**캡스톤디자인 주간 보고서**

|  |  |
| --- | --- |
| 팀원 | 이예준, 박기홍 |
| 주제 | 미리내 – ML based Korean Educational Reconjugator using seq2seq |
| 기간 | 2020. 10. 05. ~ 2020. 10. 11 |
| 진  행  내  용 | 1. **Model 설계 및 학습 테스트 (10.05)** - Pytorch 공식 홈페이지에 나와 있는 seq2seq model을 참고하여 형태소 단위로 문장을 나누어서 학습시켜 보았다. 하지만 형태소 단위로 vocab-dictionary에 들어가서 완전히 새로운 형태소에 대해선 word embedding 단계에서 막혀버렸다. 2. **Training data level 변경 (10.06)** - 기존의 Morpheme(형태소) level에서 더 적은 양의 training set으로 vocab – dictionary를 구축하고자 jamo-level로 더 분해하여 학습을 진행시켜보았다.   Text  Description automatically generated  - 새로운 test data에 대해서도 model을 돌릴 수 있었으나 학습한 데이터양이 부족해서인지 새로운 tensor를 제대로 예측하지 못 하는 모습을 보여준다.   1. **Wild-card 도입 (10.09 회의)** - 위의 jamo-level로 학습 데이터 양이 많아진다면 좋은 결과를 기대할 수도 있겠지만 한국어 특성상 불규칙적인 부분이 다소 많이 존재한다. 따라서 CTO이신 John wainwright께서 제안해주신 wild-card rule을 설정하기로 했다. 나름의 rule을 설정하고 wild-card를 사용해 변칙적으로 대응할 수 있는 pre-processing이다. 2. **Training set (10.09 회의)** - 또한 training set(pairs) 역시 더 모아야 하기 때문에 이에 대한 방안도 다음주에 고민해보기로 하였다. |