제공 서비스
국회기록보존소	국회기록보존소 통합 검색	국회기록보존소 보유 기록물 통합 검색 서비스
전문가정보 데이터베이스	국회의원/전문가 인적정보 검색	국회 입법 활동 지원과 방향성 제시를 위한 다양한 분야의 전문가 인적정보 데이터베이스 검색
지능형 의회정보 융합 분석 (Argos)	키워드 기반 이슈 분석	이슈 키워드에 대한 뉴스, 소셜미디어 빅데이터 분석 결과를 통해 입법 활동 지원
	의정활동 분석	국회 관련 뉴스, 소셜미디어의 언급량을 기준으로 의정활동 분석 결과 제공
	국회의원별 광장	Argos 수집 데이터와 국회 자료를 종합하여 국회의원 관련 정보 서비스
	빅데이터 분석	최신 뉴스 데이터를 활용한 텍스트 마이닝 지원 서비스
	Argos 스토리보드	대수별 국회회의록 데이터의 통계정보, 키워드 정보 등 분석 결과 시각화
국회회의록 빅데이터	국회도서관 회의록 데이터 검색	국회회의록 빅데이터에 대한 통합 검색과 분석 결과 제공

2. 국회 전자도서관

□ 국회도서관 소장자료 검색[18)]

○ 국회도서관의 데이터베이스를 이용하여 약 1,200만 건의 도서, 학위논문, 학술기사, 연속간행물, 인터넷 자료, 외국법률 DB, 비도서 자료들의 검색이 가능함

- 데이터베이스로 국회도서관이 보유한 목록 및 원문정보를 제공함
- 외부기관 소장자료는 외부 지식 정보 전문기관을 통해 검색이 가능한 자료로서 검색 범위를 확대하여 이용할 수 있음
- 소장자료 검색방식은 기본검색, 전체(초록·목차·본문포함), 상세검색, 연산자 검색으로 분류됨

○ **(자동완성)** 검색어 입력 시 동일하게 시작하는 단어를 나열하는 기능

- 맞춤법 오류를 정정하는 기능은 제공하지 않아 정확하게 일치하는 단어만 검색할 수 있음



<그림 5-1> 국회도서관 소장자료 자동완성 검색 예시

○ **(인포그래픽)** 검색어의 연관어를 기준으로 분석하여 다양한 인포그래픽 형태로 제공하는 기능

- **(연관어 트랜드)** 검색어와 관련된 연관어의 출현빈도를 기준으로 연도별 시각화를 제공함
- **(연관어 분포)** 검색어가 중앙에 위치한 워드 클라우드 형태로 연관어는 연관성이 높을수록 글자가 크게 표현됨
- **(연관어 인물)** 검색어와 연관어와 함께 출현하는 인물을 그래프 형태로 제공함
- **(연관어 인기도)** 검색어와 연관어의 검색 빈도를 '월간', '주간'으로 나눠서 분석한 시각화 결과를 제공함

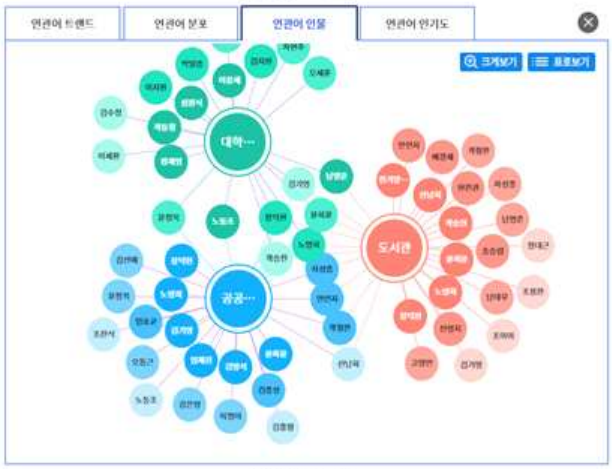
18) https://dl.nanet.go.kr/



<연관어 트랜드>



<연관어 분포>



<연관어 인물>



<연관어 인기도>

<그림 5-2> 검색 결과 인포그래픽 예시

○ **(음성지원)** 국회전자도서관에서 서비스하는 원문 자료 중 음성지원이 가능한 자료를 음성으로 읽어주는 기능

- 음성지원서비스는 일반인도 이용할 수 있으나, 시각장애인용 실시간 음성지원서비스는 회원가입 시 '시각장애인용 전용서비스'로 신청하여 승인 후 이용이 가능함
- 시각장애인용 실시간 음성지원서비스는 광학문자판독기술(OCR)을 도입하여 이미지로 된 원문자료도 실시간으로 변환하여 음성지원 서비스를 제공함

□ 지능형 질의응답[19)]

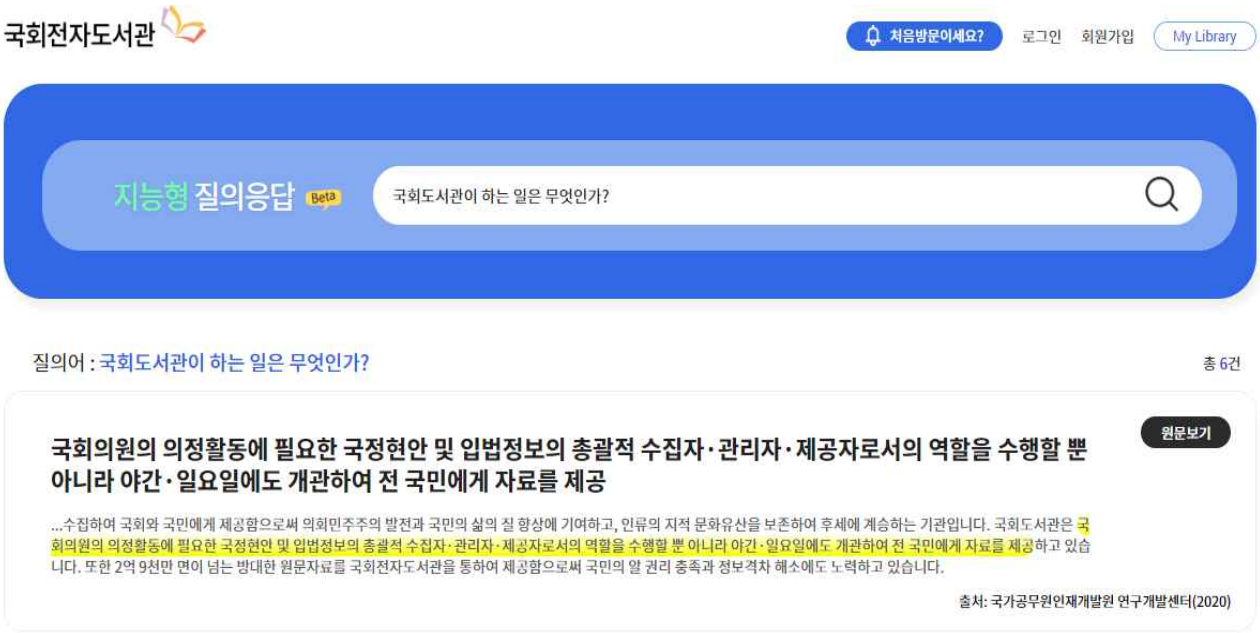
○ 국회도서관이 보유한 원문정보와 위키백과 자료를 대상으로 기계독해 기반의 질의응답 시스템을 제공하는 서비스

- 문서만 찾아주는 키워드 매칭 기반의 검색을 넘어서서 사용자의 질문과 문서를 이해하고 질

19) https://dl.nanet.go.kr/qas/qasMain.do

문에 적합한 답을 찾아 결과로 제공함

- 기계독해(Machine Reading Comprehension, MRC)는 자연어 질문과 문서를 딥러닝 모델이 이해하고 분석하여 사용자 질문에 적합한 대표 답변을 자동으로 찾는 인공지능 기술임



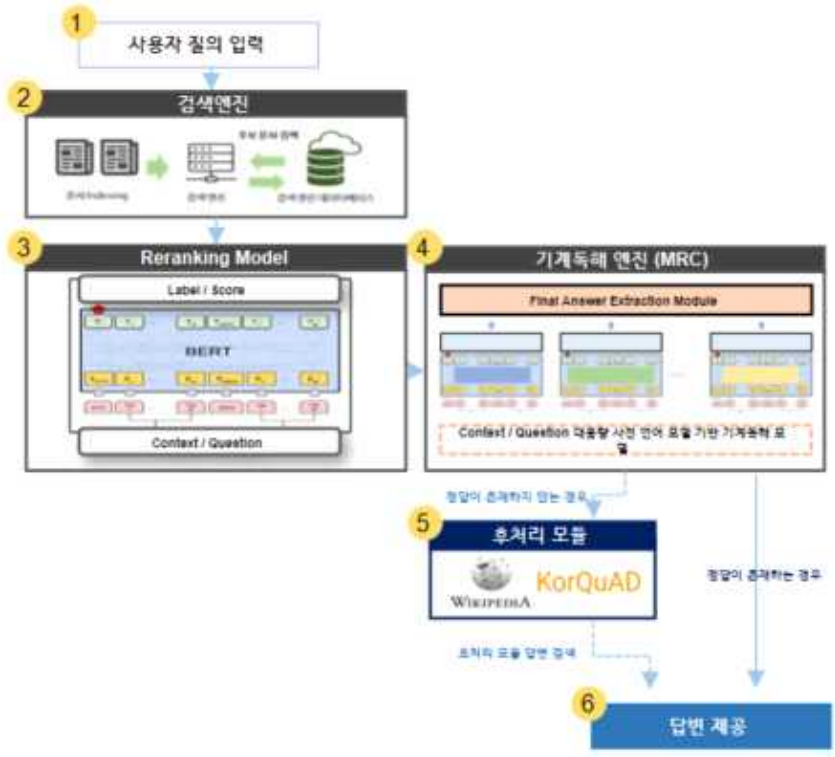
<그림 5-3> 지능형 질의응답 예시

○ 학습데이터는 국회 발간자료, 공공기관에서 생산된 정책정보, 위키백과를 기반으로 약 60,000건의 데이터가 구축됨

- 딥러닝 언어모델의 전이학습을 통한 기계독해 모델을 구축하여 단답형, 서술형 답변이 모두 가능함

○ **(서비스 구조)** 사용자가 질문을 입력하면 분석 후 딥러닝 언어모델 전이학습을 통해 답변을 순위화하여 정밀도/재현율 기반으로 대표 답변을 제공하는 기능

- 사용자가 질의를 입력하면 검색엔진을 통해 질의에 적합한 맥락을 추출함
- 추출된 맥락은 BERT를 활용한 Reranking Model을 통해 순위화되어 기계독해 엔진으로 처리됨
- 정밀도와 재현율을 계산하여 대표 답변을 제공하고 정답이 존재하지 않는 경우 위키백과와 KorQuAD를 사용한 후처리 모듈 내에서 검색 후 답변을 제공함



<그림 5-4> 지능형 질의응답 엔진 흐름도

3. 국회법률도서관

□ 법률정보 통합 검색[20]

○ 국내외 법률자료를 체계적으로 수집하고 분석하여 양질의 법률정보를 제공하는 기능

- 국회 입법 과정에서 생산되는 법률안, 회의록, 심사보고서 등을 수집하고, 국회도서관의 법률 관련 자료를 '국가법률자원표준분류체계'에 따라 주제별로 자료를 분류함
- 사회적 관심 법안의 입법 배경과 논의 과정을 정리하여 소개하고, 법률쟁점DB를 통해 법률안이 제정·개정될 때 논의의 쟁점을 요약한 자료를 제공함
- 외국의 입법사례를 제공하기 위해 주요국의 법률을 수집하여 외국법률번역DB를 구축하고 최신외국 입법정보 서비스를 통해 국내외 법률정보를 얻을 수 있는 기능을 제공함

□ AI 일본법 번역기[21]

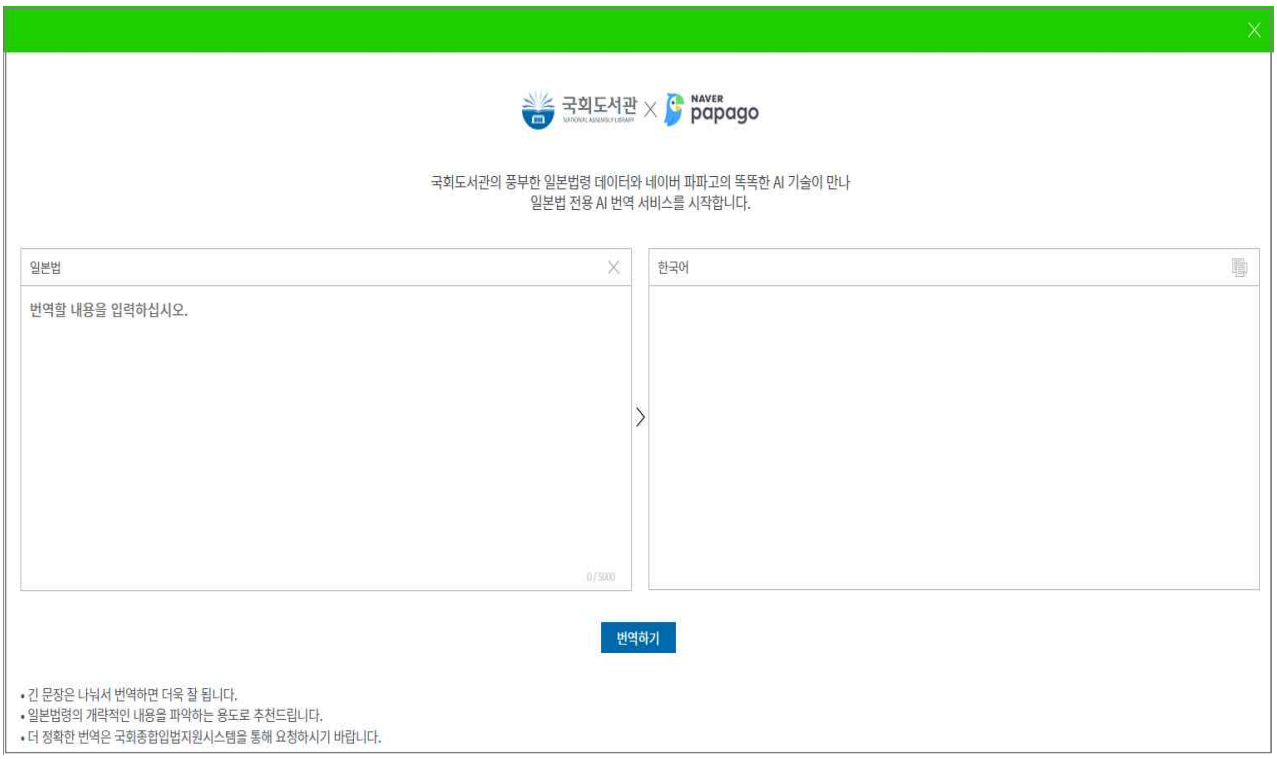
○ 국회도서관과 네이버가 공동개발한 일본법 자동번역 서비스

- 국회도서관은 약 20만 건의 일본법 학습데이터를 제공하고 네이버는 파파고 번역 서비스를

20) https://law.nanet.go.kr/index.jsp
21) https://law.nanet.go.kr/translation.do

기반으로 자동번역모델을 개발함

- 구축된 AI 일본법 자동번역기는 일본법 체계와 법률용어를 학습하여 국내 법률용어에 최적화된 번역결과의 산출이 가능함
- 국회의 입법 과정에 인공지능 기반 자동번역을 접목해 법률 분야의 인공지능 관련 산업 생태계 구축 인프라를 제공함



<그림 5-5> AI 일본법 번역기의 서비스 화면

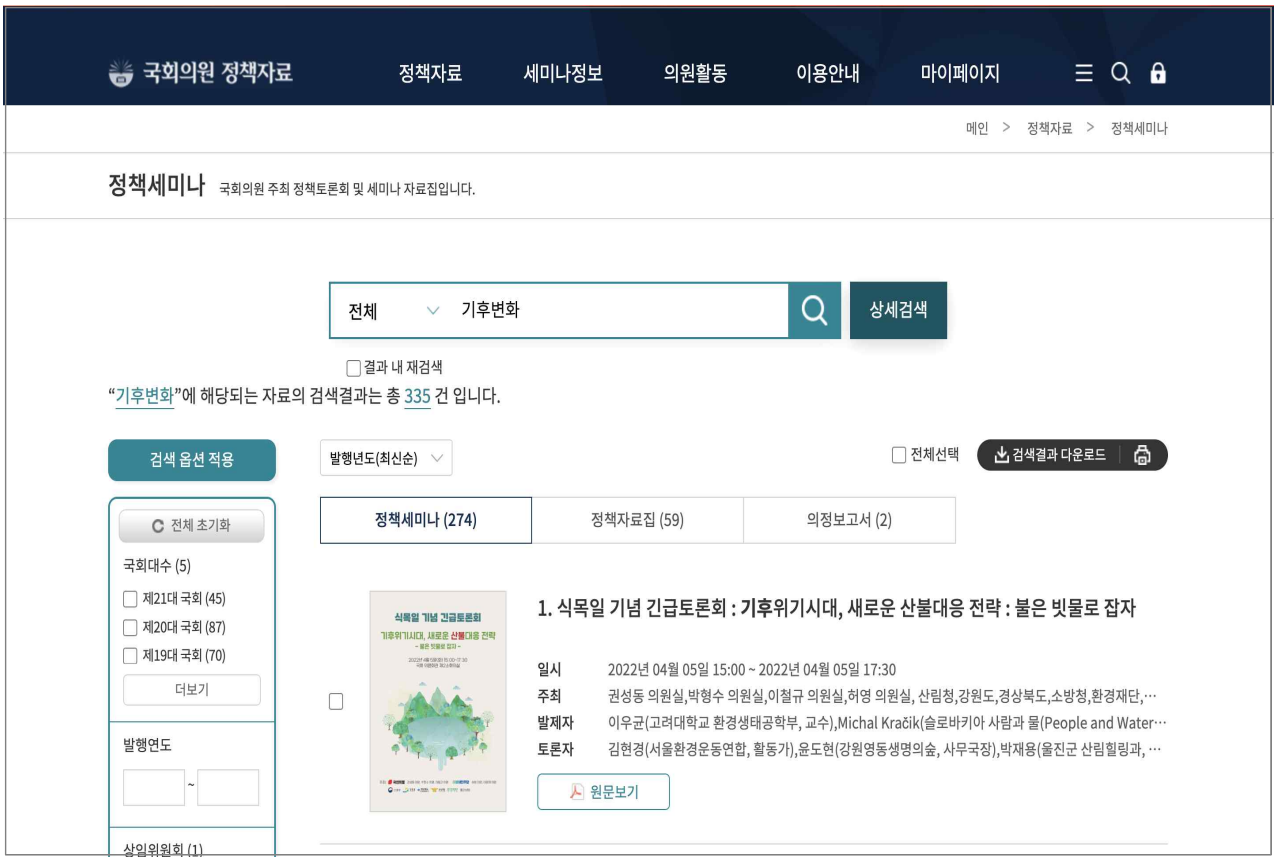
4. 국회의원 정책자료

□ 국회의원 정책자료 검색[22)]

