**캡스톤디자인 주간 보고서**

|  |  |
| --- | --- |
| 팀원 | 이예준, 박기홍 |
| 주제 | 미리내 – ML based Korean Educational Reconjugator using seq2seq |
| 기간 | 2020. 10. 26. ~ 2020. 11. 01 |
| 진  행  내  용 | 1. **화요일 회의 (10.27)** - CTO분과 미리내 팀 동기의 아이디어 제안으로 공백 단위가 아닌 형태소 그 자체를 단위로 해서 word를 split하고 이를 통해 문맥이 고려된 word embedding vector를 만드는 방안이 채택되었다. CTO분께서 저번 주차에 만들어주신 Mirinae pipeline에 training-pair를 추출해주는 기능을 통해 같은 데이터지만 word embedding vector를 morpheme단위로 사전 훈련을 시킨 뒤에 인코더 및 디코더를 통해 학습을 진행하자는 결론이 나왔다. 2. **10월 지도교수님과의 미팅 (10.28)**  - 이번 10월에도 9월과 마찬가지로 지도교수님인 이영구 교수님과 미팅을 진행하였고 현재 미리내 팀의 진행상황, 즉 현재까지 시도해본 여러 모델과 데이터셋 등에 관련해서 보고드렸고 교수님께서도 아이디어를 제안해주셨다. 3. **using pre-trained embedding vector (10.30)**   - word embedding에 있어서 문맥이 고려되면서도 OOV(Out Of Vocabulary)에 대해서도 대응이 가능한 gensim의 fastText를 활용하여 문장에서의 각각의 형태소들에 대해 embedding vector 를 학습시켰다.  그리고 이를 인코더와 디코더에서의 embedding part에 초기 embedding vector로 사용할 수 있도록 weight를 지정해주었고 학습시킬 문장들에 대하여 각각의 tensor들을 생성해주기 위해서 embedding vector들에 대해 vocab을 맞춰주었다.    결과는 상당히 좋았고 앞의 VV나 VHV와 같은 문장의 의미에 있어서 가장 크게 작용하는 부분에 대해서는 fix시키고 뒤의 높임이나 시제 부분의 미세한 오차는 vector dimension을 높이면 해결될 것 같은 결론이 도출되었다. |