NC soft 블로그 포스트 리뷰 (자연어 처리 기술의 다양한 상황들)

NLP Application Gap

Aa 영역	연구 수준	상용화 수준
환경	이상적인 환경을 가짐	다양한 비 정상 상황 가정
입력 관련	정확한 입력 → 사용자 입력에 모호성이 없고, 문법에 맞는 발화 / 텍스트	부정확한 입력 → 입력의 모호성, 생략, 비문, 신조어, 욕설 등 다양한 구어체적 현상
학습 데이터 관련	학습데이터 존재 → 도메인 특화된 학습데이터 미리 정의된 작은 데이터	학습데이터 구축의 어려움 → 적용하고자 하는 도메인에서의 대용량 학습 데이터가 필요. 빠르게 도메인에 특화시키는 것이 중요
정답	정답이 존재 → 정답이 알려져 있거나 많은 사람들이 정답에 동의 할 수 있는 쉬운 문제	정답을 알 수 없는 문제도 있음 → 문제 자체가 정답을 알기 어렵고, 모든 사람을 만족시키기 어려움
대상 데이터	대상 데이터가 고정	실시간으로 대상 데이터 변함 → 대상이 되는 데이터가 지속적으로 변화하며 미리 예측하지 못한 형태의 데이터가 발생됨

위 표의 상용화 수준의 자연어처리 연구를 하면서 생기는 문제에 착안한 엔씨소프트 언어 AI 랩의 세 가지 연구 소개

1. 구어체 인식: Multi-channel CNN을 이용한 한국어 감성분석 (제30회 한글 및 한국어 정보처리 학술대회 논문)
2. UnSupervised Style Transfer (제30회 한글 및 한국어 정보처리 학술대회 논문)
3. UnSupervised Narrative Learning (EMNLP 2018 논문)

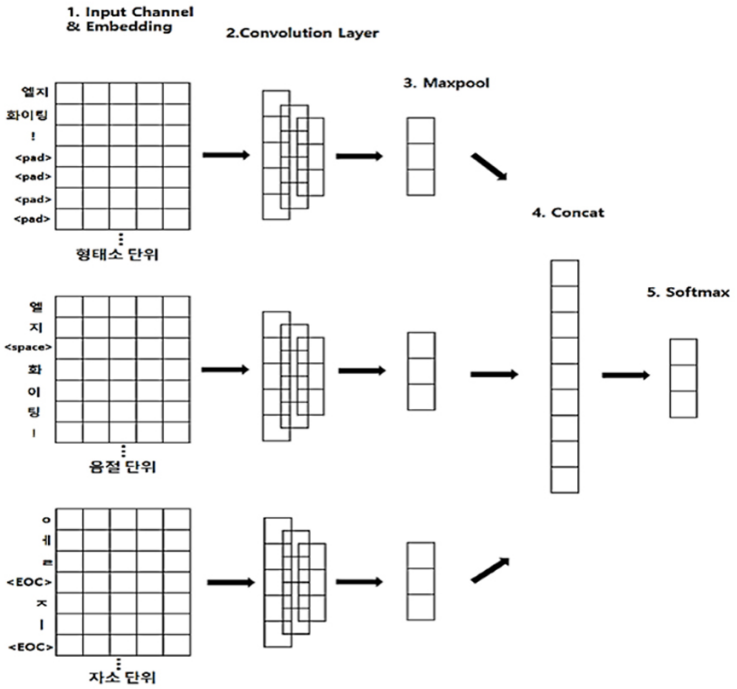
1. 구어체 인식: Multi-channel CNN을 이용한 한국어 감성분석

목표 : 온라인의 실제 구어체 한국어 텍스트의 감성을 텍스트의 품질과 관계없이 인식

문제 : 온라인 구어체 텍스트는 줄임말이나 맞춤법 오류가 많고 띄어쓰기가 맞지 않는 경우가 많음 → 이러한 구어체를 처리하기 위해 별도의 교정기나 형태소 분석기 등을 사용하여 데이터를 처리할 수 있지만, 이러한 분석기 역시 성능이 100%가 아니기에 오류전파의 위험이 있고, 구어체의 특성상 계속 변화하는 구어체 현상을 모델에 반영하기도 어려움

방안 : Multi-channel CNN 방법 고안

Multi-channel CNN은 간단하게 하나의 문장이 형태소, 음절, 자소로 각각 나누어져 세 개의 입력 channel로 사용하고, 임베딩과 CNN을 거쳐 나온 각각의 vector를 concat한후 softmax.



이 네트워크는 자소와 음절에서 feature vector를 추출함으로써 자소 단위의 오타 및 문법적 오류와 음절단위의 합성어와 줄임말에서 손실된 정보를 추출할 수 있음. 또한, 학습하지 못한 새로운 단어에 대해서 형태소가 추출하지 못하는 특징을 자소와 음절에서 추출할 수 있게 됨.