○ 국회의원이 생산하는 정책자료의 열람과 다운로드 서비스를 제공하고, 정책자료의 키워드 기반 검색을 지원하는 기능

- 국회의원이 주최한 세미나 정보나 국회의원이 발간한 정책자료집, 의정 보고서 등의 정책자료 전반을 제공하는 데이터베이스 역할을 수행함
- 국회의원이 정책과 입법 개발을 위해 주최하는 세미나와 공청회 정보를 종합적으로 검색할 수 있는 맞춤형 검색 서비스를 제공함

22) https://ampos.nanet.go.kr:7443/main.do



<그림 5-6> 국회의원 정책자료의 '기후변화' 검색 결과

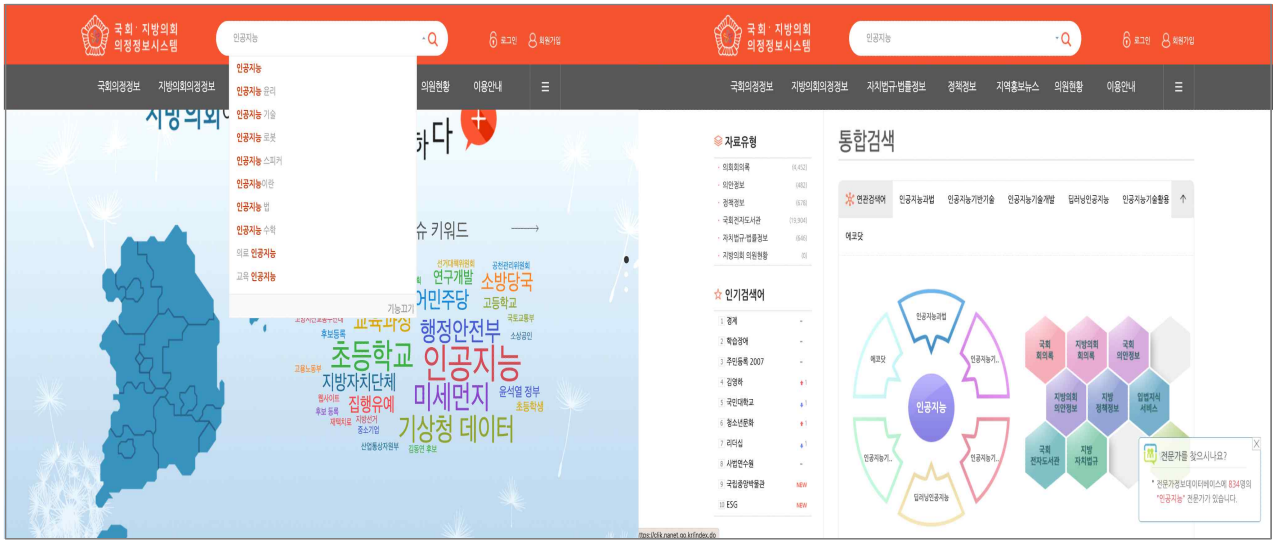
5. 국회 지방의회 의정정보시스템

□ 국회·지방의회 의정정보 통합 검색[23)]

○ 전국 지방의회의 의정자료를 국회 입법정보와 함께 통합적으로 검색할 수 있는 서비스를 제공함

- 전국 광역시도의회에서 발의된 각종 의안의 원문과 심사보고서, 회의록 등을 국회의 모든 의안 정보와 연계하여 지방의회와 국회의 정보를 한 번에 제공할 수 있음
- 통합 검색 시 텍스트 자동생성 기능을 제공하여 검색의 효율을 향상시킴(<그림 5-1> 왼쪽)
- 검색 결과, 연관 검색어 목록과 국회와 지방의회의 회의록, 의안정보 등을 제공하고, 전문가 정보데이터베이스와 연계하여 키워드와 관련된 전문가 정보를 함께 제공함(<그림 5-2> 오른쪽)

23) https://clik.nanet.go.kr/index.do



<그림 5-7> 국회 지방의회 의정정보 시스템의 인공지능 검색 결과

6. 국회기록보존소

□ 국회기록보존소 통합 검색[24]

○ 입법부의 영구기록물관리기관인 국회기록보존소의 일반문서, 국회간행물, 국회회의록, 국회의안 등의 기록물을 통합적으로 검색할 수 있는 서비스를 제공함

- 검색 결과에 따른 국회기록물 열람 서비스를 제공하고 국회 연표, 국회특별위원회 기록, 국회 주요기록, 역대 국회의장단과 주요 인사의 구술채록 검색을 제공

○ **(임시의정원 아카이브)** 국회기록보존소에서 수집한 임시의정원 관련 기록을 제공하는 디지털 아카이브로, 총 4,558건의 디지털 기록을 제공함

- 임시의정원 관련 사건의 타임라인을 제공하고, 타임라인의 사건과 관련된 기록물은 링크로 연결되어 탐색할 수 있음
- 국회기록보존소 검색 시 임시의정원 아카이브에서 제공되는 디지털 기록도 함께 검색할 수 있음

7. 전문가정보 데이터베이스

□ 국회의원/전문가 인적정보 검색[25]

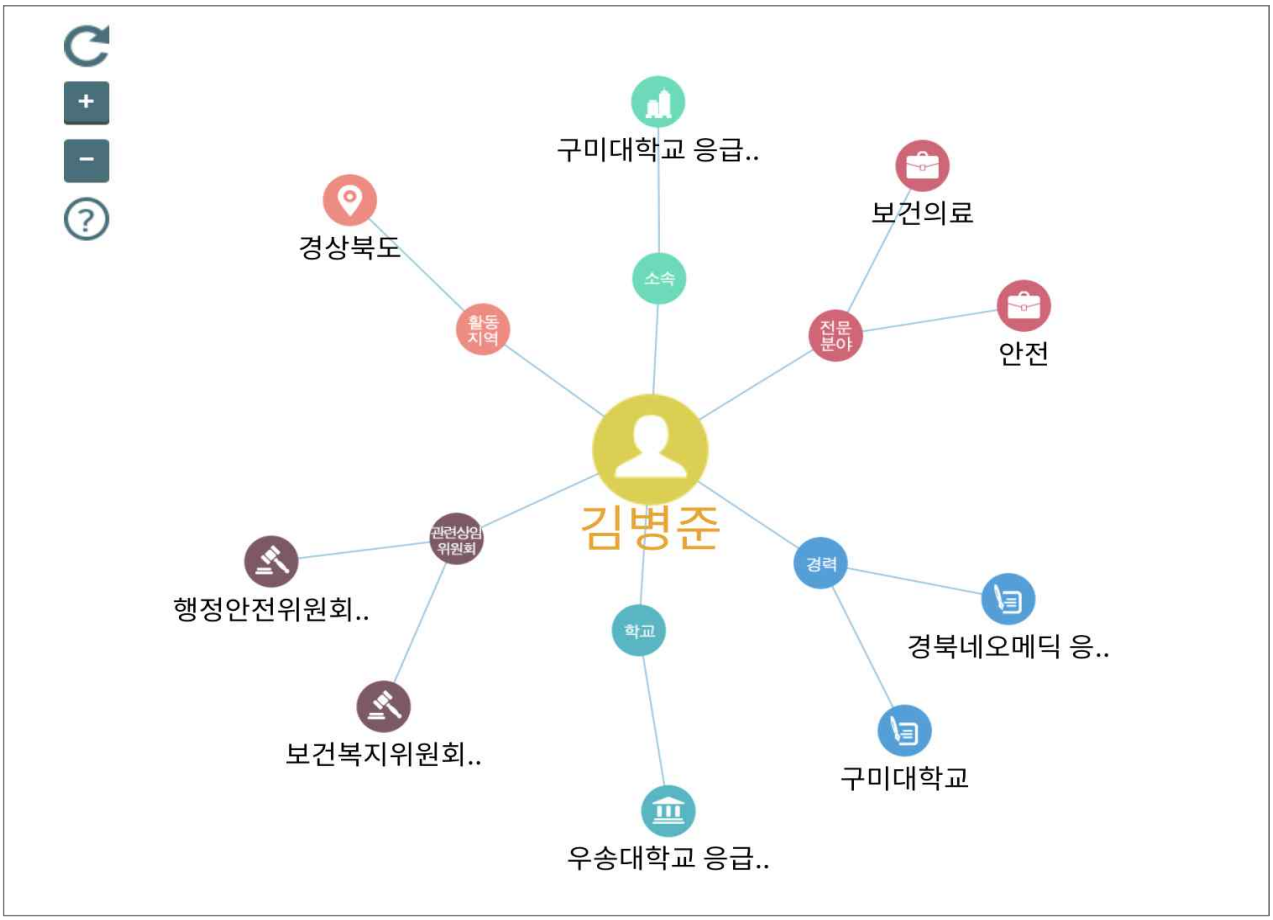
24) https://archives.nanet.go.kr/main.do
25) https://hn.nanet.go.kr/

○ 다양한 분야의 전문가 인적정보를 데이터베이스로 구축하는 사업으로, 국회입법활동 지원과 입법 방향성 제시에 중요한 역할을 맡고 있으며 전문가의 지식과 경험을 공유할 수 있는 장

- 국회의원 찾기 서비스는 국회의원명, 국회대수, 소속정당, 관련 위원회 기준과 선거구를 기준으로 검색 서비스를 제공함
- 전문가 찾기 서비스는 전문가명, 전문분야, 활동지역, 전체검색어를 기준으로 검색 서비스를 제공함
- 전문가 참여마당 탭을 이용하여 등록된 전문가들의 신간과 채널, 최근 동정 정보를 제공함

○ 각 국회의원과 전문가 상세정보 페이지를 통해 인물에 대한 소속, 연령대, 전문분야 등 자세한 정보를 제공하며 그래프 데이터베이스 형태로 관련상임위원회, 연구 키워드, 활동지역, 학교, 직업 정보 등을 한눈에 알아볼 수 있는 관계정보를 제공함

- 관계정보를 확인할 수 있다는 그래프 데이터베이스의 이점을 이용하여, 연구 키워드나 학교 노드를 클릭하면 관련 있는 인물들의 노드가 추가로 생성되어 인물들의 목록을 확인할 수 있고 인물 상세정보 페이지로 바로 이동 가능한 링크를 제공함



<그림 5-8> 전문가 상세보기에서 제공하는 그래프 데이터베이스 정보

8. 지능형 의회정보 융합분석 시스템 Argos

□ 키워드 기반 이슈 분석[26)]

○ 뉴스, 소셜미디어 등 외부자료에 대한 빅데이터 분석을 통해 국회 입법 활동에 참고 가능한 이슈를 추출하고, 이슈 키워드, 국회 활동, 국회의원별 분석 결과를 제공함

- 국회 17개 상임위와 관련성이 높은 핵심 이슈를 파악하고, 해당 이슈의 일간, 주간, 월간 추이를 비교 분석하는 기능
- 소셜미디어는 트위터, 블로그, 커뮤니티, 페이스북이 포함됨

○ **(키워드 검색)** 기본검색은 키워드 검색으로 검색어 입력 시, 검색어에 대한 이슈 키워드 목록을 제공하고 해당 키워드 기반의 분석 결과를 제공함

- 상세검색 기능은 최대 10개 검색어와 연산자(||, && 등) 사용을 지원하고, 분석 기간을 한 달 간격으로 조정할 수 있음



<그림 5-9> 이슈 키워드 예시

○ **(텍스트 분석)** 수집 데이터와 국회도서관 소장자료를 바탕으로 이슈 키워드와 관련된 텍스트 분석을 수행하고, 분석 결과를 인포그래픽, 목록으로 제공함

- 2020년 최신 머신러닝 분석엔진을 적용하여 팩트북, 싱크탱크 등의 소장자료 학습을 통해 이슈 분류의 기능을 개선함
- **(이슈 히스토리)** 이슈 키워드가 등장하는 빈도와 연관어를 고려해 정확도가 가장 높은 일자별 대표 뉴스를 선별하여 일자, 뉴스 제목, 내용 일부, 신문사명, 댓글을 제공하고, 클릭 시 원문으로 이동할 수 있음
- **(언급량 추이)** 순위가 가장 높은 이슈 키워드 또는 이용자가 선택한 이슈 키워드에 대한 언급량 추이를 조회할 수 있으며, 결과는 선형, 막대형, 영역형 중에서 선택할 수 있음

26) https://argos.nanet.go.kr/main/monitoring/simpleMonitoring.do

- **(연관어 분석)** 이슈 키워드에 대한 연관어 분석 결과를 워드 클라우드 형태로 제공하고, 자료 구분(SNS, 소장자료)과 분류 카테고리(국회의원, 인물, 법/제도, 기관/단체, 장소, 시사, 생활/문화)별로 분석 범위의 필터링이 가능함



<그림 5-10> 연관어 분석 결과

- **(뉴스 데이터와 소장자료)** 이슈 키워드가 포함된 관련 뉴스, 소셜 데이터의 원문 목록과 소장자료 목록을 제공하고, 다운로드 버튼을 통해 테이블 형태의 데이터를 제공함
- 소장자료 목록은 법률도서관, 공공정책자료, 정책 세미나, 의정자료 공유, 전자도서관으로 원문 출처가 구분됨
- **(TOP 인용문)** 수집한 뉴스 데이터에서 이슈 키워드와 관련된 인용문을 추출하여 상위 5개 인용문과 발화자에 대한 정보를 제공함

□ 의정활동 분석

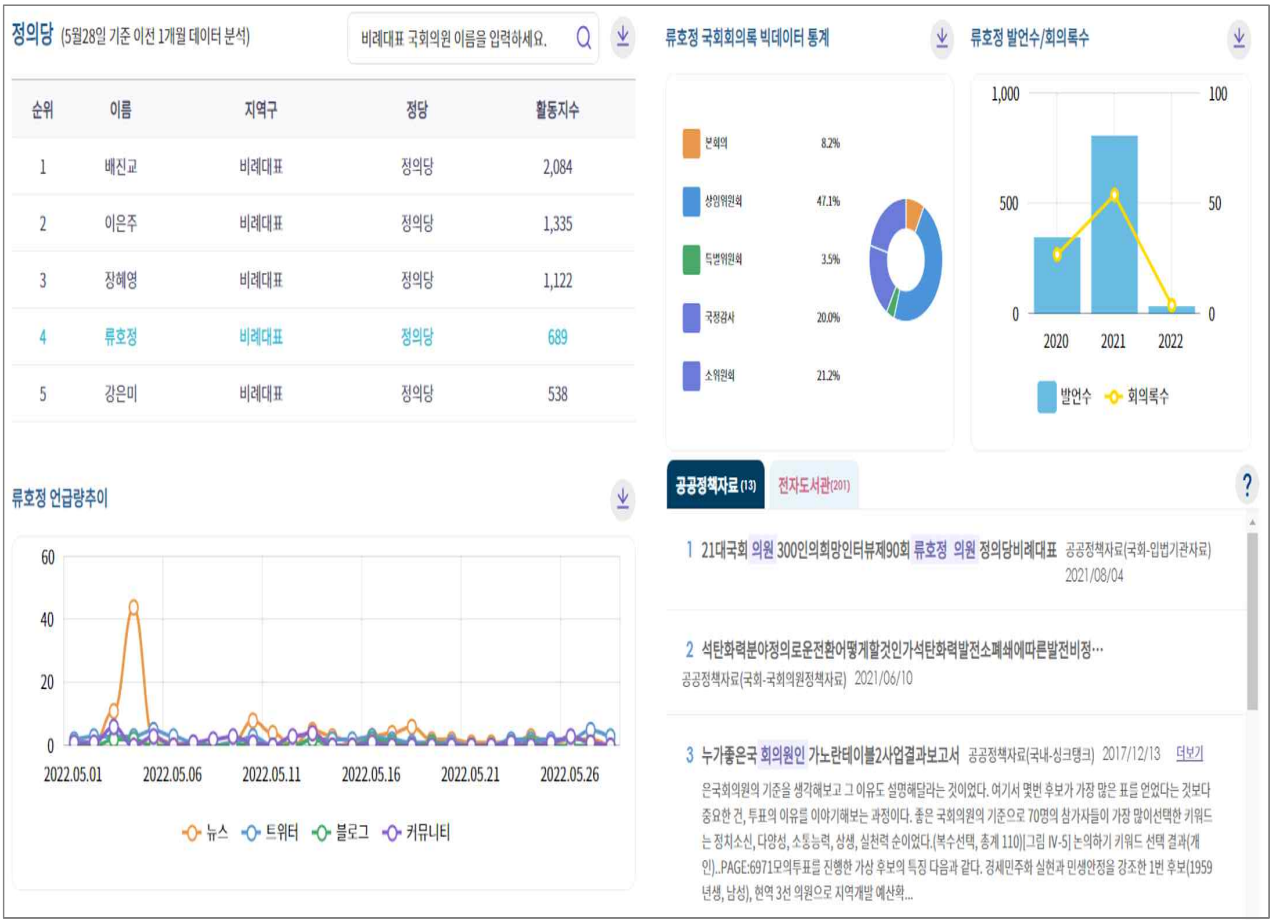
○ 국회 관련 뉴스(200여 개 신문사, 60여 개 지역 신문사), 트위터, 블로그, 커뮤니티 등 SNS 언급량을 기준으로 국회의원, 정당, 법률, 상임위원회별 활동 현황 분석 결과를 제공함

