2021 NAVER AI NOW

내용정리: 양석환

Part 1. CLOVA, 커다란 가능성을 열다

1. Keynote

- 기존 AI를 뛰어넘는 **Big AI** 출현 ⇒ 모델의 플랫폼화, 일반화 및 확장이 가능해 짐
  - 모델 크기(파라미터 수 기준)의 증가 ⇒ 새로움 문제의 해결 가능성 커짐
    ⇒ 더 큰 모델, 더 빠른 학습 추구 ⇒ 글로벌 주도권 경쟁
  - 슈퍼컴퓨터 도입 ⇒ 초 대규모 AI 추진

- **HyperCLOVA** : 네이버가 추진하는 새로운 빅 AI 개발 방법론
  - 핵심요소
    ① 슈퍼컴 인프라: 2020년 슈퍼컴퓨터 인프라 도입 및 구축
    ② 데이터: (네이버 플랫폼 운영 노하우를 통한) 고품질 데이터 확보
    ③ AI 전문가: 지속적인 전문가 확보, 국내 기업 중 최다 실적의 기술논문 발표

  - 초 대규모 언어모델. 공개 당시의 GPT-3을 넘어섬

- **글로벌 AI 연구생태계 구축**
  - HyperCLOVA의 활용, 확산
    ① 활용 예 1: 기획 작업 시 마케팅 문구 고민 중 ⇒ HyperCLOVA가 키워드, 예시 구문 제시 ⇒ 제시된 자료들을 이용하여 마케팅 문구 기획
    ② 활용 예 2: 공부 ⇒ 백과사전 같은 지식을 논리적으로 풀어서 설명

  - SME(중소기업, 벤처기업 등) + Creator ⇒ 서비스와 기회 제공 ⇒ 10개 이상의 서비스에 기술을 적용하여 곧 공개할 예정
  - 더 많은 협력이 필수 ⇒ 서울대 카이스트 등 협력, 공동연구소 설립 등
  - 연구, 윤리: 첨단 기술 ⇒ 일상에 적용, 실행가능하고 구체적인 기술

- HyperCLOVA를 함께 만들어 나갈 파트너를 기다림

2. 새로운 AI의 시작, HyperCLOVA

- HyperCLOVA
  - 최초의 한국어 특화 초 대규모 AI
  - 5,600억개의 한국어 토큰 구축
  - 한국어 데이터 학습량: GPT-3의 6,500배, 기존 네이버 언어모델의 3,000배 규모
  - 700PF(Peta Flops)급의 슈퍼컴퓨터 기반 인프라
  - 204B(2040억개)의 파라미터 규모(GPT-3은 175B)

- **Big AI 가능해진 이유**
  - 비지도 학습 기술연구의 활성화 ⇒ 사람이 알려주는 데이터가 없어도 큰 모델의 학습이 가능해짐 ⇒ 대규모 AI 개발이 가능해 짐
  - 인공지능 성능은 데이터 양 연산규모 파라미터 수에 좌우됨
  - 병목 지점이 없다면 무한히 성능 향상 가능
  - 큰 모델이 더욱 경제적, 효과적

- **HyperCLOVA를 활용할 경우, 기존 AI 방법론보다 수백 배의 속도로 개발 가능**

  ① 맥락의 이해, 공감 유도, 자연스러운 대화 가능
  - 데이터 셋 구축 불필요
  - 사용자의 만족감 인식, 호응, 디테일한 요소 파악 ⇒ 연결된 대화가 가능

  ② 창작을 도와주는 글쓰기
  - 여러 글을 학습, 창작자를 도와줄 수 있는 능력 확보
  - 상품의 이름, 속성, 소개를 보고 홍보문구를 만들어 낼 수 있음
  - 내부 품질 적합률 99%

  ③ 정보 요약
  - 주제에 대한 여러 의견을 조사하는 과정 대신 요약을 이용하여 처리 가능
  - 처리 수준도 기존 기술보다 뛰어남

  ④ 데이터 생성
  - 학습을 위한 데이터 셋 자동생성
  - 기존 지도학습 레이블링 과정 삭제 또는 최대한 축소
  - AI가 만든 문장의 필터링만으로도 연설문 작성 가능

- **HyperCLOVA 활용 성과**
  - GPT-3 이상의 결과가 이미 도출되고 있음
  - AI의 활용이 훨씬 빠르고 쉬워지고 있음
  - AI 시스템을 만들기 위한 데이터 전문성 시간 노력 획기적으로 감소 ⇒ 누구나 쉽게 AI를 사용 가능
  - 이후에는 기술자가 아닌 기획자가 AI 시스템을 만들어나갈 수 있게 지원 가능
  - 플랫폼의 개발 및 공개, 활용 지원