UnSupervised Style Transfer

문제 : 자연어 문체 변환에서는 크게 두가지 문제가 있음

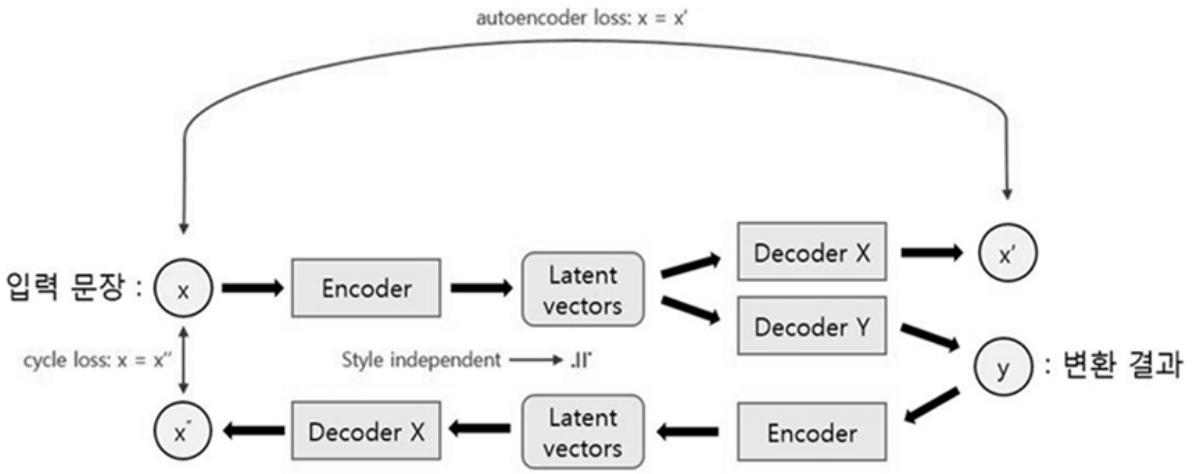
1. 학습데이터로 사용하기 위하여 동일한 의미이나 문체만 다른 문장 쌍을 대량으로 구축하기 어려움

2. 스타일은 변화시키면서 실제 문장의 콘텐츠, 의미는 유지하는 네트워크 구축이 어려움

방안 :

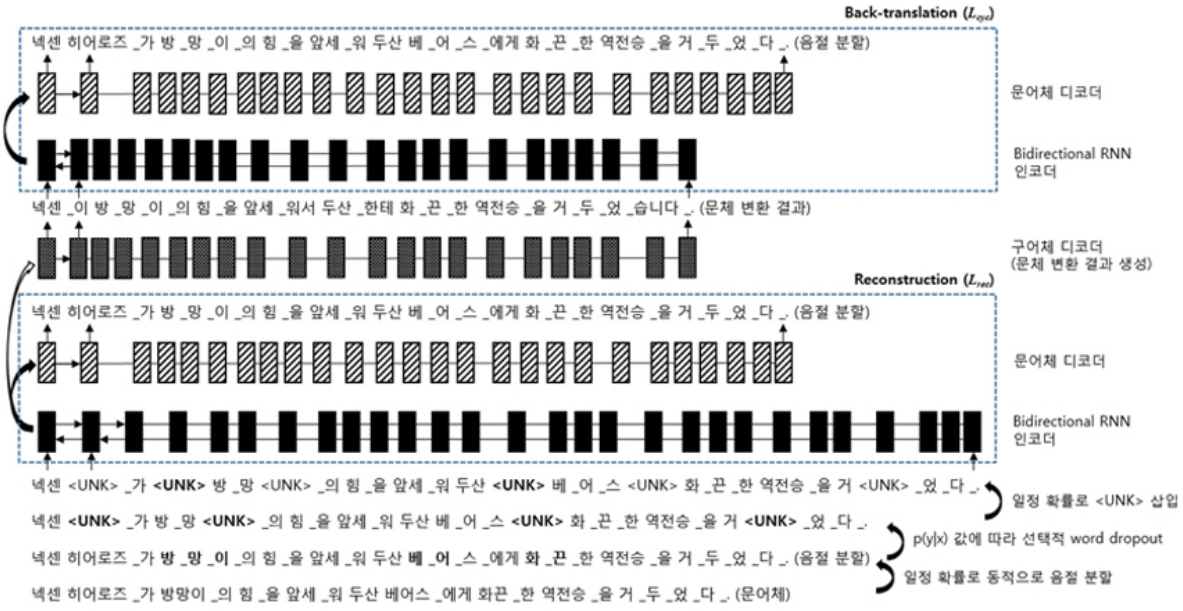
1. 문체 변환 학습데이터 구축 어려움 → Unsupervised Generation (Re-construction + Back Translation)
   즉, 문어체를 구어체로 변환하기 위한 Supervised Learning이 <문어체 문장1, 구어체 문장2>, <문어체 문장2, 구어체 문장2>로 이루어진 대량의 학습데이터가 필요로 된다면, 제안하는 방법은 문어체 문장 왕창, 구어체 문장 왕창 따로따로 있기만 하면 됨.
2. 문체는 바뀌되, 내용이 변해서는 안됨 → Controlled Generation (Attention + Data Nosing)

모델 :