○ **(국회의원 활동지수 현황)** 국회의원의 활동지수*와 언급량을 지역구, 비례대표, 당선

횟수별로 탐색 또는 검색할 수 있는 서비스[27]

* 국회의원 활동지수: 1개월간 뉴스, SNS에서 각 국회의원의 언급량을 기준으로 산출하되, 뉴스 언급량에 큰 가중치를 둠

- 국회의원별 1개월분의 데이터를 사용하여 분석하며, 분석 결과는 1주 단위로 갱신됨
- 지역구별 국회의원 탐색을 위한 지도 시각화를 제공함
- 검색 결과, 국회의원 관련 정보와 활동지수, 1개월 동안 SNS별 언급량 추이, 관련 뉴스 목록, 소장자료 목록을 제공하고, 국회회의록 빅데이터를 기반으로 발언 수, 회의록 수에 대한 통계정보를 제공하는 서비스
- 분석 결과는 웹상에서 인포그래픽 형태로 시각화되며, 다운로드 버튼을 통해 엑셀(XLSX) 형태로 다운로드가 가능함



<그림 5-11> 비례대표 검색 결과 일부

○ **(이슈법률 Top 10)** 최근 이슈가 된 법률에 대한 주요 뉴스, 연관어 분석 결과 등의 정보를 제공하는 기능[28]

- 법률 관련 1개월분의 데이터를 기준으로 분석하며, 분석 결과는 1주 단위로 갱신됨

27) https://argos.nanet.go.kr/main/fusionanalysis/snsDataStatusGuest.do
28) https://argos.nanet.go.kr/main/fusionanalysis/mentionLawTopGuest.do

- 언급량을 기준으로 이슈법률 상위 10개 목록을 제공하고, 선택한 법령에 대한 뉴스와 SNS별 언급량 추이, 연관어 분석 결과와 관련 뉴스 목록을 제공함

○ **(정당별 의정활동 현황)** 6개 정당에 대한 주요 뉴스, 관련 인물 등의 정보를 제공하는 기능[29)]

- 기본소득당, 국민의힘, 정의당, 더불어민주당, 시대전환을 포함한 6개 정당별 1주일 데이터를 바탕으로 분석하며, 분석 결과는 1주일 단위로 갱신됨
- 선택한 정당의 관련 뉴스 목록, 주요 언급 인물에 대한 워드 클라우드, 뉴스와 SNS별 언급량 추이 정보를 제공함

○ **(상임위원회별 데이터 분석)** 17개 상임위원회*와 관련한 주요 뉴스, 최신발의 법률 등의 정보를 제공하는 기능[30)]

* 17개 상임위원회: 국회운영위원회, 법제사법위원회, 정무위원회, 기획재정위원회, 교육위원회, 과학기술정보방송통신위원회, 외교통일위원회, 국방위원회, 행정안전위원회, 문화체육관광위원회, 농림축산식품해양수산위원회, 산업통상자원중소벤처기업위원회, 보건복지위원회, 환경노동위원회, 국토교통위원회, 정보위원회, 여성가족위원회

- 상임위원회별 언급량 추이, 소속 국회의원 목록, 관련 뉴스, 최신발의 법률 목록을 검색 결과로 제공하고, 그중 소속 국회의원 관련 정보는 Argos의 국회의원별 광장 서비스와 연결되어 있으며, 최신발의 법률은 의안정보시스템과 링크로 연결됨

○ 해당 키워드와 관련된 텍스트 분석을 바탕으로 관련 국회의원, 일자별 뉴스, 키워드 언급량 추이, 연관어 분석, 관련 뉴스 목록, 소장자료 목록, 인용문, 통계정보 등 다양한 분석 결과를 제공함

□ 국회의원별 광장[31)]

○ Argos에서 직접 구축한 데이터와 국회 내 자료를 종합하여 개별 국회의원 관련 정보를 제공하는 기능

- 국회의원 이름 또는 지역명 검색과 6개 정당과 무소속 필터링을 통해 탐색 가능함
- 해당 국회의원에 대한 기본 정보(예: 이름, 생년월일, 정당, 선거구, 약력, 사무실 정보, 보좌관 등)와 국회회의록 빅데이터, 관련 국회의원과 키워드 분석 결과를 시각화함

29) https://argos.nanet.go.kr/main/fusionanalysis/partyActivitiesGuest.do
30) https://argos.nanet.go.kr/main/fusionanalysis/committeeDataGuest.do
31) https://argos.nanet.go.kr/main/fusionanalysis/dataSquareGuest.do



<그림 5-12> 국회의원별 광장 예시

□ 빅데이터 분석

○ 사용자가 입력한 검색어를 바탕으로 최신 뉴스 데이터를 텍스트 마이닝 할 수 있도록 지원하는 서비스로, 상세 분석과 비교분석을 지원하는 기능

- 검색어 입력 시 최대 10개 검색어와 연산자(예: ||, &&, ())를 사용할 수 있고, 분석 기준일자를 설정할 수 있음
- 도출된 이슈 키워드에 대한 기존 분석 결과를 제공하는 키워드 기반 이슈 분석서비스와는 달리, 이용자가 원하는 검색어를 바탕으로 소셜 데이터 제공사로부터 분석 결과를 실시간으로 제공함

○ **(상세분석)** 최근 한달 동안의 뉴스를 바탕으로 사용자 검색어에 대한 언급량 추이, 연관어, 뉴스 원문 목록 제공을 제공하는 기능[32)]

- 검색어에 대한 뉴스 언급량 추이를 테이블 형태의 목록 또는 선, 막대그래프로 시각화하여 제공하고, 사용자가 자유롭게 기간을 설정할 수 있도록 인터랙티브 그래프로 구현함
- 연관어 분석은 입력한 키워드에 대한 SNS 데이터와 국회도서관 소장자료를 분석해 최대 50자의 연관어를 워드 클라우드 형태로 시각화하고, 자료 구분(SNS, 소장자료)과 분류 카테고리별로 필터링 기능을 제공함

32) https://argos.nanet.go.kr/main/analysis/detailAnalysisGuest.do

- 검색어와 연관된 뉴스 원문, 소장자료 원문 목록을 제공하고, 원문 조회 시 검색어를 강조하여 표시함

○ **(비교분석)** 최근 한달 동안의 뉴스를 바탕으로 2가지 검색어의 연관어를 비교분석하고, 검색어 사이의 관계를 파악하는 기능[33)]

- 검색어의 언급량 추이에 대한 선 그래프, 언급량 비율에 대한 파이 차트를 통해 2개 검색어의 언급량 비교분석을 지원함
- 검색어에 대한 연관어 네트워크 그래프를 통해 개별 검색어의 연관어와 두 검색어 사이의 연관어를 동시에 파악 가능함



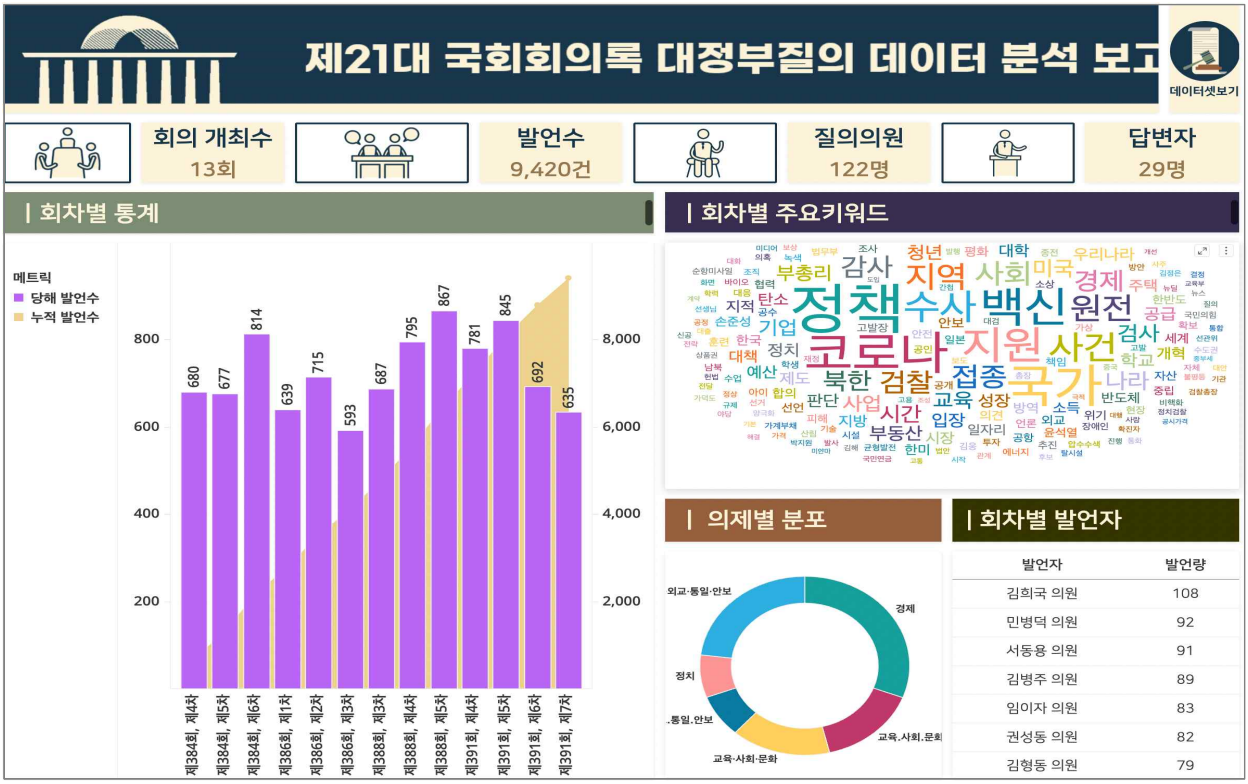
<그림 5-13> 비교분석 결과 예시

□ Argos 스토리보드

○ Argos 스토리보드를 이용하여 대수별 국회회의록 대정부질의 데이터를 제공하는 기능

- 전체 탭은 전체 회의 현황과 회차별 통계정보, 주요 키워드정보를 분석해 시각화하고, 의제 분포와 발언자 목록에 대한 정보를 제공함
- 분야별 탭은 국회의원과 답변자별 발언 내용 키워드를 워드 클라우드로 제공함
- 연도/차수별 탭은 날짜와 차수를 지정하여 발언한 국회의원과 답변자 정보, 키워드를 워드 클라우드로 구현함

33) https://argos.nanet.go.kr/main/analysis/compareAnalysisGuest.do



<그림 5-14> Argos 스토리보드와 연계한 국회회의록 데이터 분석 자료

9. 국회회의록 빅데이터

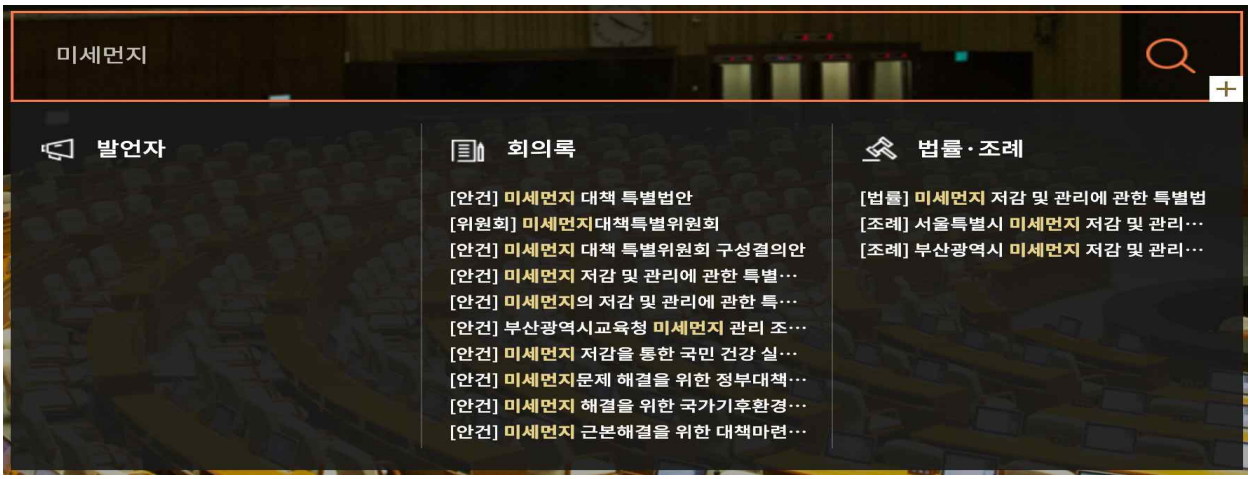
□ 국회회의록 데이터 검색[34]

○ 인공지능/빅데이터 기술을 활용하여 제16대부터 제21대 국회까지의 국회회의록 약 2만 건과, 의원별 발언 단위로 분리한 1,200만여 건의 데이터를 구축하고 내용 분석을 통한 시각화 프로그램을 제공하는 서비스[35]

○ 통합 검색 서비스를 이용해 발언자명, 발언 내용, 의원회명, 회의명, 안건, 법률명 검색이 가능하며 발언자, 회의록, 법률/조례 기준으로 자동완성 기능을 제공함

- 발언자를 검색할 경우, 자동완성 창에 뜨는 특정 발언자를 클릭하여 검색하면 해당 발언자에 대한 정보만 제공하지만, 검색 버튼을 클릭하여 검색하면 키워드로 인식되어 다른 발언자의 발언에 언급되었을 경우 발언자 정보를 추가하여 제공함

34) https://dataset.nanet.go.kr/
35) https://www.readersnews.com/news/articleView.html?idxno=103992



<그림 5-15> 자동완성 예시

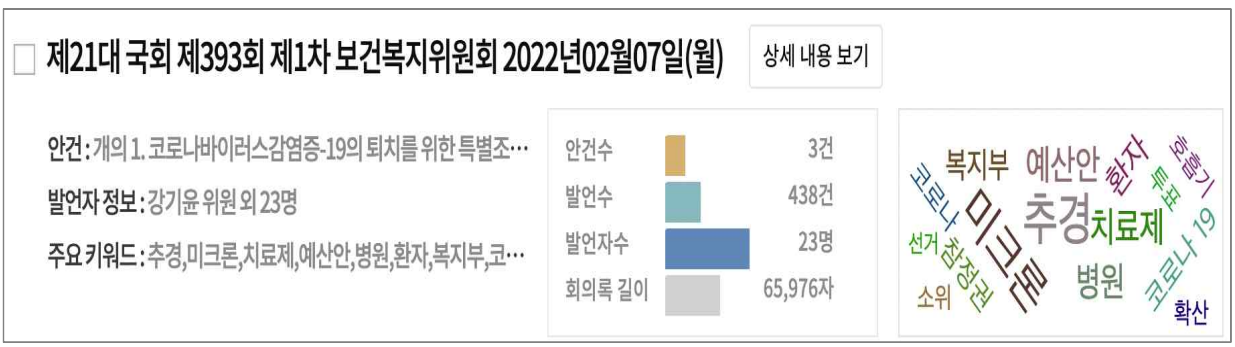
○ 검색 결과는 발언자 탭과 회의록 탭으로 나눠서 제공함

- 발언자 탭은 대수별로 인물의 발언 수와 회의록 수를 집계한 표와 회의록 통계와 발언 수/ 회의록 수를 시각화한 그래프 정보를 제공함



<그림 5-16> 발언자 탭 검색 결과 예시

- 회의록 탭은 회의마다 회의명, 안건, 발언자 목록, 키워드 목록 정보를 제공하며 통계정보와 회의의 키워드를 시각화한 자료 확인이 가능함



<그림 5-17> 회의록 탭 검색 결과 예시

- 각 시각화 자료에 대한 데이터셋과 시각화 이미지를 다운로드할 수 있는 기능을 제공함
- 각 탭은 특정 발언자 또는 회의록을 선택하여 엑셀 형식으로 데이터셋을 다운로드할 수 있는 기능을 제공하며, 국가학술정보 분석서비스와 연계하여 데이터 분석을 수행함



