Survey : Open Source + Input Preprocessing

2조 HIM
TAEKOAN YOO 유태관

0. Background

- **Chat-bot building process**
  - 1. 도입 타당성 검토 : 사용자 편의 제공, 매출 상승…
  - 2. 적용 분야 선택 : 숙박 검색, 숙박 예약 등…
  - 3. 데이터 활용 가능성 : 데이터 확보
  - 4. 설계
    - 적용 범위 : 오프라인 서비스, 홈페이지 서비스…
    - 서비스 지역 및 대상 : 카카오톡 유저, 페북 유저 …
    - 인터페이스 : 대화형, 템플릿(Visual), STT&TTS
  - **5. 개발**
    - **Open API 활용**
    - **Builder 사용 : 노코드 빌더, 코드 활용 빌더**

Outline

- **Outline**
  - 1. General Open Source (Korean LM)
  - 2. To Chat-bot
    - A. Chat-bot Builder
    - B. With open source framework
    - C. With general open source
  - 3. Preprocessing for API

1. General Open Source (Korean LM)

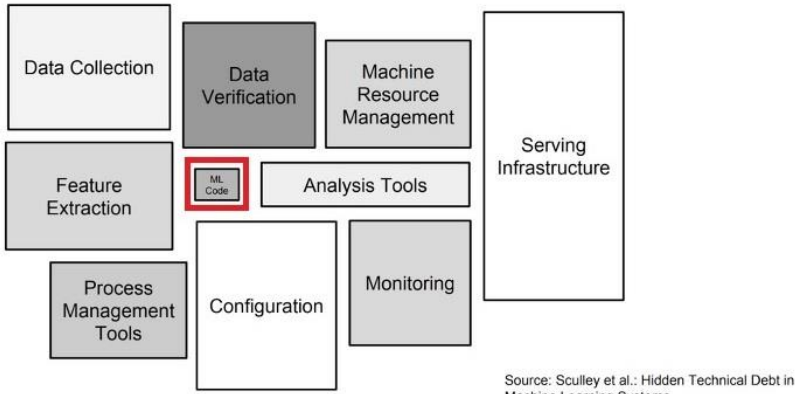
- **Korean Pre-trained Language Model(PLM)**
  - SKT의 **KoGPT-2(‘20)** ← GPT-2
    - https://github.com/SKT-AI/KoGPT2
  - SKT의 **KoBERT(‘18)** ← BERT
    - **https://github.com/SKTBrain/KoBERT.git**
  - TwoBlock AI의 **HanBERT** ← BERT
    - **https://github.com/monologg/HanBert-Transformers**
  - ETRI의 **KorBERT** ← BERT
    - **http://aiopen.etri.re.kr/service_dataset.php**
  - 박장원 **KoELECTRA**
    - https://github.com/monologg/KoELECTRA
  - 박장원 **DistillBERT** ← BERT
    - https://github.com/monologg/DistilKoBERT

2. To Chat-bot : (A) Chat-bot Builder

- **Korean**
  - 네이버 클로바(Naver Clova) https://clova.ai/
  - 다빈치봇(DAVinCIBOT) http://dayliai.com/works_ai/bot.html
  - 단비(Danbee) https://danbee.ai/
  - 라떼(LATTE.AI) https://builder.latte.ai/
  - 봇그리다(BOTGRIDA) https://www.botgrida.com/
  - 심심이(Simsimi) https://workshop.simsimi.com/
  - 에이브릴(Aibril) https://www.aibril.com/
  - 카카오i오픈빌더(Kakao i Open Builder) https://i.kakao.com/
  - 클로저(Closer) https://closer.ai/
  - 톡봇(TalkBot) http://www.saltlux.com/ai/talkbot.do?menuNumber=0202
  - 핑퐁(Pingpong) https://pingpong.us/

2. To Chat-bot : (B) With open source framework

- **Open source framework :**
  - Kochat : https://github.com/sinabro-team/kochat
    - 챗봇빌더 보다는 개발자를 타깃으로 하는 framework
    - API부터 직접 구현하는 어려움을 해결하기 위해 제작