3. HyperCLOVA를 위한 슈퍼컴퓨팅 인프라

- **슈퍼컴퓨팅 인프라의 구축**
  - Big AI 모델을 위해 구축됨 ⇒ 글로벌 TOP 500 중 상위권
  - GPT-3등의 초 대규모 AI는 일반적인 GPU서버로는 불가능 ⇒ 슈퍼컴퓨터 필수
  - 현재 한국어 일본어 초 대규모 AI 구축 중

- **슈퍼컴퓨팅 인프라의 구성**
  - 고성능 병렬 GPU 클러스터
  - 초저지연 고대역폭 네트워크: OS 접근 없이 네트워크상에서 처리 가능
  - 고성능 병렬 아키텍처 스토리지

- **네이버의 역량 활용**
  - 클라우드 인프라 운영 역량
  - 데이터 센터 구축 노하우
  - 모니터링 플랫폼과 운영 자동화

- **향후 계획**
  - 슈퍼컴퓨팅 클러스터 확장 + 다양한 AI 가속 솔루션 모색
  - 혁신적인 AI 생태계 ⇒ 서비스 핵심 플랫폼 역할 ⇒ 클라우드 플랫폼

4. HyperCLOVA를 위한 빅데이터

- **딥러닝에서 중요한 것**
  - 다양한 내용
    - 데이터의 구성

<table>
<tr><td>검색 허용된 문서</td><td>신뢰할 수 있는 출처의 오픈된 리소스</td><td>전문지식</td></tr>
<tr><td colspan="3">기반 지식</td></tr>
</table>

  - 범용의 구성
    - 유의미한 구조로 구성
    - 문서내용 + 메타정보 추가
  - 양질의 정보
    - 영역 선별 : 상위품질의 데이터를 중심으로 확보, 그 중에서도 가치, 중요성 등으로 선별 사용, 핵심 영역 데이터만 사용
    - 저품질 문서 필터링: 의미없는 단어의 나열, 비속어나 유해정보 제거, 서비스별 홍보 및 스팸 판별 결과 활용
  - 충분한 크기
    - 최종 1.96테라 데이터셋 구축 ⇒ 5600억 토큰 ⇒ 한국어 위키피디아 2900배, 50년치 뉴스, 네이버 블로그 9년치
    - 한국어 데이터 양 6500배(GPT-3 대비)

- **HyperCLOVA 활용 테스트**
  - 전문적 내용, 사투리, 욕설 등의 순화 작업 테스트에도 좋은 성능 보임
  - 언어 외의 음성, 이미지 등 데이터 셋 구축도 계속 추진 중

5. 새로운 글로벌 AI R&D 리더십

- **HyperCLOVA의 개발과 개선**
  - (1)공개된 기술 적용 vs (2)자체 기술 개발
    - (1)의 경우 속도경쟁에서 이길 수 없다
    - 네이버의 투자 결과, 혁신적 성과 도출 중
    - 연구 결과는 엔진으로 구현, 서비스/프로덕트/디바이스 등을 통해 제공
    - 결과 데이터는 다시 연구에 반영하는 선순환 구조

- **개발을 위한 새로운 생태계 구축**
  - AI 연구동향 핵심 ⇒ Big AI, Big Model로 진행 중
  - 모델, 데이터, 전문가 ⇒ BIG ⇒ 기업중심의 연구 생태계 ⇒ 새로운 산학 연구 생태계 도출 중

- **AI R&D 리더십 제시**
  - 연구관점 ⇒ 전문 기술논문 발표건수 ⇒ 꾸준히 증가 중, 국내 최대 실적
  - 대부분의 연구결과는 직간접적으로 실제 서비스에 활용
  - 일반 기업은 응용에 치우침
  - 네이버 ⇒ 최적화, 모델, 기법, 인프라, 데이터 등 전 분야에서 강세 보임

- **현재의 상황**
  - 최선을 다하고 있으나 자체만으로는 부족 ⇒ 글로벌 생태계 구축 중
  - 서울대 AI 연구원, 카이스트 AI 대학원 각각과 함께 초거대규모 AI 구축 중
  - 자동화된 창작 능력을 가진 AI 개발 등 연구
  - 한국어 자연어 이해에 대한 데이터가 매우 부족한 상황
    ⇒ 다양한 협업으로 데이터와 모델 공개 예정
    ⇒ 전체 생태계가 글로벌 지속 성장할수 있도록 지원
  - 선지원 No고민 부탁~