<그림 5-18> 국가학술정보 분석서비스와 연계한 데이터 분석 서비스 제공 기능 예시

○ 발언자 기본정보 페이지는 인물에 대한 기본정보와 발언 정보를 제공함
- 발언 정보를 그래프와 워드 클라우드 형태로 제공하며, 회의 구분별 발언 회의록 수와 발언 내 법률 리스트 정보의 확인이 가능함
- 발언 정보 데이터셋 다운로드와 분석 기능을 제공함

○ 회의록 기본정보 페이지는 회의에 대한 기본적인 정보의 확인이 가능함
- 그래프 형태로 발언자별 발언 수와 워드 클라우드 형태로 회의록 키워드, 법률정보와 대화 형식의 발언 정보를 제공함

제 6 장 국회전자도서관 도입을 위한 인공지능 서비스 모델 제안

1. 국외 도서관 서비스 벤치마킹

□ 국외 주요도서관 서비스 분석

○ 미국, 일본, 프랑스, 네덜란드 주요도서관 서비스에서 검색 기능을 제공하고 있음

- 일본 MIMA 검색, 네덜란드 DBNL ngramviewr를 제외한 5개 검색 서비스에서 상세검색 기능 지원. 특히, 일본 재팬 서치는 검색어에 대한 자동완성 기능을 제공하고 있음
- 미국의 Newspaper Navigator와 프랑스의 GallicaPix는 이미지 인식 기술을 적용하여 멀티미디어 자료를 검색할 수 있는 기능을 제공하고 있고, 이미지 데이터로부터 추출한 메타데이터를 이미지 자료 검색에 활용하고 있음

○ 국외 주요도서관 대부분은 문서 분류와 이미지 인식 기술을 활용하고 있음

- 일본 NDC Predictor, 노르웨이 도서관은 표준십진분류법에 머신러닝 기술을 적용하여 분류의 자동생성을 지원함
- 광학문자인식 (OCR) 기술은 디지털 자원의 변환과 생성을 위해 광범위하게 활용되고 있음. 특히, 신문 자료에서 텍스트 데이터의 추출과 신문 자료의 자동분류를 목적으로 적용됨
- 스웨덴 Advertisement Classification은 딥러닝 기반 이미지 인식 기술을 적용하여 디지털로 변환된 신문기사에서 일반 기사문과 광고문을 자동으로 분류하는 서비스를 제공함
- 미국 Newspaper Navigator, 일본 차세대 디지털 도서관은 머신러닝 기반 이미지 처리 기술을 시각 자료의 검색에 활용하고 있음. 특히 검색 결과에 연관 이미지를 함께 제공함

○ 미국의 Citizen DJ에서 음성 인식 서비스를 제공함

- Citizen DJ는 의회도서관 소장 오디오와 동영상 자료를 탐색하고 편집할 수 있는 서비스이며, 서비스 구현을 위해 파이썬 머신러닝 라이브러리인 Scikit-learn을 활용함

○ 지식그래프 기반 서비스는 5개로, 데이터의 연계를 위한 온톨로지를 설계하고 통합검색을 위한 지식베이스를 구축하는 데 활용되고 있음

- 일본 재팬 서치는 지식그래프 기술을 적용하여 다양한 기관이 제공하는 디지털 아카이브, 도서관이 보유하고 있는 디지털 자원을 연계하여 통합 검색을 지원하고 있음
- 핀란드는 다양한 주제 분야 연계를 위해 온톨로지와 어휘 사전을 구축하여 문서 자동색인을 위한 서비스 FintoAI 개발

[표 6-1] 국외 주요도서관의 서비스 분석 결과 (●: 기능 제공, ▲:일부 기능 제공)

	서비스명	정보 검색	질의 응답 (QA)	문서 요약	문서 분류	이미지 인식	음성 인식	기계 번역	지식 그래프
미국	Newspaper Navigator	●				●			
	Citizen DJ						●		
일본	차세대 디지털 도서관	●				●			
	NDC Predictor				●				●
	MIMA 검색	▲							
	재팬 서치	●							●
프랑스	GallicaPix	●				●			
스웨덴	Swedish BERT								
	Swedish spaCy								
	Advertisement Classification				●	▲			
	스웨덴-노르웨이 연합 언어모델								
핀란드	Annif				●				
	Finto								●
	FintoAI				●				●
네덜란드	Demosaurus				●				
	Frame generator				▲				
	DBLNL ngramviewer	▲							
	PoliMedia	▲							●
노르웨이	머신러닝을 이용한 듀이 십진분류법 생성				●				

□ 국회전자도서관 서비스 분석

○ 모든 웹사이트에서 정보검색을 위한 기능을 제공하고 있고, 국회전자도서관의 지능형 질의응답을 통해 QA 서비스를 제공함

- 국회전자도서관, 국회·지방의회 의정정보시스템, 국회회의록 빅데이터 서비스는 상세검색 기능과 자동완성 기능을 제공하고 있음
- 국회법률도서관, 국회의원 정책자료, 국회기록보존소, 전문가정보 데이터베이스, 지능형 의회정보 융합분석 시스템 Argos는 정확히 일치하는 키워드만 검색할 수 있음
- 국회전자도서관의 지능형 질의응답 서비스는 키워드 매칭 기반의 검색을 넘어서 사용자 질문을 이해하고 적합한 답을 찾아 제공함. 국회도서관이 보유한 자료와 위키백과를 기반으로 학습데이터를 구축하여 답변과 함께 출처 정보를 제공함

○ 국회전자도서관, 국회·지방의회 의정정보시스템, 지능형 분석(Argos)에서 검색 결과를 토대로 자료를 분류하여 시각화하는 서비스를 제공함

- 국회전자도서관은 키워드와 유사한 연관어를 탐색하여 연관어 분석, 연관어 인물, 연관어 빈도를 시각화한 인포그래픽 서비스를 제공함
- 국회·지방의회 의정정보시스템은 키워드 검색 시 관련 연관어를 시각화하여 제공하지만 검색 결과를 분석하는 서비스는 제공하고 있지 않음
- 지능형 분석(Argos)는 이슈 키워드에 대한 언급량 추이, 연관어 분석을 시각화함. 한편, 최근 의정활동을 분석하여 국회의원별 언급량, 정당별 언급량, 상임위별 언급량, 법률별 언급량을 분석한 시각화를 제공하고 있음

○ 이미지 인식과 기계번역 기술은 특정 서비스에만 적용되고 있음

- 이미지 인식 기술은 광학문자인식(OCR)을 활용한 원문 읽어주기 서비스(국회전자도서관)에 적용되어 있고, 이미지 기반 검색과 유사한 이미지 검색 또는 추천 기능을 제공하지 않음
- 기계번역은 국회법률도서관의 AI 일본법 번역기에서만 구현되어 있음

○ 국회는 국가학술정보 클라우드 환경을 구축하여 지식그래프 기반 검색 서비스를 제공하고 있음

- 학술기사, 학술지, 단행본 RDF 데이터, 연구자 RDF 데이터, 주제어 RDF 데이터를 통해 국가학술정보 클라우드 환경 구축
- 국회회의록 빅데이터와 연계하여 검색 서비스를 제공하지만, 국회도서관에서 구축한 데이터베이스와 통합적으로 관리하고 있지 않음

[표 6-2] 국회전자도서관의 서비스 분석 결과 (●: 기능 제공, ▲:일부 기능 제공)

서비스	정보 검색	질의 응답 (QA)	문서 요약	문서 분류	이미지 인식	음성 인식	기계 번역	지식 그래프
국회전자도서관								
국회도서관 소장자료 검색	●			●	▲			●
지능형 질의응답		●						
국회법률도서관								
법률정보 통합 검색	▲							
AI 일본법 번역기							●	
국회의원 정책자료								
국회의원 정책자료 검색 서비스	▲							
국회 지방의회 의정정보시스템								
국회·지방의회 의정정보 통합 검색	●			▲				
국회기록보존소								
국회기록보존소 통합 검색	▲							
전문가정보 데이터베이스								
국회의원/전문가 인적정보 검색	▲							●
지능형 분석 (Argos)								
키워드 기반 이슈 검색	▲			●				●
의정활동 분석	▲			●				
국회의원별 광장	▲			●				●
빅데이터 분석	▲			●				
Argos 스토리보드								
국회회의록 빅데이터								
국회도서관 회의록 데이터 검색	●							●

□ 벤치마킹 결과

○ 멀티미디어를 포함한 디지털 자원을 검색할 수 있는 기능을 지원해야 함

- 국회전자도서관 서비스 14개 중 11개 서비스에서 검색 서비스를 제공하고 있으나, 키워드 기반의 텍스트 검색 기능에 한정되어 있음
- 미국 Newspaper Navigator, 일본 차세대 디지털 도서관의 사례와 같이 이미지, 음성 인식 기술을 기반으로 멀티미디어 자료 검색 서비스를 제공하여 이용자의 접근성을 높이고, 편집 기능을 제공해 멀티미디어 유형의 자료를 활용할 수 있는 기반을 마련하는 것이 필요함

○ 분산적으로 존재하는 정보의 통합적 검색을 위한 방안을 마련해야 함

- 법률정보, 국회의원 정책자료, 국회·지방의회의 의안과 회의록은 국회도서관의 다양한 사이트에서 분산적으로 존재하기 때문에, 이용자가 해당 정보의 맥락과 연결된 자료들을 탐색하는 데 한계가 존재함
- 예를 들어, 국회회의록 빅데이터는 국회의원별 회의록 내용을 면밀히 파악할 수 있지만, 회의 결과 의안의 진행 방향을 파악하기 어려움. 의안정보시스템은 의안의 진행 과정을 알 수 있지만, 의안을 발의한 국회의원의 풍부한 정보나 의안 처리과정에서 발생하는 다양한 논의를 파악할 수 없음
- 일본 국립국회도서관의 재팬 서치는 다양한 기관의 자료를 연계한 검색 결과를 제공하여 풍부한 검색 결과를 제공함. 재팬 서치와 같이 국회도서관의 분산된 정보의 연계를 통해 의안 발의에서 법률안 공표에 이르는 과정의 다양한 맥락 정보를 검색할 수 있고, 통합적인 검색 결과를 제공할 수 있음

○ 텍스트 요약 기술을 활용한 검색 결과의 요약 서비스를 제공하는 것이 필요함

- 현재 대부분의 사이트에서 소장자료에 대한 검색을 제공하고 있으나, 검색 결과는 목록 형태로 메타데이터와 원문의 일부 정보만 제공하고 있음
- 원문 자료의 요약 정보를 제공해 이용자가 검색 결과를 한눈에 이해하고, 원하는 결과의 빠른 탐색이 가능함. 특히, 국회·지방의회 회의록이나 법률 쟁점, 법률안에 대한 요약 서비스를 통해 국회의원의 입법 활동을 지원할 수 있음
- 특히, 전문 서지 검색 결과에는 복합 어휘 형태로 표현되는 개체명, 법령 용어, 기술명 등이 다수 포함되어 있으므로 이러한 핵심 개체가 포함된 한국어 텍스트 요약 서비스가 필요함
- 이를 위해서 전문용어 사전, 개체명 사전 등의 방대한 한국어 전문 코퍼스 구축을 통해 텍스트 요약을 위한 AI 모델의 성능을 높일 수 있어야 함
- 사용자 서비스의 만족도 향상뿐 아니라 구축된 사전 및 요약 기술은 한국어 자연어처리 AI 모델의 기반 기술로 적용할 수 있음

○ 기계번역 적용 범위의 확대를 통한 회답 서비스의 효율성을 제고해야 함

- 국회전자도서관은 일본법 체계와 일본 법률용어에 특화한 번역 서비스인 AI 일본법 번역기

를 개발하고, 국회도서관은 의회·법률정보 회답 서비스의 일환으로 전문조사관, 업체를 통한 다양한 언어권의 정책참고자료, 법률정보 번역 서비스를 제공함

- 외국법률번역DB의 텍스트 데이터를 통해 기계번역 모델을 개발·학습하여 다국어 번역 서비스를 제공하여 기존 회답 서비스의 적용 범위를 확장하고, 서비스 효율성과 이용자 만족도를 높일 수 있음

2. 서비스 도입을 위한 분석 프레임워크

□ 성과와 도입 난이도에 따른 서비스 모델의 분석 프레임워크 적용

○ 페이오프 매트릭스(Pay-Off Matrix; POM)* 기준으로 국회전자도서관이 도입 가능한 인공지능 서비스 모델의 현황을 제공함

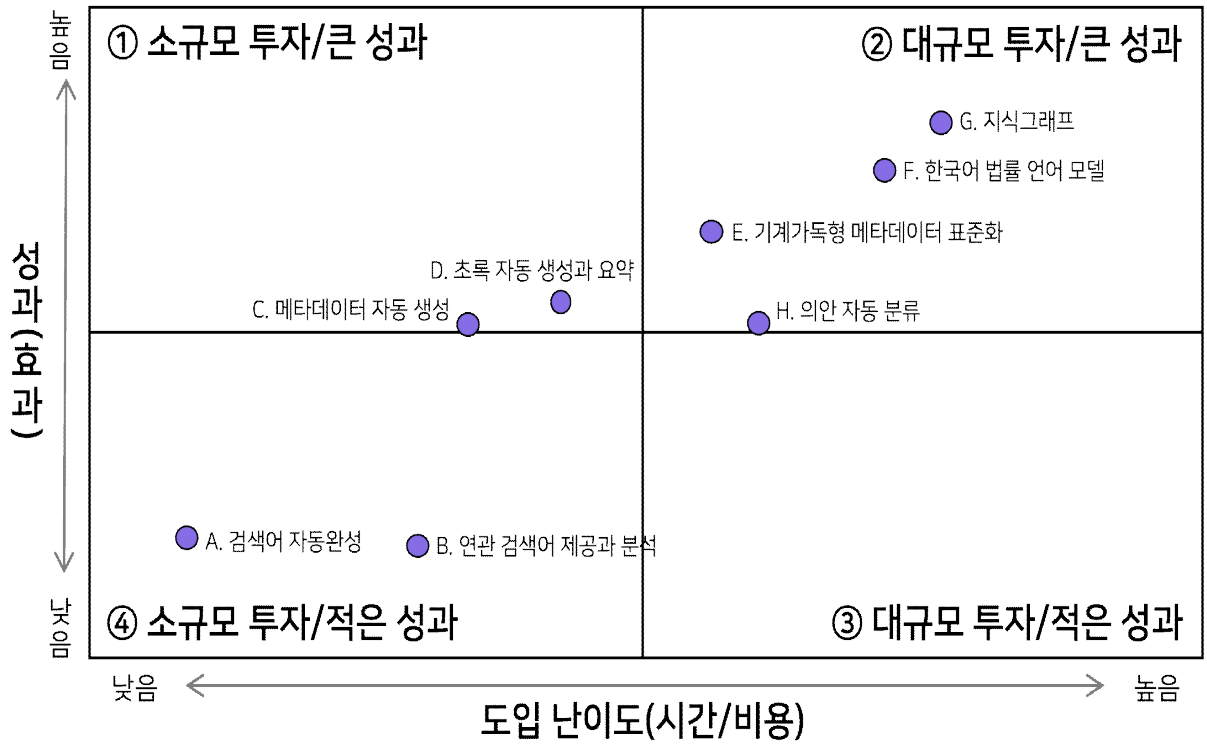
* 페이오프 매트릭스: 팀이나 기업의 의사결정 과정에 사용되는 도구로, 행동경제학자 다니엘 카네만(Daniel Kahneman)의 제한적 합리성을 기반으로 최선의 의사결정을 할 수 있도록 도움

- 성과(효과)는 서비스의 도입으로 예상되는 효과의 정도를 평가하고, 도입 난이도(시간/비용)는 서비스 도입 시 예상되는 개발 시간과 투자의 정도를 평가함

※ [표 6-3]은 국회전자도서관에 도입할 수 있는 인공지능 기반 서비스 목록임

[표 6-3] 도입 가능한 인공지능 기반 서비스 목록

항목	서비스 이름	서비스 설명
A	검색어 자동완성	검색창에 단어 입력 시 예상되는 검색어를 추천해주는 기능
B	연관 검색어 제공과 분석	검색창에 단어 입력 시 검색어와 연관된 단어를 제공하고, 연관 검색어에 대한 분석과 시각화를 제공하는 기능
C	메타데이터 자동 생성	디지털 자원의 메타데이터를 자동으로 생성해주는 기능
D	초록 자동 생성과 텍스트 요약	텍스트 내용을 분석하여 자동으로 초록을 생성하는 기능
E	기계판독형 메타데이터 표준화	메타데이터를 일관적으로 기술하는 체계를 구축하고, 기계가 읽고 처리할 수 있는 형태로 메타데이터를 구축하는 서비스
F	한국어 법률 언어모델	국회전자도서관이 서비스하는 법률 텍스트 데이터를 기반으로 법률에 특화된 한국어 언어모델을 구축하는 서비스
G	의안 정보 통합 지식그래프	의안 정보를 의미적으로 연계하고, 인공지능이 처리 가능한 그래프 형태의 지식베이스를 구축하는 서비스
H	의안 자동 분류	유사도를 기반으로 의안을 자동 분류하고, 특정 의안과 관련된 의안을 제공하는 서비스