1. Re-construction + Back Translation : 2개의 디코더는 각 하나의 문체를 학습. 예시 : 문어체(즉, 입력 문장(x)의 문체.)를 구어체로 바꾸는 경우. 문어체 디코더는 문어체를 생성하도록, 구어체 디코더는 구어체를 생성하도록 먼저 학습. **이 때 Re-Construction된 x'를 auto encoder loss를 통해 학습시킴.** 그리고 나서, 문체 변환 학습 시에는 구어체 디코더의 생성 결과를 다시 원문 문어체로 변환하게 하여 그 내용이 유지되도록 학습. 이 때 **Back Translation된 x''를 cycle loss를 통해 학습시킴.**
2. Attention + Data Nosing : Attention 기반의 생성 방법은 정보 보존에서 강점을 지니지만 반대로 문체에 따라 변환되어야 할 어절들을 과도하게 보존하는 현상이 나타남 → 이러한 과보존을 방지하기 위해 확률적 word drop-out을 사용. 여기서 확률이라는 것

은 문체의 특성이 뚜렷한 어절들에 대해 높은 빈도로 word dropout하게 함. (아래 그림 참조)

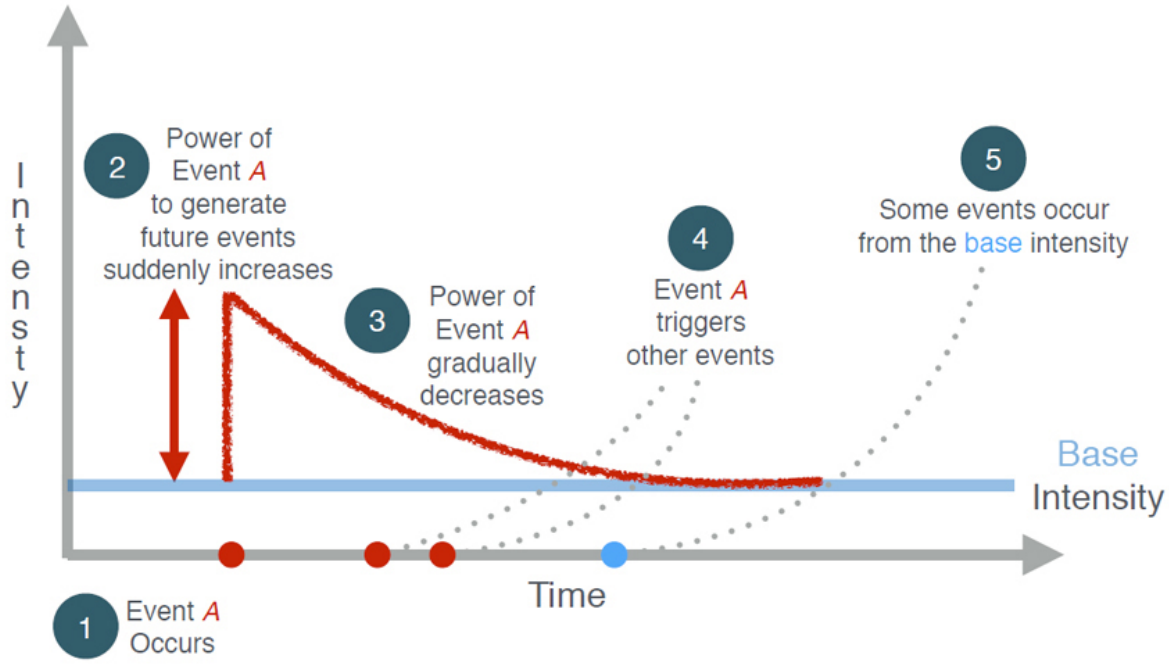


review : Decoder Y에서 어떻게 특정한 문체의 학습하는지, 극단적으로는 입력 문장 x와 똑같이 문장을 생성할 수도 있지 않은지 생각하였는데 각각의 decoder의 vocab을 달리하면 해결됨.

Unsupervised Narrative Learning

정답이 없는 문제에 대해 AI가 수 많은 현상들을 관찰하면서 "...는 이런 것 같아."라고 그럴 듯한 의견을 제시하는 것. 즉, 왜 이러한 일이 일어났으며, 어떤 사건으로부터 시작되었는지, 어떤 사건과 유의미하게 연결되어 있는지를 AI가 분석하고 전달할 수 있도록 함. 이때, 이러한 일련의 연관된 시퀀스를 Narrative라고 함

방안 : 이러한 일련의 인과관계를 모델링하기 위하여 Poisson Point Precoess의 일종인 Hawkes Process[8, 9]를 확장하여 VHP(Vectorized Hawkes Process)을 개발



위에 그림은 하나의 Narrative 하에 있는 사건들의 인과관계를 모델링하기 위한 이 Hawkes Process를 도식화한 것. 이 때 사건들의 인과관계를 모델링하기 위해서는, 사건의 vector 표현을 만들어 Hawkes process에 적용해야 함.
이때 뉴스기사(문서)의 토픽모델링을 통해 사건/문서의 벡터표현을 만듦.

참조

- https://blog.ncsoft.com/커뮤니케이션과-ai-6-자연어처리-extreme-setting/