6. AI, 사람을 위한 일상의 도구

- **아젠다 리서치**
  - 학계와 협업 중

- **네이버 AI 윤리 준칙**
  - 사람을 위한 AI 개발
  - 다양성의 존중
  - 합리적 설명과 편리성 조화
  - 안전을 고려한 서비스
  - 프라이버시, 정보 보호

- **사례**
  - CLOVA Care Call: 코로나 분석 관리 방역 도구
    - 보건소 도입, 접촉자 자동 관리/증세확인
    - 성남시 기준 10만 건 이상 수행
    - 2천명 이상 유증상자 신규 파악 성과
    - 일상에서 SME(중소기업)의 사업을 도와주는 도구인 CLOVA Ai Call을 기반으로 개발됨

  - CLOVA 램프
    - 아동, 초등학교 저학년에게 종이책을 즐겁게 읽을 수 있도록 도와주는 디바이스
    - 문자인식, 이미지인식, 음성인식, 음성합성, 자연어처리 기술 활용

  - 다양한 영역에서 활용되는 AI
    - 챗봇, 문자인식, 상품추천, 객체추적, 기계번역, 음성인식, 합성, 얼굴인식, 음악추천, 클라우드 등
    - 문서번역, 대화, 상식 등, 그림과 음성의 혼합 활용

- **AI 기술과 일상**
  - AI가 완벽할 수는 없다
  - 사람을 위한 일상의 도구가 되도록 지속 개선 노력, 학계와 협력 추진
  - 사용자의 눈높이에서 AI를 쉽게 설명하는 노력 지속
  - 사례 중심의 AI리포트 작성, 공개
  - 윤리준칙, 협력 등 스타트업에도 제공, 연계
  - AI는 우리 사회의 구성원 모두가 공유하는 기술, 사회 자산이 되어야 함

Part 2. NAVER AI Technology

1. HyperCLOVA의 새로운 한국어 모델

- GPT-3와 HyperCLOVA
  - GPT-3는 학습 데이터 구성상 한국어 성능은 매우 제한적
  - GPT-3는 약 93%가 영어, 한국어는 0.1% 미만 ⇒ 사실상 영어 전용 모델
  - 그러나 내재된 가능성은 매우 높음
  - 대부분의 NLP에서 뛰어난 성능을 보이며 과거 불가능했던 영역 중 일부를 가능하게 만듬
  - 만약 한국어용 Big AI, Big Model을 확보하지 못하면, 기술 종속에 빠질 수 있으며 한국어 기반 서비스 성장의 걸림돌이 됨
  - 한국어 97% 비율의 한국어 특화 모델 개발 필요성 ⇒ HyperCLOVA 개발

- **HyperCLOVA의 한국어 모델**
  - 데이터 말뭉치
    - 어휘 집합 학습
    - 코퍼스 믹서(전처리 시 데이터 종류별 비율 자동 조절) ⇒ 시리얼라이저(하둡 스트리밍 적용 처리속도 10배 개선)

  - 개발을 통해 얻고자 하는 기술성과 및 문제점 개선
    - 다양한 형태의 글을 이해, 생성하는 능력 ⇒ 문장 형태 생성을 기본 아키텍처로 선정
    - 모델이 1000배 증가하는 동안 GPU메모리는 4~5배 증가 ⇒ 3중 병렬화를 통해 극복
    - 모델이 커지면 학습의 난이도 증가, 실패 기회비용 증가 ⇒ 극복
    - 모델을 키울수록 초반에는 속도가 느리지만 시간이 지남에 따라 더 뛰어난 성능과 속도를 보임

  - 하이퍼클로버가 한국어를 읽는 방법 ⇒ 토큰화
    - 기계가 글을 이해하려면?
      - 문장을 특정 단위로 끊어 읽는 기능. 기계는 능력을 정해주어야 함 ⇒ 토큰화
      - 서브워드로 끊어 읽으면 규칙 기반의 장점들을 취하면서 단점을 줄임
      - 말뭉치로부터 병합의 대상이 도는 문자열에 적절한 전처리 수행 ⇒ 서브워드 학습 방식에 변화를 줄 수 있음