<그림 6-1> 서비스 모델의 페이오프 매트릭스 현황

○ 페이오프 매트릭스의 사분면은 도입 난이도 대비 성과에 따라 구분되고, 각 사분면은 아래와 같이 해석될 수 있음

※ 각 서비스 모델이 페이오프 매트릭스에 매핑된 현황은 <그림 6-1>에서 확인

- **(1. 소규모 투자/큰 성과)** 이미 구현도가 높아 유지하면 되는 서비스 또는 비교적 쉽게 구현할 수 있지만, 아직 지원되지 않은 서비스가 포함된 영역임. C와 D는 구현이 어렵지 않지만, 현재 국회전자도서관에서 제공하지 않는 서비스 모델임
- **(2. 대규모 투자/큰 성과)** 완성도 있는 서비스가 구현된다면 큰 성과를 얻을 수 있지만, 완성도 높은 서비스 구현을 위해 관련 기술, 데이터와 인프라가 뒷받침되어야 하는 영역임. 현재 국회전자도서관에서 도입되지 않은 서비스 모델은 E, F, G, H가 해당됨
- **(3. 대규모 투자/적은 성과)** 서비스의 목표와 적절성부터 기획해야 하는 서비스 또는 적은 노력으로 유지와 보완은 가능함. 기능적 필요성이 떨어지는 서비스가 포함된 영역임
- **(4. 소규모 투자/적은 성과)** 중요도와 활용성은 비교적 낮지만 구현도와 완성도가 높은 서비스가 포함된 영역임. 국회전자도서관에서 부분적으로 제공하고 있는 서비스 중 확대 적용할 수 있는 모델은 A, B가 해당됨

○ 서비스 모델의 도입 가능성과 관련 기술의 시장 상황을 고려하여 단기, 중기, 장기 과제로 구분

- **(단기 과제)** 6개월 이내로 도입할 수 있는 서비스 모델. 국회전자도서관의 일부 서비스에서 제공되어 확대 적용이 가능한 서비스 모델 A, B가 단기 과제에 해당됨

- **(중기 과제)** 서비스 도입까지 1년에서 3년의 기간이 필요한 서비스 모델. 서비스 모델과 관련된 라이브러리 또는 API가 있어 비교적 구현이 어렵지 않지만, 현재 국회전자도서관에서 서비스되지 않는 모델 C, D, H가 중기 과제임
- **(장기 과제)** 서비스 도입까지 3년에서 5년의 기간이 필요한 서비스 모델. 단기 과제나 중장기 과제가 선행되어야 도입할 수 있는 서비스 모델 또는 관련 라이브러리 또는 API를 직접 개발해야 하는 서비스 모델 E, F, G가 장기 과제로 분류할 수 있음

[표 6-4] 단기·중기·장기 과제의 개발 소요 시간과 서비스 모델

분류	단기	중기	장기
개발 소요 시간	6개월 이내	1년~3년	3년~5년
서비스 모델	A. 검색어 자동완성 B. 연관 검색어 제공과 분석	C. 메타데이터 자동 생성 D. 초록 자동 생성과 요약 H. 의안 자동 분류	E. 기계판독형 메타데이터 표준화 F. 한국어 법률 언어모델 G. 의안 정보 통합 지식그래프

3. 단기 과제

A. 검색어 자동완성

○ 이용자가 검색창에서 특정 키워드를 입력할 때 자동으로 예상 검색어를 추천하여 빠른 검색을 지원하는 서비스

- **(세부 내용)** 여러 분야의 데이터를 통합적으로 검색하는 서비스인 경우, 세부적으로 각 분야에 가장 적합한 단어들을 추천해주는 기능
- **(도입 효과)** 총 3개의 웹 서비스*에서 검색어 자동완성 기능을 제공하지만, 국회법률도서관, 국회의원 정책자료, 국회기록보존소, 전문가정보 데이터베이스는 자동완성 기능을 제공하지 않음. 특히, 국회법률도서관의 법률정보 통합 검색 서비스에 적용이 된다면 법률용어가 생소하거나 어려운 이용자는 법률용어의 검색어 입력이 수월해지고 검색어와 연관된 맥락 정보의 습득이 가능함

* 국회회의록 빅데이터, 국회전자도서관, 국회·지방의회 의정정보시스템

○ 현재 국회전자도서관이 부분적으로 제공하는 서비스이므로 지원되지 않는 웹 서비스에 확대 적용할 수 있음

- **(관련 기술)** 네이버 클라우드 검색 서비스는 한국어에 특화된 형태소 분석 처리기를 이용하

여 자동완성 검색을 설정하거나 수정할 수 있는 기능을 제공함. 아마존의 CloudSearch는 구조화된 데이터와 일반 텍스트를 인덱싱하고 검색하는 기능과 자동완성 기능을 제공함

B. 연관 검색어 제공과 분석

○ 검색 결과에 비슷한 주제의 단어인 연관 검색어를 제공하고, 단어 사이의 연관성을 분석하여 시각화하는 서비스

- **(세부 내용)** 검색 결과에 다른 검색어로 활용할 수 있는 단어들을 함께 제공할 수 있고, 별도의 메뉴를 제공하여 연관 검색어들을 통합한 내용을 분석하여 시각화하는 방법
- **(도입 효과)** 현재 국회·지방의회 의정정보시스템은 연관 검색어 기능을 제공하고 있으며, 국회전자도서관은 인포그래픽 메뉴를 통해 연관어를 분석한 내용을 트렌드, 분포, 인물과 인기도 기준으로 상세히 제공하고 있음. 이용자가 검색한 키워드와 의미적 연관성을 갖는 키워드를 추천하여 탐색을 확장할 수 있고, 이용자에게 검색어와 관련된 현황을 제공함

○ 국회전자도서관, 국회·지방의회 의정정보시스템에서 제공하는 연관 검색어 서비스를 확대 적용하여 많은 웹 서비스에서 지원 가능함

- **(관련 기술)** Amazon Comprehend는 머신러닝을 사용하여 텍스트에서 인사이트를 얻는 자연어 처리 서비스로 기능별 API를 통해 다방면 분석이 가능하며, Naver의 Data Analytics Service는 검색어 기반의 데이터 분석 환경을 제공함
- **(해외 사례)** 네덜란드 국립도서관연구소의 DBNL Ngram Viewer는 ngram 알고리즘을 사용해 기간 내 특정 키워드가 언급된 빈도를 시각화하고, Frame generator는 토픽 모델링 기반으로 연관 검색어와 관련된 분석 서비스를 제공함

4. 중기 과제

C. 메타데이터 자동 생성

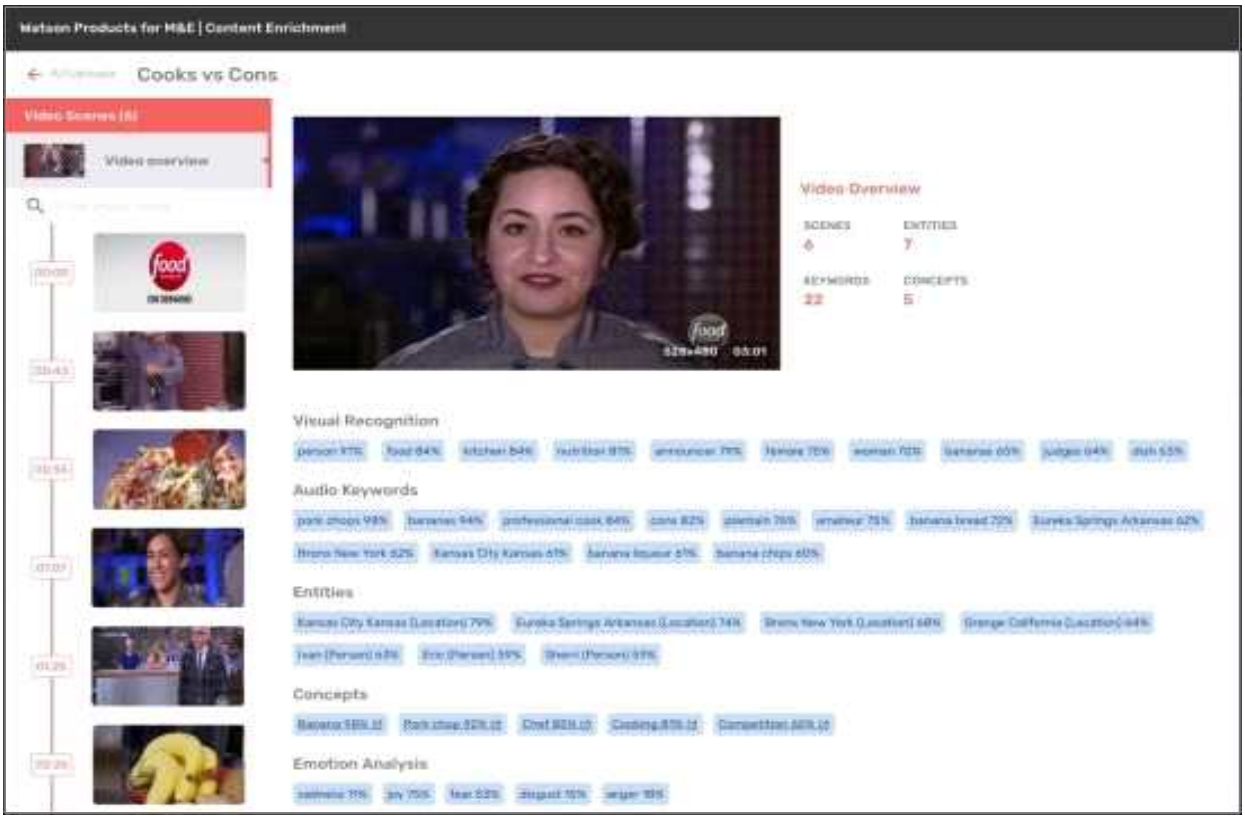
○ 기계가 자동으로 디지털 자원의 메타데이터(예: 출처 정보, 저자)를 생성하는 기능

- **(세부 내용)** 메타데이터는 데이터에 대한 데이터로, 데이터를 효율적으로 탐색할 수 있는 정보를 제공함. 그러나 대규모의 디지털 자원은 사람이 모든 자료의 메타데이터를 생성하는데 어려움이 있음. 효율적인 메타데이터 생성을 위해 인공지능 기반의 메타데이터 생성 도구가 필요함
- **(도입 효과)** 대부분의 국회전자도서관 서비스는 디지털 자원의 메타데이터가 충분하게 제공되지 않음. 특히, 국회·지방의회 의정정보시스템의 회의록 정보, 국회의원 정책자료의 정책자료는 충분한 메타데이터가 제공되지 않아 자료의 검색과 내용 이해에 제한이 있음. 풍부한 메타데이터는 디지털 자원을 이해하는데 필요한 맥락 정보를 제공하고, 디지털 자원의

효율적인 검색을 도움

○ 국회전자도서관이 보유한 모든 유형의 데이터 특성을 파악하여 적합한 메타데이터 항목을 설계해야 하므로 완성도 있는 서비스를 구현하기 위한 중기 계획이 필요함

- **(관련 기술)** IBM 왓슨은 동영상의 메타데이터를 자동 추출하는 클라우드 서비스로, 동영상을 심층 분석하여 핵심어, 개념, 비주얼 이미지, 어조, 정서적 맥락과 같은 메타데이터를 추출하는 기능을 제공함. Content Analysis on AWS 솔루션은 미디어 콘텐츠를 분석하기 위한 포괄적인 기능 세트를 제공하며, 메타데이터를 자동으로 추출하는 기능을 제공함
- **(도서관 사례)** Demosaurus(네덜란드 국립도서관연구소)는 인공지능을 사용해 메타데이터를 자동으로 생성하는 파이썬 라이브러리임. Demosaurus는 디지털 자원의 저자, 출판 시기, 출판사, 형식 등의 메타데이터를 자동 생성할 수 있음



<그림 6-2> IBM 왓슨의 동영상 메타데이터 자동 생성 기능 예시

D. 초록 자동 생성과 요약

○ 주어진 텍스트 내용을 분석하여 자동으로 핵심 요약 글을 생성하고 제공하는 기능

- **(세부 내용)** 원문을 요약하는 방법은 핵심적인 문장이나 단어를 추출하는 추출요약과 원문에 없는 문장이라도 문맥을 반영하여 새로운 문장을 생성하는 생성요약으로 구분됨. 추출요약 방식은 비교적 구현이 수월하지만, 정확한 요약을 위해 생성요약 방식을 고려해야 함

- **(도입 효과)** Argos와 국회전자도서관의 지능형 질의응답 서비스는 검색 결과로 뉴스나 소장자료와 연결된 링크와 원문의 부분적 내용만 제공함. Argos와 국회전자도서관에서 원문의 일부가 아닌 요약된 글 자체를 제공한다면, 외부 웹 사이트의 이동 없이 국회전자도서관 서비스 내에서 정보의 지속적인 탐색이 가능함



<그림 6-3> Argos 서비스: 초록 자동 추출을 제공하지 않음

○ 한국어 자동 요약과 관련된 라이브러리나 API를 사용할 수 있으나, 서비스 도입을 위한 범위를 선정하고 데이터를 확보하기 위한 중기 계획이 요구됨

- **(관련 기술)** 줌인터넷과 연합뉴스가 함께 개발한 '뉴스 세 줄 요약'이나 네이버 뉴스 요약봇의 CLOVA Summary와 같이 의안이나 회의록의 정보를 검색어 입력 시 요약하여 결과로 제공하는 서비스를 도입할 수 있음. Microsoft Azure Cognitive Service LUIS나 AYLIEN Text API, MeaningCloud와 같은 API를 사용하여 의안이나 회의록의 초록을 자동적으로 생성하는 기능을 개발할 수 있음

H. 의안 자동 분류

○ 의안 내용의 유사도를 기반으로 의안을 자동으로 분류하고, 특정 의안과 주제가 비슷한 의안을 추천하는 기능

- **(세부 내용)** 의안은 다양한 주제의 내용을 포함한 텍스트 데이터로, 단일 레이블 다중 분류(single-lable, multiclass classification)로 의안을 분류하고 특정 의안을 확인할 때 내용이 유사한 의안을 추천해주는 기능을 제공함
- **(도입 효과)** 입법자가 의안을 작성하기 전 발의예정인 의안과 유사한 의안을 검색할 수 있고, 중복적인 의안 발의를 방지할 수 있음. 일반 시민은 관심 있는 주제의 의안을 확인할 때, 주제 분류와 같은 의안을 함께 탐색할 수 있음

○ 의안 텍스트 전처리와 의안의 분류체계 설계에 많은 시간이 소요될 것으로 예상되고, 의안에 특화된 자동 분류 학습을 위해 중기 계획을 수립하는 것이 필요함

- **(관련 기술)** 딥러닝(예: CNN, RNN, LSTM, BERT) 기반의 분류 알고리즘을 사용해 문서의

유사도를 측정할 수 있음. API 서비스는 사용자가 직접 분류 모델을 생성할 수 있는 IBM 왓슨과 사전에 훈련된 모델을 제공하는 Google Cloud NLP가 있음

- **(도서관 사례)** 일본 국립국회도서관의 NDC Predictor와 노르웨이 국립도서관의 듀이 십진 분류법 자동 생성 사례는 딥러닝 기반으로 서지의 자동 분류 서비스를 제공함

5. 장기 과제

E. 기계판독형 메타데이터 표준화

○ 국회전자도서관이 보유한 모든 자원의 메타데이터 항목과 그 값의 형식을 통일하여 기계가 읽고 이해할 수 있는 형식으로 표현

- **(세부 내용)** 국회도서관이 보유한 디지털 자원에 표준 메타데이터 체계를 사용하고, 기계가 자동으로 메타데이터를 읽고 처리하는 RDF 방식의 메타데이터 체계를 검토해야 함
- **(도입 효과)** 대표적으로 국회의원 정보는 4개의 국회전자도서관 웹 서비스에서 동일한 정보를 중복적으로 제공하고 있고, 서로 다른 메타데이터로 기술하고 있음. 기계가 읽을 수 있도록 표준화된 메타데이터는 모든 디지털 자원의 상호운용과 재사용이 가능하여 통합 검색의 성능을 향상시킬 수 있음

※ <그림 6-4>는 국회의원 정책자료, 지능형 분석 Argos, 전문가정보 데이터베이스, 국회회의록 빅데이터에서 국회의원의 정보가 중복적으로 제공되고, 재사용되지 않은 형태로 기술됨



<그림 6-4> 강기윤 국회의원 정보의 서로 다른 메타데이터 기술

○ 기계판독형 메타데이터 구축은 기존의 메타데이터와 표준 어휘 사이의 매핑 체계를 설계하기 위한 광범위한 작업이 요구되므로 장기적인 전략 수립이 필요함

- **(관련 기술)** 표준 RDF 어휘는 디지털 자원의 표현을 위한 더블린 코어와 Schema.org, 기록 자원의 표현을 위한 RiC-O, 서지 정보의 표현을 위한 Bibframe 등이 있음
- **(도서관 사례)** 미국의회도서관의 웹 아카이브는 개별 항목에 대한 상세한 정보를 제공하고, 메타데이터는 더블린 코어 중심의 표준 어휘로 표현되어 기계가 읽고 처리할 수 있음

F. 한국어 법률 언어모델

○ 언어모델은 문장 또는 단어에 확률을 할당하여 기계가 처리할 수 있도록 하는 모델로, 법률에 특화된 언어모델을 구축해야 함

- **(세부 내용)** 법률 분야를 다루는 국회전자도서관의 전문성을 증진시키고자 국회전자도서관이 보유하는 법률 데이터를 학습시키고, 법률 분야에 특화된 언어모델을 생성해야 함
- **(도입 효과)** 법률 분야에 특화된 한국어 자연처리 모델을 적극 활용하여 Argos 외의 시스템에 전체적으로 적용하고, 이를 기반으로 텍스트 생성 최신 기술을 도입하여 이용자의 편의성과 시스템 성능을 향상시킬 수 있음