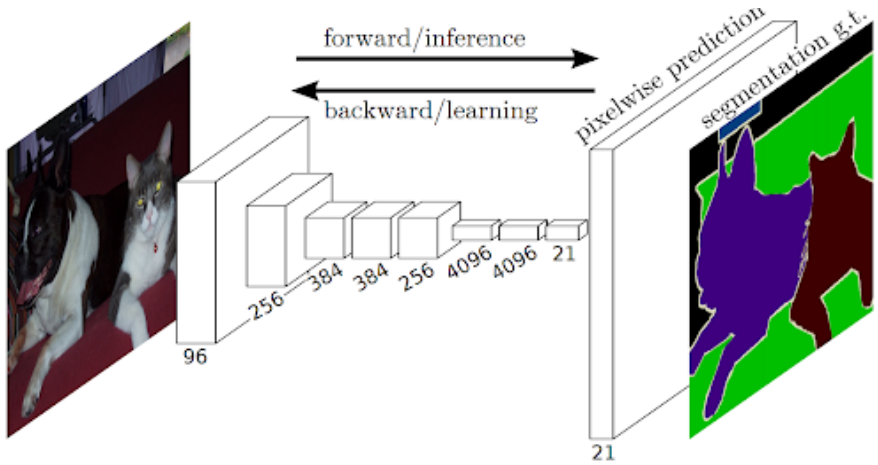


2. To Chat-bot : (C) With general open source

- **Korean Language Model**
  - https://github.com/seunghee63/ChatBot
    - Hard Coding (Seq2Seq)
  - https://github.com/haven-jeon/KoGPT2-chatbot
    - -> KoGPT2
  - https://github.com/nawnoes/WellnessConversationAI
    - -> KoGPT2, KoELECTRA, KoBERT

3. Preprocessing for API

- **Preprocessing = Pre-training:**
  - task에 맞게 전처리 기준을 세우는 것
  - Keypoint for Korean
    - 한자, 일부 특수무자 제거 필수
    - 한국어 문장 분리기 사용 (kss)
    - 뉴스 관련 문장 제거
      - Ex) 무단전재,
        (서울=뉴스1) 등 포함시



- **참고: https://monologg.kr/2020/05/02/koelectra-part1/**

3. Preprocessing for API

- **Corpus to Vocab**
  - Character level : ㄱ,ㄴ,ㄷ,ㄹ,ㅏ,ㅑ …
  - Space level : 책이, 책을, 책에, 책은 …
    - 빈도수 낮은 단어는 학습이 잘되지 않는다.
  - Subword level : 겨울, 이, 되어, 서, 날씨, 가 …
    - OOV를 해결하기 위해
    - Algorihms
      - BPE(Byte-Pair Encoding) : 점진적으로 병합
        - » Byte-level BPE : GPT2
      - Wordpiece : 점진적으로 병합
        - » BERT, ELECTRA
      - Sentencepiece : subword에서 시작해 점차 줄여나가는 식
        - » 대부분의 한국어 개발자
      - Unigram Language Model : 언어에 특화

3. Preprocessing for API

- **Subword-based Tokenizer**

```
from transformers import GPT2Tokenizer
```

  - Huggingface.tokenizer
    - Wordpiece vocab을 만들 때
    - https://github.com/huggingface/tokenizers
  - Google.sentencepiece
    - https://github.com/google/sentencepiece
  - ETC :
    - Huggingface.tranformer에 BertTokenizer, XLNetTokenizer
    - OpenNMT
    - Rule-based : spaCy, Moses
- **Korean 형태소 분석기**
  - Mecab, Khaiii(Kakao), OKT(Open Korean Text), 한나눔, 꼬꼬마, KOMORAN