    - 대용량 말뭉치로 서브워드 토크나이저 학습하기
      - 말뭉치 전체로 서브워드 토크나이저 학습은 현실적 불가능
      - 어휘집합 간 중복되는 토큰을 감안, 말뭉치 양 조절하면서 진행 ⇒ 1%만 사용해도 가능함을 확인
      - 1%도 매우 많은 데이터 ⇒ 메모리 문제 등 다양한 구조적 문제 발생

    - 언어모델을 위한 서브워드 토크나이저
      - 미등록 단어 문제 극복
      - 고빈도 형태소 기반 학습 Morpheme-Aware Byte-Level BPE 적합 판단
      - 미등록 단어 존재할 경우 학습 난이도가 증가함
      - 좋은 서브워드 토크나이저로 만든 언어모델이 생성한 문장은 사람의 문장과 비슷하다고 가정 ⇒ 지표화를 통해 판별모델 도입
      - 학습용 말뭉치의 1%로부터 학습한 Morpheme-Aware Byte-Level BPE 토크나이저로 만든 언어모델이 좋음

  - 지표
    - 설계목적
      - 모델이 생성한 문장은 얼마나 유창한가? 유창성, 정도의 수치 측정

    - 신 모델의 능력에 비하여 평가지표는 구형
- 문제점
    - 생성문장과 레퍼런스 문장간의 유사성이 모델의 품질을 보장하지 않는다 (Blue Score 활용 계산하여 측정)
    - 서로 다른 설정(특히 어휘집합)에서 학습한 모델들을 퍼플렉시티(PPL)로 비교하는 것이 불가능함

- 개선 아이디어
    - 레퍼런스 없는 평가지표 ⇒ 판별자를 GAN 응용
    - 생성문장 기반 평가지표
    - 판별정확도 / 판별 F1

- 평가결과
    - Blue 점수보다 DA 점수가 더 적합한 평가가 가능함
    - 모델이 충분한 학습을 수행한 경우 모델이 클수록 성능이 뛰어남

2. HyperCLOVA Studio: 나에게 필요한 인공지능, 내 손으로 손쉽게 만들기

- HyperCLOVA Studio
    - HyperCLOVA 개발환경을 위한 도구
    - 코딩 없이 서비스의 기반이 되는 머신러닝 기반 시스템의 개발을 지원
    - 기존의 개발 프로세스를 완전히 바꿈
    - 자연어로 개발하고자 하는 머신러닝 시스템을 설명(입력) 후 어떤 방법으로 생성할지 선택하면 AI 모델 자동 생성됨
    - 예시
        - 몇 개의 예제만으로 감정을 분석하는 시스템을 개발하는 화면. 문법에 맞지 않거나 은유적인 표현도 인식
        - 몇 개의 예제만으로 이메일에 대한 답장을 작성하는 시스템 개발의 예제

3. HyperCLOVA의 활용 1: 검색 어플리케이션

- Null 검색 질의(Query) 재작성
    - 결과가 없는 질의를 재작성하여 결과를 얻을 수 있게, 더 유연하고 명확하게 검색할 수 있도록 바꿔줌
- 쇼핑 리뷰 요약
    - 많은 리뷰 데이터를 클러스터링하고 그 중 대표적인 데이터를 선택하여 한줄 요약으로 문장을 생성함
    - 실제 적용 시 약 21%의 입력 리뷰에 없는 데이터가 나오는 경우 발생 ⇒ 대응 처리에서 NLI가 더 좋은 성과를 얻음을 확인
- 질의응답
    - 지식 백과 기반의 질의응답 작업

- 질문에 대한 정보의 검색, 선택, 신뢰성 판단 등의 대한 검증 등, 답변에 대한 질문을 생성하여 교차 팩트체크 등 수행 ⇒ 최종적으로 답변의 신뢰성 평가
- 아직 미진한 부분이 있지만 기존에 수행하기 어려웠던 다양한 질문, 요청 등에 대한 처리가 가능해 졌음

4. HyperCLOVA의 활용 2: AI 어시스턴트

- 목적지향형 대화 + 질의응답 + 일상대화 ⇒ 결과를 병렬 생성하여 최적의 결과를 선택하는 방식
- 생성 기반 모델인 HyperCLOVA
- 자연스러운 대화, 보다 많은 지식 ⇒ 기술 개발
- 주요 기능
  - 대화 이해: 대화 이해를 위한 컨텍스트 주입. 대용어 해소, 대화상태 추적기술 등이 필요함. 생성모델 활용
    - 질문이 있을 때 이전의 질문 문장을 현재의 질문에 주입하여 처리함으로써 문맥의 연결 등을 추구함
    - 언어모델의 크기가 성능 개선에 큰 영향을 미침