○ 언어모델 생성을 위해 전문 인력의 도입이 필요하고, 법률 언어모델을 향상시키고 이를 국회도서관의 다양한 웹 서비스에 도입하기 위한 장기적인 도입 전략이 필요함

- **(관련 기술)** 최근 문장의 의미 추출에 강점을 가진 BERT 계열 모델(RoBERTa, SpanBERT 등)과 이전 단어가 주어졌을 때 다음 단어를 맞추는 학습 모델인 GPT 계열(GPT-2, GPT-3)의 성능이 우수함.
- 최근 한국어의 언어학적 특징을 반영한 koBERT와 korBERT, KoGPT-2와 GPT-3 모델 기반의 KoGPT, HyperCLOVA 등이 개발됨. 법률 분야의 한국어 언어모델은 LBox의 Legal-BERT가 대표적임
- **(도서관 사례)** 스웨덴 국립도서관은 스웨덴어에 최적화된 언어모델인 Swedish BERT와 spaCy를 직접 구축하여 활용하고 있음

G. 의안정보 통합 지식그래프

○ 다양한 의안 정보를 의미적으로 연결하여 관계를 생성하고, 인공지능이 처리 가능한 형태로 그래프 형태의 지식베이스 생성

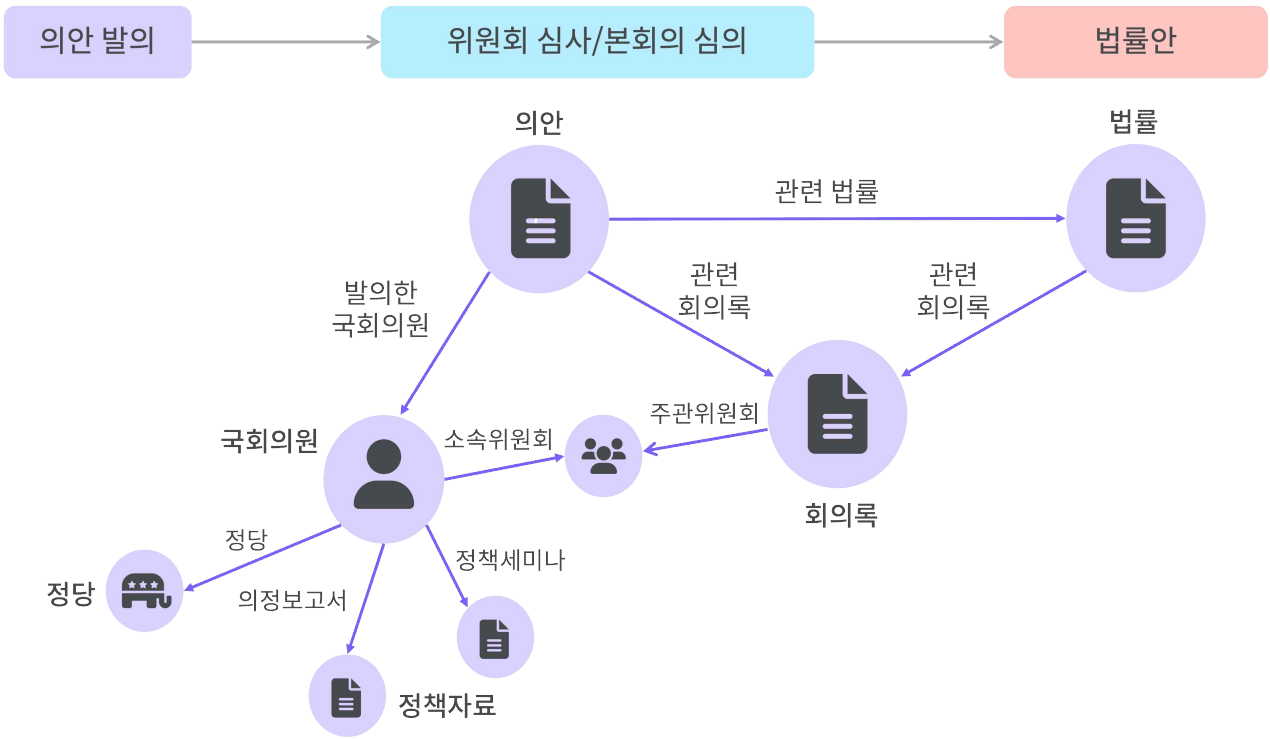
- **(세부 내용)** 현재 국회전자도서관이 소장하는 디지털 자원은 대부분 DB에 저장되어 데이터 사이의 연결에 제한이 있고, 의안 관련 정보는 국회전자도서관의 서로 다른 웹사이트에 존

재하여 종합적으로 의안 처리 과정을 파악하기 어려움

- 인공지능이 읽고 이해할 수 있는 표준 어휘를 사용하여 의안 관련 정보의 관계를 의미적으로 표현하고, 파편화되어 존재하는 서로 다른 의안 정보를 통합할 수 있음

※ <그림 6-5>는 국회전자도서관이 보유한 데이터 중심으로 의미적으로 연계된 의안 정보 지식그래프의 예시임

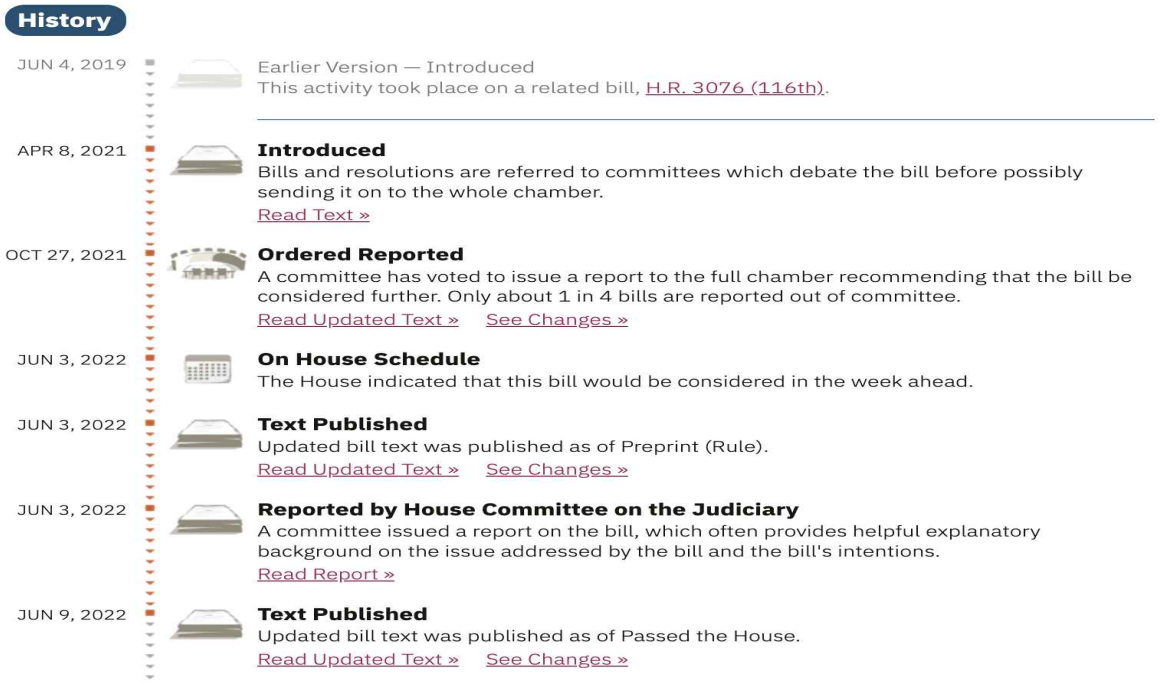
- **(도입 효과)** 의안 관련 데이터가 유기적으로 연결되어 의안의 발의부터 법률안이 만들어지기까지의 과정을 한 번에 제공할 수 있는 서비스를 제공할 수 있음
- 의안 관련 데이터가 연결될수록 국회회의록 빅데이터나 Argos에서 제공하는 데이터 분석이 더욱 풍부해질 수 있음



<그림 6-5> 의미적으로 연계된 의안 정보 지식그래프

○ 지식그래프 구축을 위해 디지털 자원의 메타데이터 표준화가 선행되어야 하고, 데이터 사이의 유기적 연결을 위해 데이터를 파악·정제하고 그 관계를 지식 모델로 표현하는 프로세스가 필요함

- **(대표 사례)** 미국의 GovTracks.us는 <그림 6-6>과 같이 의안의 진행 과정을 추적한 정보를 제공하고 있음. 지식그래프 구축된 의안 정보는 의안 관련 정보가 연계되어 의안의 처리 과정을 함께 제공할 수 있음



<그림 6-6> GovTracks.us의 H.R.2377 의안 처리 과정

6. 요약

○ 인공지능 기술 기반의 검색 서비스는 대용량 데이터 도입과 활용, 머신러닝 기반의 인공지능 기술과 지식 그래프 기술을 융합하여 성능 고도화에 집중할 필요가 있음

○ **(공통기술)** 검색 서비스를 위한 공통 기술은 데이터 수집·전처리·정제에 적용할 수 있고, 대용량 빅데이터의 분산 환경에서 안정적인 수집, 실시간 데이터 처리와 정제, 효율적인 저장과 색인에 활용할 수 있음

- 인공지능 모델은 사전학습 기법을 적용하고 있고 이를 위해 대용량 코퍼스가 필요함. 한국어에 특화된 대용량 코퍼스를 활용하여 모델의 학습, 검증, 테스트를 진행하면 검색 성능의 향상과 AI 기반 서비스를 제공하기 위한 기반을 확보할 수 있음
- 한국어는 영어와 구조와 표현방식이 다르기 때문에 검색에서 활용하는 자연어를 처리하기 위해 신조어, 다양한 어형 변화를 지원하기 위한 한국어 형태소 분석 기술의 확보와 최적화가 필요함

○ **(단기 서비스 기술)** 최근 이미지, 음성과 같은 서로 다른 유형의 자원 사이의 검색을 지원하기 위해 메타검색, 시맨틱 검색을 적용하기 위한 기술을 검토해야 함

- 사용자 질문에 대한 후보 검색 정보들을 군집화하는 것을 넘어 임베딩 기법(유사 자원을 인접한 벡터 공간에 사상)을 적용하면 풍부한 검색 결과를 제공할 수 있음

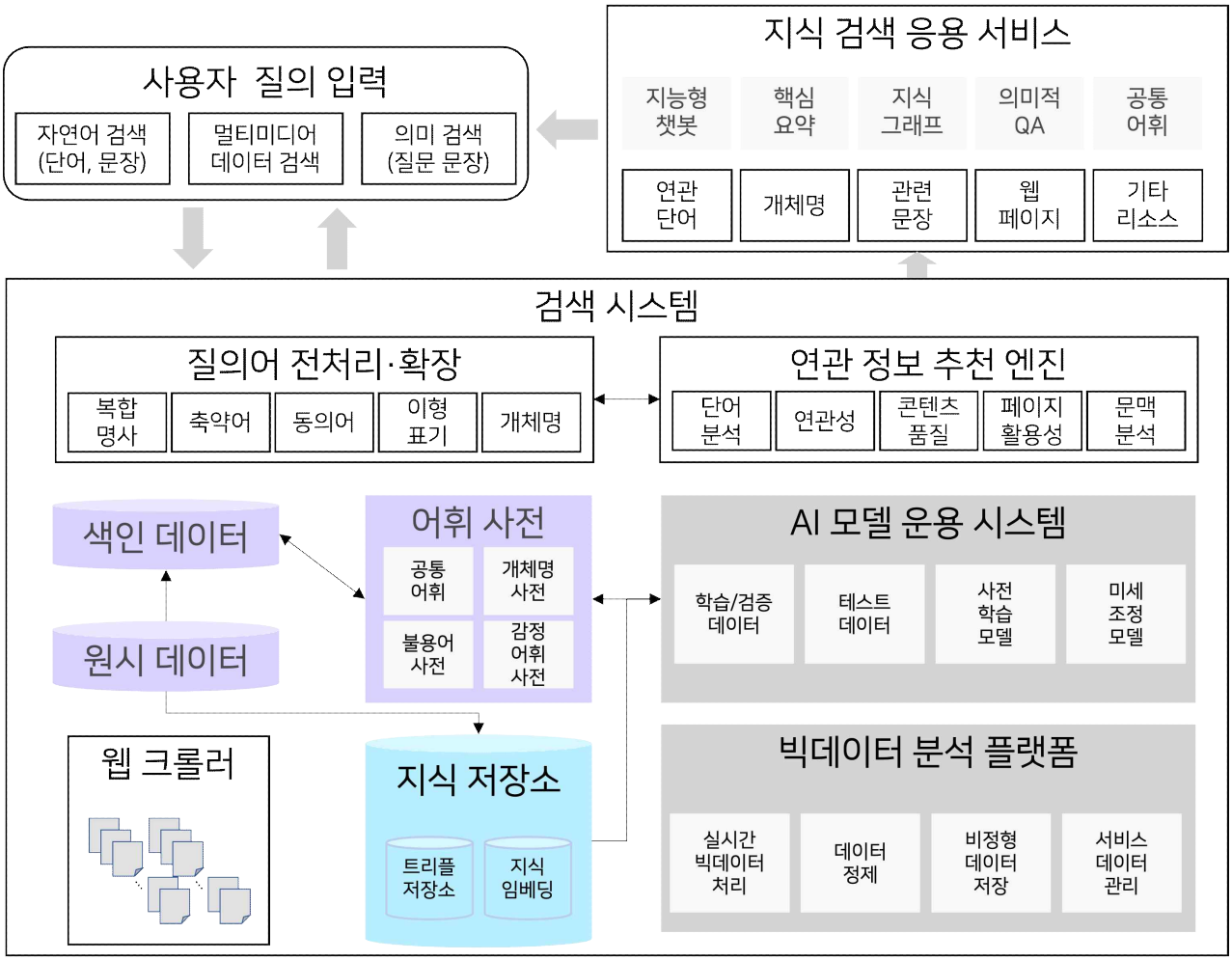
- 자원 사이의 관계 정보는 지식그래프 기술을 통해 모델링할 수 있고, 관계 정보를 유연하게 처리하기 위해 지식베이스가 필요함. 특히 다양한 자원의 메타데이터 정보 제공, 자원 사이의 관계 식별, 암묵적인 관계를 추론할 수 있는 지식그래프 기술(지식그래프 저장소, 지식그래프 모델링, 지식그래프 URI 설계 등)이 필요함

○ **(중장기 서비스 기술)** 지식그래프와 함께 기존의 텍스트 기반 AI 기술을 융합하기 위한 멀티 모달 퓨전 언어모델의 도입을 검토해야 함

- 기존의 텍스트에서 문맥 정보의 표현이 충분하지 못한 한계를 지식그래프의 임베딩(GNN, KG Embedding)과 지식 퓨전 AI 기술을 적용하여 해결할 수 있고, 기계 독해, 텍스트 생성 요약, 관계 예측 등의 서비스에 적용할 수 있음

○ **(지능형 검색 시스템)** 중장기적 관점에서 지식검색을 지원하기 위한 목표 시스템을 설계하고, 인공지능 기술을 포괄적으로 적용할 수 있는 추진 전략을 수립해야 함

- 국회도서관에서 제공하는 인공지능 서비스는 공통 기술과 데이터를 활용하여 확장 가능한 방식으로 추진되는 것이 바람직함



<그림 6-7> 인공지능 기술 기반 지능형 검색 시스템을 위한 기능 중심 개요도

[표 6-5] 지능형 검색 서비스 제공을 위한 주요 기술

서비스	주요 기술	
	공통 기술	AI 기술
A. 검색어 자동완성	- **분산 병렬 데이터 수집 도구** · 웹 자원에 포함된 텍스트, 이미지, 동영상 등의 대용량 자원을 신속하게 수집하고 저장할 수 있는 분산형 웹 크롤러 · 감정 어휘 및 개체명을 선택적으로 수집할 수 있는 인식, 색인, 수집 기술 - **대용량 AI 학습 데이터** · 서비스 목적에 맞는 AI 모델 학습을 위한 전문용어사전, 다양한 분야의 개체명 사전, 축약어/신조어 사전 · 학습/검증/테스트용 레이블 학습 데이터 - **데이터/메타데이터 표준화** · 자원에 대한 메타데이터 정의를 위한 표준 가이드와 관리 도구 - **질의어 확장 기술** · 다양한 사용자 질의 응답 데이터 (AI 모델 학습용) · 검색어에 대한 확장을 위한 키워드, 문장, 문서, 트리플(지식그래프) 사이의 유사도 계산 기술 - **검색 성능 향상을 위한 실시간 빅데이터 플랫폼** · 검색 결과의 신속한 제공을 위한 고성능 색인 기법 · 실시간 데이터 처리 및 서비스 제공을 위한 API	- 신조어, 특수어 등에서 사용되는 자소, 자간에 대한 자동 분류 및 예측을 위한 딥러닝 기반 형태소 분석기 (신규 개발 필요)
B. 연관 검색어 제공과 분석		- 연관된 키워드, 문장, 문서, 리소스의 유사도를 계산하기 위한 임베딩 기법 - 연관성에 대한 관계 도출을 위한 관계 분류 및 관계 예측 모델 (트랜스포머 기반의 한국어 언어모델, KG 임베딩, GNN 기술)
C. 메타데이터 자동 생성		- 데이터 및 웹 자원이 포함하고 있는 문맥(상황)에 대한 이해를 위한 자연어 이해(NLU)를 위한 AI 기술 (WSD 기술, SACE, DeBERTa-v3 등) - 레이블 데이터 자동 확장을 위한 AI 기술 (AutoML, Active Learning, Reinforce-learning 등)
D. 초록 자동 생성과 요약		- 핵심 개체가 포함된 초록(짧은 문장) 요약 및 생성을 위한 언어 모델 (BARTSum, BertSum 등)
E. 의안 자동 분류		- 문서를 구성하는 키워드 사이의 상호 관계를 파악하여 문서의 대표성을 도출할 수 있는 트랜스포머 기반의 AI 모델 (BERT, RoBERTa, Span-BERT 등)
F. 기계판형 메타데이터 표준화		- RDF 형태의 데이터 자동 생성 시 표준화된 메타데이터 추천을 위한 AI 모델 (GNN, KG-embedding 등)
G. 한국어 법률 언어모델		- 한국어의 교착어 특성을 반영하고 장기 문맥 정보를 반영할 수 있는 AI 법률 모델 (KoBARTSum, KoBertSum 등)
H. 의안정보 통합 지식그래프		- 지식그래프에 포함된 자원 사이의 관계 정보를 통해서 사전학습 모델의 문제를 해결하고, 이미지, 동영상의 메타데이터와 텍스트 데이터를 융합할 수 있는 멀티 모달형 AI 언어 모델 필요 (ERNIE, CoLake 등)

