**대본**

#1

안녕하세요. 오늘 발표를 맡은 VIVA Pro조의 조원 손승현이라고 합니다.

#2

오른쪽에 보시면 저희 팀 수행계획서에 있는 간트차트가 있는데, 저희는 9월에 말뭉치 자료 확보, 전처리 과정 및 정제과정, 그리고 기존 프로젝트 및 관련기술 학습을 진행하기로 했기 때문에 발표자료도 이 순서에 맞추어 다음과 같이 첫번째, 저희가 그동안 확보한 말뭉치들을 정리하고, 두번째 전처리 과정을 어떻게 하기로 하였는가에 대한 내용을 발표할 것이고, 마지막으로 저희 팀이 작년 팀의 결과물, 교수님의 머신러닝, 그리고 지난 미팅 때 강의해주신 내용에서 생긴 질문들을 정리하여 발표할 것입니다.

#3

저희가 확보한 첫번째 말뭉치는 UCORPUS\_NE입니다. 이것은 울산대에서 나온 문장 단위의 텍스트에서 단어, 특히 의미 기준으로 나눈 의미역을 기본 단위로 해서 tokenization을 진행한 것입니다. 또 저희가 확보한 두번째 말뭉치는 AIHUB 말뭉치입니다. 이것은 AIHUB 홈페이지에서 다운로드 받은 데이터들이며, 이것 또한 UCORPUS\_NE와 마찬가지로 단어의 의미역을 기본 단위로 해서 tokenizing을 진행한 것을 확인할 수 있었습니다.

사진을 보시면 위가 UCORPUS이고, 아래가 AIHUB의 문장 데이터의 일부입니다.

(애니메이션)

방금 보여드린 문장 데이터들을 tokenizing한 결과는 다음과 같습니다. 전체적으로 둘 사이에 크 차이는 없지만, AIHUB 쪽의 tokenizing이 더 세분화되어 진행된 것을 확인했습니다.

#4

그 다음으로 소개드릴 것은 국립국어원 표준 국어 대사전에 있는 단어 중 명사들을 활용하여 만든 개체명 사전입니다. 보통 사전을 보시면 첫번째 사진처럼 건축, 경제, 교육 등으로 분야가 분류된 경우가 있는데, 표준국어대사전도 그런 방법을 사용하였기 때문에 저희들이 태그를 새로 생성할 때 각 분야들의 명칭을 참고해서 저희가 사용할 수 있는 형태의 데이터로 만들었습니다. 그것이 두번째 사진입니다.

저희가 확보한 말뭉치들을 분석한 결과 대부분이 단어의 의미역 단위로 tokenizing이 되어 있는 것을 확인할 수 있었습니다. 따라서 저희 팀도 향후 전처리 과정에서 단어의 의미역 단위로 tokenizing을 진행할 계획입니다. 다만, 아직 프로젝트를 진행하기에는 조건에 맞는 말뭉치 데이터들이 턱없이 부족하다고 판단하여, 기존에 저희 팀이 설정한 기간인 연말까지는 말뭉치 데이터들을 확보하는 작업을 꾸준히 해야 할 것 같습니다.

#5

다음은 머신러닝을 공부하다가 나온 질문들을 정리했습니다.

우선 첫번째 질문입니다. 작년 팀은 HanBERT를 pre-trained model로 사용하여, BERT에 classification layer를 부착해 학습을 했다고 했는데 Google에서 발표한 BERT와 HanBERT 내부의 BERT는 모델 구성은 같으나, weight만 다른 것 인지 아니면 또 다른 부분이 있는지 궁금해서 질문드리고 싶습니다.

그 다음 질문은, pre-trained model을 사용하는 경우 model의 weight를 이용하는 것으로 이해했는데, 작년팀을 예로 들어서 만약에 HanBERT를 pre-trained model로 가지면, 이번 프로젝트의 model을 위해서는 HanBERT의 classification layer를 제거하고 새로운 classification layer를 정의해서 model을 구성하는 것인지 아닌지에 대해서 질문드리고 싶습니다.

그 다음 질문은, 작년 팀처럼 preprocessed module에 Mecab과 KoNLPy를 붙였듯이, RNN에서 문장을 더 세세하게 분할한다면, 최종 성능에서 차이가 있을지에 대해서 질문드리고 싶습니다.

#6

다음은 역전파의 hidden layer에 관련된 질문입니다.

(사진 띄우고 볼 것!)