  - 시스템 응답 선택: 시멘틱 검색 이용 응답 선택
    - 사례: 미세먼지 좀 알려줘
      - 목적지향형 대화(날씨 정보 등)
      - 질의응답(미세머지에 대한 전문 지식)
      - 일상대화

      ⇒ 3개의 내용을 함께 고려하여 가장 적절한 대답을 선택, 생성하여 제시

  - 시스템 응답생성: 조율 방법
    - 이후 세션에서 다룸

- 향후
  - 실 서비스 적용을 위한 최적화
  - 멀티턴 대화에 자연스러운 적절한 VUX 탐색
  - 품질향상과 효과적 제어 위한 최신기술 적용

5. HyperCLOVA의 활용 3: 대화

- 기존 대화 어플리케이션: 스피커, 해피콜, AiCall 등
  - 말투나 대화체의 특징이 없음 ⇒ 어시스턴트로만 생각해서 개발한 것이 원인

- 캐릭터 대화 AI (내 어시스턴트에 캐릭터를 입힐 수 있을까?)
  - 캐릭터 대화는 초대형언어모델 AI 만으로는 구현 불가능
  - 캐릭터 대화의 요소

        - 일관적 캐릭터 대화체 유지 + 유창성
            - 답변의 대화체(캐릭터 페르소나 탐지기) + 캐릭터 대화체 변환기 + 초 대규모 언어 모델
        - 캐릭터의 세계관 유지(프로필, 배경, 철학)
            - 특정 질문에 대해서 일관성 있는 답변 요구
            - 입력 프롬프트에 동일한 답변을 하는 특정 질문을 다수 준비하여 입력, 학습

    - 기본적인 대화 모델 구성
        - 캐릭터 대화체에 퓨샷 러닝 적용
        - PCU 적용 제안

- AiCall의 미래와 HyperCLOVA
    - 멀티턴 목표지향 대화 구축의 어려움
        - 사람 간 대화 로그가 많아도 그대로 사용어려움
        - 대화 디자이너가 설계한대로 처리됨
        - 실제 사람이 가진 권한 문제
        - 시나리오 관리, 분기 등에 따른 처리
        - DB 정보에 따라 다른 반응을 하도록 처리
    - HyperCLOVA를 이용하여 다양한 예시 발화를 생성하여 학습, 처리 등에 활용, 어울리는 답변의 추천에 응용, 대화 분기 추천 등
    - 연속대화 생성, 대화 시스템 평가 등에 활용
    - HyperCLOVA의 기능을 순차적으로 챗봇 빌더에 적용할 계획

6. HyperCLOVA의 활용 4: 데이터 증강

- 자연어 처리 패러다임의 변화
    - 컴퓨팅 환경의 접근성 문제
    - 초대형 언어모델에서 사전학습된 언어모델을 파인튜닝 패러다임에 적용하는 방식의 한계
    - NLP 파이프라인의 변화
        - 기존 방식(프롬프트를 활용한 방안)의 한계점
            - 최신 프롬프트 기술 활용 위해서는 HyperCLOVA에 역전파 계산 필요
            - 언어모델은 프롬프트 설계에 따라 민감하게 반응하므로 제한된 모델의 통제성 등의 문제가 있음
        - 개선 방안
            - HyperMix(HyperCLOVA를 이용한 텍스트 증강기법)의 활용

- HyperMix
    - 파인튜닝 패러다임과 프롬프트 기반 방법의 대안으로 데이터 증강기법 HyperMix 제시
    - 데이터 수집의 공수 감소 및 NLP 모델 경량화에 초점

- HyperMix 작동방식
  - HyperCLOVA를 통해 기존 예시로부터 새로운 예시 생성
  - HyperCLOVA가 알고 있는 언어적, 상식적 지식이 첨가되어 현실적인 예시 및 높은 정확도의 분류정보 생성
- HyperMix의 효용성 실험결과 소개

7. HyperCLOVA의 조율

- HyperCLOVA를 이용한 개발의 예제를 통해 문제점 및 개선 방향 소개
  - 최근성 편향 문제
  - 다수 레이블 편향 문제
  - 범용 토큰 편향 문제
  - 이산공간과 연속공간에서의 프롬프트 엔지니어링 방법

8. HyperCLOVA를 위한 서비스 기반

- 개발한 모델을 시스템으로 올려서 서비스하기 위한 방법
- 해당 시스템을 위한 하드웨어적 사양, 기술적인 사양 등
- 개발하는 앱의 구조