제 7 장 결론

1. 결과 요약

□ 인공지능 기술 동향

○ 4차 산업 혁명 시대에 도서관이 도입할 수 있는 인공지능 기술을 검토하고 디지털 기술 분야의 전체적인 동향을 조사·분석함

- 미국·중국·일본 등 주요 국가는 국가 경쟁력 강화를 위해 디지털 기술을 활용한 국가 정책과 디지털 트랜스포메이션과 관련된 정책을 본격적으로 추진하고 있음
- 주요국들이 추진한 13개의 디지털 트랜스포메이션 관련 정책을 분석한 결과, 11개의 정책에서 인공지능 기술을 주요하게 언급하고 있음

○ 인공지능 도입 방안을 구체적으로 계획하기 위해 인공지능 기술에 대한 분석 수행

- 3개의 인공지능 분야*를 역할에 따라 구분하고, 국제동향과 핵심 기술 동향에 대해 심층적으로 조사를 수행함

* 인공지능의 하위 분야 3개: 머신러닝, 딥러닝, 자연어 처리

○ 인공지능 분야에서 자연어 처리와 머신러닝 기술을 결합하여 국회전자도서관에 도입하는 것이 적절하다고 판단됨

- 자연어 처리 기술은 도서관에서 범용적으로 활용될 수 있고, 머신러닝 기술은 인공지능에서 지능형 서비스를 구현하는 데 필수적으로 고려할 필요 있음

□ 자연어 처리 기술 동향

○ 심층 머신러닝 기술과 자연어 처리 기술을 연계하여 적용 방안을 검토

- 자연어 처리 관련 최신 기술들에 대한 개념적 내용과 각 기술의 특징, 성과를 세부적으로 정리하여 현재 국내외 자연어 처리 기술 발전 동향 파악
- 4개의 다국어 자연어 처리 언어모델과 8갱의 한국어 처리를 위한 언어모델, 7개의 자연어 처리 관련 알고리즘에 대해 조사를 수행함

○ 자연어 처리 기술 활용 사례를 파악하기 위해 6개의 응용서비스* 조사

- 사례별로 사용된 언어 기술 모델, 현재 상업적으로 서비스하고 있는 예시에 대해 정리하여 국회전자도서관에 도입 가능 여부를 검토함

* 응용서비스 사례: 질의응답/검색, 문서분류와 요약, 이미지 인식, 음성 인식, 기계번역, 지식그래프

□ 국외 주요도서관 사례조사

○ 머신러닝, 인공지능 관련 연구소를 활용해 기술개발을 진행하거나 관련 로드맵, 또는 비전을 발표한 국외도서관*을 대상으로 진행

* 조사 대상 국외도서관: 미국, 일본, 프랑스, 노르웨이, 스웨덴, 핀란드, 네덜란드

- 총 19개의 인공지능 관련 기술의 도입 사례를 분석하고 서비스 사례를 검토함
- 미국, 일본, 스웨덴, 네덜란드의 도서관은 자체적으로 연구소를 설립하여 도서관에 적용 가능한 머신러닝 기술에 대한 연구를 수행하고 있음
- 프랑스, 노르웨이, 핀란드의 도서관은 소장하고 있는 디지털 자료의 활용과 접근성을 높이기 위한 목적으로 인공지능 기술의 활용을 검토하고 있음

□ 국회전자도서관 서비스 현황 조사

○ 현재 국회도서관의 서비스와 개선점을 파악하기 위해 8개 웹사이트 서비스 조사

- 8개 사이트에서 14개의 서비스가 존재하고, 대부분의 사이트에서 검색 서비스를 기본적으로 제공하고 있음
- 국회도서관 자료를 통합적 제공하는 사이트는 총 4개 (국회전자도서관, 국회 지방의회 의정정보 시스템, Argos, 국회회의록 빅데이터)에서 제공하고 있음
- 각 웹사이트의 자료를 기반으로 키워드 기반 검색만 제공하는 사이트는 4개이며, 국회법률도서관, 국회의원 정책자료, 국회기록보존소, 전문가정보 데이터베이스에서 제공

□ 차세대 전자도서관 서비스 모델

○ 국외 주요도서관 서비스 벤치마킹

- 이미지, 오디오 등 멀티미디어 자료와 분산적으로 운영되는 개별 사이트의 자료를 연계한 통합적 검색을 지원할 필요 있음
- 이용자 편의성과 법률회답서비스의 효율성을 높이기 위해 원문 검색 결과의 요약 서비스와 다국어 기계번역 서비스 도입을 검토할 필요 있음

○ 국회전자도서관에 적용할 수 있는 서비스 모델 제안과 분석 프레임워크의 적용

- 국외 주요도서관 서비스 사례와 국회전자도서관의 제공 서비스를 기반으로 신규 서비스 모델을 제안함
- 성과와 도입 난이도에 따른 페이오프 매트릭스를 적용하여 서비스 모델의 개발 소요 시간과 효과 판단
- 서비스 모델의 도입효과, 기술의 성숙도를 기준으로 단기·중기·장기 과제를 각각 2~3개로 구분하여 제안함

2. 시사점

○ 인공지능과 자연어 처리의 기술 동향에 따라 지속적인 기술 분석이 필요함

- 인공지능과 자연어 처리 관련 기술은 매우 빠르게 발전하고 있는 분야이고, 응용 가능한 분야도 광범위하게 확대될 것으로 예상됨. 따라서, 인공지능과 자연어 처리 기술의 동향과 국외도서관 적용 사례는 향후 지속적으로 검토하고 업데이트해야 함

○ 국외 주요도서관은 최신 인공지능 기술을 이용하여 맞춤형 서비스를 제공하고 있음

- 도서관 이용자들을 대상으로 유용한 신규 서비스를 제공하고 있으며, 분류기 구축처럼 사서들이 도서관 내부적으로 유용하게 사용할 수 있는 기반을 마련하고 있음
- 국외 주요도서관은 다량의 텍스트 데이터를 갖고 있고, 모국어 말뭉치와 언어처리 모델의 구축을 서비스하고 있음. 도서관 이용자를 넘어 국민 전체에게 서비스를 제공하고, 더 나아가 인공지능 기술 자체의 발전에 기여하고 있음
- 기술이 발전함에 따라 도서관이 제공할 수 있는 서비스의 종류도 더욱 다양해질 것이라고 예상됨

○ 대국민 지식정보 지원기관이자 의정활동 지원기관인 국회도서관의 특성을 고려한 서비스 개발이 필요함

- 국회도서관은 일반도서만 다루는 것이 아니라, 국회 입법 관련 자료도 다루고 있음. 국회도서관만이 보유하고 있는 장서의 특성을 잘 활용한 서비스 모델을 구현한다면, 국회도서관의 위상을 높이고 이용률이 증가할 것으로 예상됨
- 국회도서관은 입법 관련 직무에도 직접적인 영향력을 끼칠 수 있는 도서관이며, 국가차원에서 국민들이 주목하고 있는 도서관임. 국회도서관이 앞장서서 인공지능과 자연어 처리 기술을 활용하여 우수한 서비스 모델을 발굴한다면, 국내 여러 도서관들에게 모범 사례가 될 수 있을 것이라고 예상됨

○ 4차 산업혁명시대에 사람들이 유용하게 이용할 수 있는 서비스 종류와 범위를 고려하여 로드맵을 설계하는 것이 우선적으로 필요함

- 구체적인 서비스를 구현하기 전에 도서관의 관점에서 이용자와 관계자, 국민들의 수요를 우선적으로 파악하여 단계적으로 실현가능한 계획의 마련이 필요함

○ 단편적인 기술 도입보다 도서관 내부에서 체계적으로 인공지능 서비스 도입이 가능한 기반을 구축하는 것이 중요함

- 단발성 프로젝트로 기술을 도입하지 않고, 도서관의 정보화 정책에 따라 관련 기술과 사례, 현황을 모니터링하여 도입할 수 있는 체계를 갖추어야 함
- 기존의 서비스에 새로운 기술을 도입하는 일은 시간이 많이 소요된다는 것을 염두에 두고,

장기적으로 적절한 서비스를 설계하여 점진적으로 실행할 수 있는 기반을 마련해야 함

○ 원활한 인공지능과 자연어 처리 서비스 도입을 위해 고품질의 데이터를 공유하고, 재사용할 수 있는 데이터 관리(data stewardship) 체계를 마련해야 함

- 본 연구에서 제안한 개선 사항 중에서 메타데이터 표준화, 한국어 법률 언어모델, 지식그래프는 데이터 거버넌스 수립을 통해 추진하는 것이 바람직하며, 국회전자도서관이 보유하고 있는 자료와 데이터가 중요한 기반이 된다는 공통점을 갖고 있음
- 국회전자도서관이 보유하고 있는 모든 유형의 데이터를 고품질로 유지하는 것은 인공지능, 자연어 처리 도입의 필수적인 전제조건이므로 국회도서관 내부에서 자료와 데이터를 저장하고 관리할 수 있는 효과적인 체계를 마련하는 과정이 필요함
- 체계적인 데이터 구축을 위해 FAIR 데이터 원칙(FAIR data principles)*을 적용하고, 상호운용과 재사용이 용이한 데이터 체계를 구축해야 함

* FAIR 데이터 원칙: 데이터의 공유와 재사용성을 강화하기 위해 고안된 가이드 원칙(Wilkinson et al., 2016)으로, 데이터를 보존하고 관리하기 위한 보편적인 프레임워크 역할을 하고 있음

붙임 1. 응용서비스별 오픈소스 라이브러리

응용서비스	레포지토리	인용 논문 수	깃허브 Star 수	URL
정보검색	jingtaozhan/DRhard	3	87	https://github.com/jingtaozhan/DRhard
	shenweichen/DeepCTR	2	6,036	https://github.com/shenweichen/DeepCTR
	stanford-futuredata/ColBERT	2	470	https://github.com/stanford-futuredata/ColBERT
	xue-pai/FuxiCTR	2	194	https://github.com/xue-pai/FuxiCTR
	WING-NUS/scisumm-corpus	2	191	https://github.com/WING-NUS/scisumm-corpus
	jingtaozhan/LearningToRetrieve	2	87	https://github.com/jingtaozhan/LearningToRetrieve
	jingtaozhan/repconc	2	53	https://github.com/jingtaozhan/repconc
	jingtaozhan/JPQ	2	32	https://github.com/jingtaozhan/JPQ
질의응답(QA)	huggingface/transformers	23	62,469	https://github.com/huggingface/transformers
	facebookresearch/ParlAI	5	8,813	https://github.com/facebookresearch/ParlAI
	PaddlePaddle/ERNIE	4	4,866	https://github.com/PaddlePaddle/ERNIE
	dmlc/gluon-nlp	4	2,412	https://github.com/dmlc/gluon-nlp
	google-research/tapas	4	817	https://github.com/google-research/tapas
	ZHEvent/ZHEvent.github.io	4	16	https://github.com/ZHEvent/ZHEvent.github.io
문서 분류	huggingface/transformers	6	62,469	https://github.com/huggingface/transformers
	dmlc/gluon-nlp	2	2,412	https://github.com/dmlc/gluon-nlp
	QData/TextAttack	2	1,954	https://github.com/QData/TextAttack
	wikipedia2vec/wikipedia2vec	2	774	https://github.com/wikipedia2vec/wikipedia2vec
	facebookresearch/xformers	2	712	https://github.com/facebookresearch/xformers
	alibaba/EasyNLP	2	462	https://github.com/alibaba/EasyNLP
	cair/TsetlinMachine	2	394	https://github.com/cair/TsetlinMachine
	amzn/pecos	2	295	https://github.com/amzn/pecos

문서 요약	huggingface/transformers	6	62,469	https://github.com/huggingface/transformers
	theamrzaki/text_summurization_abstractive_methods	5	441	https://github.com/theamrzaki/text_summurization_abstractive_methods
	HHousen/TransformerSum	3	280	https://github.com/HHousen/TransformerSum
	tensorflow/tensor2tensor	2	12,268	https://github.com/tensorflow/tensor2tensor
	facebookresearch/xformers	2	712	https://github.com/facebookresearch/xformers
	huffon/factsumm	2	74	https://github.com/huffon/factsumm
	google-research/longt5	2	60	https://github.com/google-research/longt5
음성 인식	pytorch/fairseq	10	16906	https://github.com/pytorch/fairseq
	PaddlePaddle/PaddleSpeech	9	3032	https://github.com/PaddlePaddle/PaddleSpeech
	mravanelli/pytorch-kaldi	8	2167	https://github.com/mravanelli/pytorch-kaldi
	espnet/espnet	7	5031	https://github.com/espnet/espnet
	TensorSpeech/TensorFlowASR	6	674	https://github.com/TensorSpeech/TensorFlowASR
	Alexander-H-Liu/End-to-end-ASR-Pytorch	5	999	https://github.com/Alexander-H-Liu/End-to-end-ASR-Pytorch
	thu-spmi/CAT	5	222	https://github.com/thu-spmi/cat
	huggingface/transformers	4	62469	https://github.com/huggingface/transformers
	PaddlePaddle/DeepSpeech	4	3033	https://github.com/PaddlePaddle/DeepSpeech
	rwth-i6/returnn	4	324	https://github.com/rwth-i6/returnn
	NVIDIA/NeMo	3	4186	https://github.com/NVIDIA/NeMo
	osmr/imgclsmob	2	2554	https://github.com/osmr/imgclsmob
이미지 인식	PaddlePaddle/PaddleOCR	17	21,364	https://github.com/PaddlePaddle/PaddleOCR
	open-mmlab/mmocr	5	2,425	https://github.com/open-mmlab/mmocr
	Media-Smart/vedastr	5	475	https://github.com/Media-Smart/vedastr
	mindee/doctr	3	1,010	https://github.com/mindee/doctr
	huggingface/transformers	2	62,588	https://github.com/huggingface/transformers
	PaddlePaddle/PaddleSeg	2	4,750	https://github.com/PaddlePaddle/PaddleSeg

기계번역	awslabs/sockeye	21	1,066	https://github.com/awslabs/sockeye
	pytorch/fairseq	15	16,906	https://github.com/pytorch/fairseq
	huggingface/transformers	11	62,469	https://github.com/huggingface/transformers
	tensorflow/tensor2tensor	10	12,268	https://github.com/tensorflow/tensor2tensor
	bzhangGo/zero	10	110	https://github.com/bzhangGo/zero
	facebookresearch/fairseq	9	3,791	https://github.com/facebookresearch/fairseq
	marian-nmt/marian	5	894	https://github.com/marian-nmt/marian
	THUNLP-MT/THUMT	5	634	https://github.com/THUNLP-MT/THUMT
	lvapeab/nmt-keras	5	525	https://github.com/lvapeab/nmt-keras
	thumt/THUMT	4	634	https://github.com/thumt/THUMT
	bytedance/neurst	4	224	https://github.com/bytedance/neurst
지식그래프	Accenture/AmpliGraph	4	1737	https://github.com/Accenture/AmpliGraph
	Sujit-O/pykg2vec	2	448	https://github.com/Sujit-O/pykg2vec
	shaoxiongji/awesome-knowledge-graph	2	1218	https://github.com/shaoxiongji/awesome-knowledge-graph
	daiquocnguyen/ConvKB	2	167	https://github.com/daiquocnguyen/ConvKB