3. Preprocessing for API

- **Example :**
  - https://github.com/haven-jeon/KoGPT2-chatbot
    - MXNet, Pytorch version
    - 'Chatbot_data/ChatbotData.csv‘ : 원하는 chatbot 데이터얻기
      - 학습할 q와 a를 얻는다
    - Kogpt2.utils.get_tokenizer : tokenizer 다운로드
      - Sentencepiece.tokenizer : 학습된 tokenizer로 생성
        - » Gluonnlp.data : MXNet
      - Q와 a에 대해 vocab형태로 바꾼다.
    - Kogpt2.mxnet_kogpt2.get_mxnet_kogpt2_model : vocab, model
      - Vocab이 들어 갈 수 있는 padder를 만들어 놓는 것
      - 위의 vcab형태의 q,a를 넣는다

3. Preprocessing for API

ChatbotData

Q	A	label
12시 땡!	하루가 또 가네요.	0
1지망 학교 떨어졌어	위로해 드립니다.	0
3박4일 놀러가고 싶다	여행은 언제나 좋죠.	0
3박4일 정도 놀러가고 싶다	여행은 언제나 좋죠.	0
PPL 심하네	눈살이 찌푸려지죠.	0
SD카드 망가졌어	다시 새로 사는 게 마음 편해요.	0
SD카드 안돼	다시 새로 사는 게 마음 편해요.	0

- 인풋 : [q_token, a_token]　　mask　　labels
- 설명 : Token_ids,　　mask,　　label
- 설명 :　　a에만 마스크　　q에만 마스크

Conclusion

- **1. 방법 선택**
  - A. Chat-bot Builder
  - B. With open source framework
  - C. With general open source
  - Or from-the-scratch?
- **2. API**
  - KoGPT2 활용하는 것이 제일 좋을 듯
    - https://github.com/haven-jeon/KoGPT2-chatbot
  - 심리상담 관련해서 Open source 참고하는 것도 좋을 듯
    - https://github.com/nawnoes/WellnessConversationAI

1. General Open Source (Korean LM)

	Task	Feature	
KoGPT2	질의 -> 답변 text		
KoBERT		Vocab size가 작다	
HanBERT		Ubuntu에서만 가능	
KorBERT		API신청 필요	
KoELECTRA			

2. To Chat-bot : Chat-bot Builder

- **ETC**
  - ActiveChat https://activechat.ai/
  - Amazon Lex https://aws.amazon.com/ko/lex/
  - Azure Bot Service https://azure.microsoft.com/en-gb/services/bot-service/
  - Botkit https://botkit.ai/
  - BotMock https://botmock.com/
  - Botsify https://botsify.com/
  - ChatBot https://www.chatbot.com/
  - ChatbotsBuilder https://www.chatbots-builder.com/
  - Chatfuel https://chatfuel.com/
  - ChatScript http://brilligunderstanding.com/
  - ChatterOn https://www.chatteron.io/
  - Collect.chat https://collect.chat/
  - Converse AI http://www.converse.ai/
  - Dialogflow https://dialogflow.com/
  - Drift https://www.drift.com/learn/chatbot/
  - Facebook Messenger Platform https://developers.facebook.com/docs/messenger-platform/
  - Flow XO https://flowxo.com/

2. To Chat-bot : Chat-bot Builder

- **ETC**
  - Gupshup https://www.gupshup.io/developer/bot-platform
  - IBM Conversation One https://www.ibm.com/us-en/marketplace/conversation-one
  - IBM Watson Conversation https://www.ibm.com/watson/kr-ko/developercloud/conversation.html
  - Intercom https://www.intercom.com/
  - Kore.ai https://kore.ai/
  - Landbot.ai https://landbot.io/
  - LivePerson https://www.liveperson.com/products/ai-chatbots/
  - ManyChat https://manychat.com/
  - Microsoft Bot Framework https://dev.botframework.com/
  - MobileMonkey https://mobilemonkey.com/
  - Pandorabots https://www.pandorabots.com/
  - PlayChat https://www.playchat.ai/
  - Reply.ai https://www.reply.ai/
  - Rundexter https://rundexter.com/
  - SAP Conversational AI https://cai.tools.sap/
  - Sequel https://www.onsequel.com/
  - Smooch https://smooch.io/
  - TARS https://hellotars.com/
  - textit.in https://textit.in/
  - Wit.ai https://wit.ai/