붙임 2. 국회도서관 제공 서비스 목록

서비스	구분	내용
의회정보서비스	의회정보회답	. 국회의원, 입법지원기관의 자료 요청에 대해 전문사서와 외국자료조사 전담직원이 국내외 정책과 현황을 조사, 검색하여 제공 . 영어, 일본어, 중국어, 프랑스어 언어권의 제도, 정책, 현한 조사 . 정책참고자료 번역요청은 외부 전문업체를 통해 번역 서비스 제공
	입법지식서비스	. 국회의원, 국회직원이 의회정보서비스를 편리하게 이용할 수 있는 통합 플랫폼 . 공공정책정보, 국회의원 정책자료 등 데이터베이스와 팩트북 등 발간자료, 외국정책사례 콘텐츠를 국회 내부에 제공
	공공정책정보DB	. 국내외 정부기관, 공공기관과 싱크댕크 등 인터넷 홈페이지의 최신 정책자료, 학술자료 등을 수집, 선정, 목록작성, DB 구축
	전문가정보 DB	. 국회휴먼네트워크 . 국회의원, 각분야 전문가를 검색하는 데이터베이스 . 2020년 홈페이지 개편을 통해 검색기능 강화, 전문가 참여 콘텐츠 추가
	국회의원 정책자료 DB	. 국회의원 주최 세미나, 공청회, 정책자료집, 의정보고서 등 입법과 정책활동 관련 자료를 국회 조서관 홈페이지를 통해 제공 . 국회의원별, 정당별, 상임위원회 주제별, 국회대수별 검색 가능
	국회(중점)국회의원 보도자료DB	. 입법활동과 정책 추진상황을 중심으로 국회의원, 입법지원기관, 정당 등의 보도자료 제공
	외국신문정보DB	. 외국 주요 언론매체에 게재된 외국 의회, 입법동향, 정치, 사회, 경제 관련 기사와 한국 관련 주요 내용을 요약, 번역하여 국회의원, 입법지원기관을 대상으로 주1회 메일링 서비스
	싱크탱크보고서DB	. 주요국 의회, 정부, 민간 부문의 싱크탱크에서 제공하는 최신 정책보고서를 번역, 요약하여 국회의원과 입법지원기관을 대상으로 주1회 메일링 서비스

법률정보서비스	**법률정보회답**	. 국회의원, 국회 내 입법지원기관이 요청하는 국내외 법률정보에 대한 회답 제공 . 영미법, 일본법, 독일법, 프랑스법 법률자료조사관과 담당사서가 주요국의 입법례, 판례 등 다양한 법률정보를 조사, 번역 . 그외 언어권, 단순번역의 경우 전문번역업체에 의뢰 . 외국법률번역DB 구축
	법률문헌주제분류	. 국내에서 생산, 유통되는 법률문헌을 주제별로 파악하기 위한 종합관리체계 구축과 검색 편의 제공
	국회법률정보서비스	. 국내외 법률정보를 전문적이고 종합적으로 제공하기 위해 국회법률정보서비스 홈페이지를 구축하여 운영 . 외국법률정보DB, 외국입법사례DB, 법률쟁점DB 등 각종 법률정도DB와 국회회의록과 법제처의 법령, 판례정보 등 외부기관의 법률관련 콘텐츠 연계하여 서비스
	외국법률번역DB	. 국회의원과 입법지원기관의 외국법률 번역 요청에 대해 자체번역하거나 외부기관이 번역한 외국법률을 수집하여 DB로 구축한 번역법률데이터베이스
	법률쟁점DB	. 국회입법과정에서 생산된 자료를 분석해 개별 법률별로 제, 개정 요약, 입법경과를 정리하고 관련 자료를 연계하려 제공하는 데이터베이스
정보봉사	**정보검색지원**	. 이용자의 요구와 질의에 대한 참고봉사업무 . 디지털정보센터 참고도서 코너의 참고자료 이용안내와 특정 주제와 현안에 대한 참고질의 . 전문사서의 정보검색 지원과 참고봉사서비스 . 도서관 홈페이지 이용자마당의 '묻고 답하기'
	장애인 및 노약자 서비스	. 장애인 이용자의 자료 열람을 위한 시설물 . 대면낭독, 원문음성지원서비스, 점자파일과 출력
데이터 융합 분석	**지능형 의회정보 융합분석**	. 뉴스, 소셜미디어 등 외부자료에 대한 빅데이터분석과 국회 내부 보고서를 연계하여

	시스템 Argos	의사결정과정 지원 . 2020년 머신러닝 분석엔진을 적용해 팩트북, 싱크탱크 등 소장자료 학습을 통해 이슈 분류 기능을 개선 . 국내외 헌법 및 국내법률에 대한 지능현 질의응답(QA) 서비스 . 시각화 서비스
	국가학술정보 클라우드 서비스	. 국가학술정보 데이터를 분야별 도서관과 연구기관 등과 공동으로 구축, 활용하여 융합, 분석 서비스 제공 . 학술정보 표준 온톨로지 구축, LOD 기반의 통합 검색 지원 . 국회, 지방의회 회의록 분석을 통한 비정형 데이터셋 구축 활용 기반 마련 . 국회 및 지방회의록 데이터를 발언자, 내용, 안건 단위로 검색하여 결과를 데이터셋으로 제공하는 서비스
전자도서관	**원문DB 구축**	. 국회입법활동 지원을 위한 학술지, 학위논문, 공공정책자료 등 원문 등록 서비스 지원
	저자전거DB 구축	. 저자 검색 시 정확한 식별을 위한 저자 전거DB 구축 . 인명과 단체명에 대한 필명(筆名), 호(號), 한자명(漢子名), 외 국어명, 약칭, 이전명칭 등 이형(異形)을 수집하고 관련 저작물 연계
	용어관계사전DB 구축	. 검색의 효율성과 정확도를 높이기 위해 시소러스 구축
	자료추천시스템	. 납본되지 않는 국내외 상업출판물에 대해 자료추천 위원들이 도서를 선별·추천하는 시스템으로 단행본 수서업무에 활용하는 시스템
	챗봇시스템	. 국회도서관 이용자의 다양한 문의사항에 대해 채팅로봇 이용하여 On-Demand 상담 서비스를 제공하는 시스템
국회기록물 서비스	**국회 주요기록물 열람 서비스**	. 국회에서 생산한 일반문서, 국회간행물, 국회회의록, 국회의안 열람 서비스 제공
	국회의장단과 주요 인사 구술채록 서비스	. 제6대부터 제21대까지의 국회 주요인사에 대한 구술채록 서비스 제공
	임시의정원 아카이브	. 국회기록보존소에서 수집한 임시의정원 관련 기록 제공하는 온라인 플랫폼

[참고문헌]

○ **논문**

곽우정, 노영희. (2021). 도서관의 인공지능(AI) 서비스 현황 및 서비스 제공 방안에 관한 연구. 한국도서관정보학회지. 52(1), 155-178. https://doi.org/10.16981/KLISS.52.1.202103.155

권순선. (2020). 인공지능과 빅데이터 기술동향. TTA Journal 187, 38-43

김성원, & 박광렬. (2022). 딥러닝을 이용한 법률 분야 한국어 의미 유사판단에 관한 연구. 정보처리학회논문지. 소프트웨어 및 데이터 공학, 11(2), 93-100.

박상언. (2021). 딥러닝 중심의 자연어 처리 기술 현황 분석. 한국빅데이터학회 학회지, 6(1), 63-81. https://doi.org/10.36498/kbigdt.2021.6.1.63

박준호, 윤경일, 민성태. (2019). 인공지능 기반의 챗봇 시스템 기술 동향. 정보처리학회지, 26(2), 39-46.

이로제, 박은환, 이재구.(2020).BERT 파생 모델의 한국어에 대한 성능 비교.한국통신학회 학술대회논문집,(),901-902.

임상범, 박찬준, & 양영욱. (2022). 언어적 특성과 서비스를 고려한 딥러닝 기반 한국어 방언 기계번역 연구. 한국융합학회논문지, 13(2), 21-29. https://doi.org/10.15207/JKCS.2022.13.02.021

임준호, 김현기, & 김영길. (2020). 딥러닝 사전학습 언어모델 기술 동향. [ETRI] 전자통신동향분석, 35(3), 9-19.

조영주, 배강민, 박종열. (2020). GAN 적대적 생성 신경망과 이미지 생성 및 변환 기술 동향. 전자통신동향분석. 35(4), 91-102. https://doi.org/10.22648/ETRI.2020.J.350409

Aziz, M. A., Islam, M. M., Munshi, M. S. P., Samshi, S., & Khan, M. M. (2022, March). Sentimental Analysis on reviews of Protein Supplement using RNN-LSTM. In 2022 6th International Conference on Computing Methodologies and Communication (ICCMC) (pp. 796-800). IEEE.

Bai, H., Zhang, W., Hou, L., Shang, L., Jin, J., Jiang, X., ... & King, I. (2020). Binarybert: Pushing the limit of bert quantization. arXiv preprint arXiv:2012.15701.

Bojanowski, P., Grave, E., Joulin, A. & Mikolov T. (2016). Enriching Word Vectors with Subword Information. Transactions of the Association for

Computational Linguistics 5(1), 135-146.

Brochier, R., Guille, A., & Velcin, J. (2019). Global vectors for node representations. In The World Wide Web Conference. pp. 2587-2593.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

Brygfjeld, S.A., Wetjen, F., & Walsøe, A.U. (2017). Machine learning for production of Dewey Decimal. http://library.ifla.org/id/eprint/2216/

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Frankenreiter, J., & Nyarko, J. (2022). Natural Language Processing in Legal Tech. Legal Tech and the Future of Civil Justice (David Engstrom ed.). http://dx.doi.org/10.2139/ssrn.4027030

Ge, T., Xia, H., Sun, X., Chen, S. Q., & Wei, F. (2022). Lossless Acceleration for Seq2seq Generation with Aggressive Decoding. arXiv preprint arXiv:2205.10350.

Gritsenko, A., Salimans, T., van den Berg, R., Snoek, J., & Kalchbrenner, N. (2020). A spectral energy distance for parallel speech synthesis. Advances in Neural Information Processing Systems, 33, 13062-13072.

Hwang, S., & Kim, D. (2020). BERT-based classification model for Korean documents. Journal of Society for e-Business Studies, 25(1).

Karatzoglou, A., Jablonski, A., & Beigl, M. (2018, November). A Seq2Seq learning approach for modeling semantic trajectories and predicting the next location. In Proceedings of the 26th acm sigspatial international conference on advances in geographic information systems (pp. 528-531).

Kun, K. W., Racharak, T., Yiming, C., Cheng, P., & Le Nguyen, M. (2021, November). KGWE: A Knowledge-guided Word Embedding Fine-tuning Model. In 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI) (pp. 1221-1225). IEEE.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). ALBERT: A lite bert for self-supervised learning of language

representations. arXiv preprint arXiv:1909.11942.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461.

Lin, X. V., Mihaylov, T., Artetxe, M., Wang, T., Chen, S., Simig, D., ... & Li, X. (2021). Few-shot learning with multilingual language models. arXiv preprint arXiv:2112.10668.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., & Stoyanov, V. (2019). RoBERTa: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Mehta, S. V., Rao, J., Tay, Y., Kale, M., Parikh, A., Zhong, H., & Strubell, E. (2021). Improving Compositional Generalization with Self-Training for Data-to-Text Generation. arXiv preprint arXiv:2110.08467.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Park, J., & Zhao, H. (2020, December). Korean Neural Machine Translation Using Hierarchical Word Structure. In 2020 International Conference on Asian Language Processing (IALP) (pp. 294-298). IEEE.

Pennington, J., Socher, R., & Manning, C. D. (2014, October). GloVe: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv. https://doi.org/10.48550/arXiv.1910.10683

Malmsten, M., Börjeson, L., & Haffenden, C. (2021). Playing with Words at the National Library of Sweden: Making a Swedish BERT. https://doi.org/10.48550/arXiv.2007.01658

Sakhrani, H., Parekh, S., & Ratadiya, P. (2021, December). Transformer-based Hierarchical Encoder for Document Classification. In 2021 International Conference on Data Mining Workshops (ICDMW) (pp. 852-858). IEEE.

Shoeybi, M., Patwary, M., Puri, R., LeGresley, P., Casper, J., & Catanzaro, B. (2020). Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism. arXiv. https://doi.org/10.48550/arXiv.1909.08053

Singh, S., & Mahmood, A. (2021). The NLP cookbook: modern recipes for transformer based deep learning architectures. IEEE Access, 9, 68675-68702.

Siriwardhana, S., Weerasekera, R., Wen, E., & Nanayakkara, S. (2021). Fine-tune the Entire RAG Architecture (including DPR retriever) for Question-Answering. arXiv preprint arXiv:2106.11517.

Stuart Russell, Peter Norvig. (2020). Artificial Intelligence: A Modern Approach(4th Ed). Pearson

Sun, Y., Wang, S., Feng, S., Ding, S., Pang, C., Shang, J., & Wang, H. (2021). Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. arXiv preprint arXiv:2107.02137.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N. & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Wang, N., Issa, R. R., & Anumba, C. J. (2022). NLP-Based Query-Answering System for Information Extraction from Building Information Models. Journal of Computing in Civil Engineering, 36(3), 04022004.

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. (2016). The FAIR guiding principles for scientific data management and stewardship. Scientific Data, 3, 160018, DOI: 10.1038/sdata.2016.18

Xu, Q., Likhomanenko, T., Kahn, J., Hannun, A., Synnaeve, G., & Collobert, R. (2020). Iterative pseudo-labeling for speech recognition. arXiv preprint arXiv:2005.09267.

Zhang, X., Chen, M. H., & Qin, Y. (2018, September). NLP-QA Framework Based on LSTM-RNN. In 2018 2nd International Conference on Data Science and Business Analytics (ICDSBA) (pp. 307-311). IEEE.

Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., & Liu, Q. (2019). ERNIE: Enhanced language representation with informative entities. arXiv preprint arXiv:1905.07129.

○ **보고서**

과학기술일자리진흥원. (2019). 인공지능(빅데이터) 시장 및 기술 동향. S&T Market Report. https://me2.kr/eat12

과학기술정보통신부. (2018). I-Korea 4.0 실현을 위한 인공지능(AI) R&D 전략.

과학기술정보통신부. (2019). 인공지능 국가전략.

과학기술정보통신부. (2021). 디지털 전환에 따른 주요국 정책 거버넌스 비교 연구. 한국행정학회.

국립외교원. (2021). 바이든 행정부의 인공지능 국가정책:평가와 함의. 외교안보연구소.

국회도서관. (2021). 2020 국회도서관 연간보고서. (31-9720087-000749-10). https://me2.kr/1ohns

국회입법조사처. (2019). 국회입법조사처 업무의 인공지능(AI) 도입 가능성 연구. 경인교육대학교 입법학센터. .https://www.nars.go.kr/report/view.do?cmsCode=CM0010&brdSeq=26466

서교리.(2020).인공지능 기반 챗봇 서비스의 국내외 동향분석 및 발전전망. (2018-2). 한국정보화진흥원

정보통신기획평가원. (2021). 인공지능과 자연어 처리 기술 동향. (제1983호).

정보통신기획평가원. (2020). 인공지능 기술청사진 2030

한국과학기술기획평가원. (2017). 인공지능 분야 국내외 동향분석 및 투자전략 수립. (2018-035). 과학기술정보통신부. https://doi.org/10.23000/TRKO201800036018

한국정보화진흥원(NIA). (2019). [IT & Future Strategy 보고서 2019-5] 디지털 트랜스포메이션 성공전략 -주요국 국가정책 분석을 중심으로-. (제2019-5호)

한국정보화진흥원(NIA). (2021). [IT & Future Strategy 보고서 2021-1] 주목받는 인공지능(AI) 9대 핵심 기술 분석 및 주요 시사점. (제2021-1호).

한국정보화진흥원(NIA). (2021). [IT & Future Strategy 보고서 2021-2] 초대규모 AI 모델(GPT-3)의 부상과 대응 방안. (제2021-2호).

Nathan Benaich·Ian Hogarth. (2020). State of AI Report 2020

○ **단행본**

사이토 고키. (2019). 밑바닥부터 시작하는 딥러닝 2. 한빛미디어.

○ **웹페이지**

송현경. (2016.03.07). 국회전자도서관, 검색어 '인포그래픽' 제공. 내일신문.
http://m.naeil.com/m_news_view.php?id_art=187330

유원준. (2022.05.27). 딥 러닝을 통한 자연어 처리 입문. Wikidocs.
https://wikidocs.net/book/2155

이기창. (2022.06.05). 트랜스포머 살펴보기. ratsgo's NLP.
https://ratsgo.github.io/nlpbook/docs/language_model/transformers/

조명의. (2021.09.08). 중국 신인프라 산업 정책 동향 및 시사점. 테크월드뉴스.
http://www.epnc.co.kr/news/articleView.html?idxno=214547

KISTEP. (2022.05.22). [이슈분석 140호] 중국 신인프라 산업 정책 동향 및 시사점. S&T GPS.
https://now.k2base.re.kr/portal/issue/ovseaIssued/view.do?poliIsueId=ISUE_000000000000938&menuNo=200&pageIndex=3

LG 경제연구원, 이승훈. (2017). 최근 인공지능 개발 트렌드와 미래의 진화 방향.
https://www.lgeri.co.kr/uploadFiles/ko/pdf/busi/LGERI_Report_20171010_20170110170112387.pdf

A.I. Architecture Intelligence. (2016.05.04). Future Architecture.
https://futurearchitectureplatform.org/news/28/ai-architecture-intelligence/

BIBFRAME model. (2016.04.21.). Library of congress.
https://www.loc.gov/bibframe/

BnF. DIGITAL SCHEME. https://www.bnf.fr/fr/schema-numerique-bnf

Convolutional neural network. (2022.06.14). Wikipedia.
https://en.wikipedia.org/wiki/Convolutional_neural_network

DCMI Metadata Terms. (2020.01.20.). Dublin Core Metadata Innovation.
https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

International Council on Archives Records in Contexts Ontology (ICA RiC-O) version 0.2. (2021.02.12.). International Council on Archives.

https://www.ica.org/standards/RiC/RiC-O_v0-2.html

Meghan Rimol. Gartner Identifies Key Emerging Technologies Spurring Innovation Through Trust, Growth and Change. (2021.08.23). Gartner. https://www.gartner.com/en/newsroom/press-releases/2021-08-23-gartner-identifies-key-emerging-technologies-spurring-innovation-through-trust-growth-and-change

Schema.org. (2022.03.17.). Schema.org. https://schema.org/