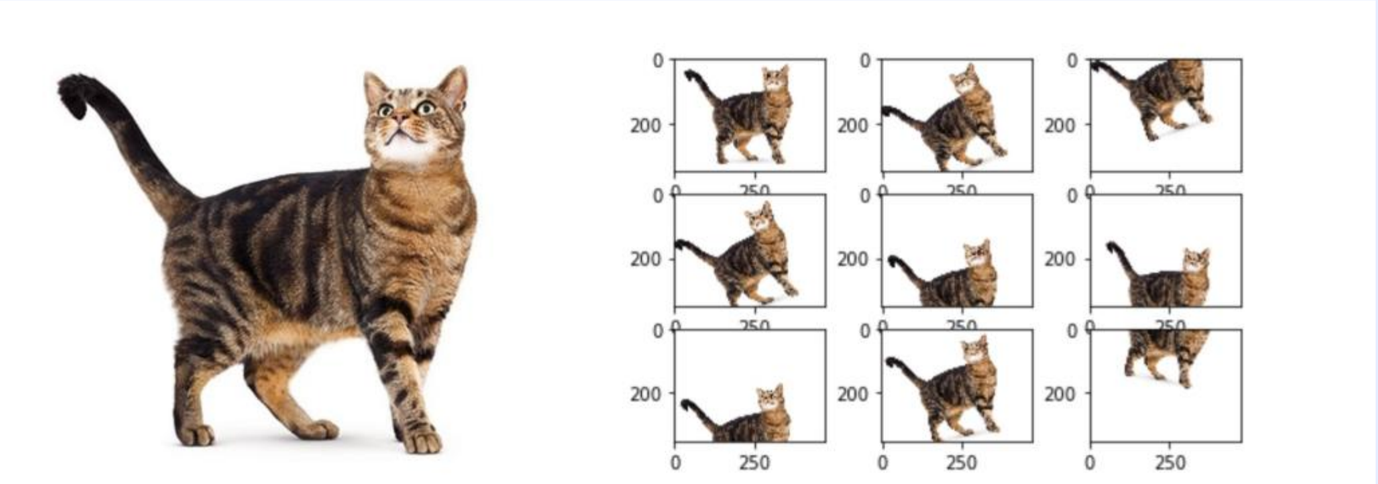
text Data Augmentation

텍스트 데이터 증강 기법

Text Data Augmentation



이미지 데이터 증강을 위해 기존 데이터에 회전을 추가하고 pixel 단위의 변화를 주듯이

텍스트 데이터에도 변형을 줌으로써 새로운 데이터를 만들어냄

텍스트 데이터 증강 기법 소개

Lexical Substitution (동의어 치환)

제가 우울감을 느낀지는 오래됐는데
점점 개선되고 있다고 느껴요

제가 우울감을 느낀지는 오래됐는데
점차 개선되고 있다고 느껴요

불용어가 아닌 단어를 랜덤하게 선택해
wordnet을 기반으로 동의어를 찾아 치환
(**워드임베딩** 혹은 **masked 언어 모델**, **tf-idf**를 사용하는 방법도 존재)

Back Translation(역번역)

ko

제가 우울감을 느낀지는 오래됐는데 점점 개선되고 있다고 느껴요

en

It's been a long time since I've felt depressed, but I feel it's getting better.

ko

우울함을 느낀지 오랜 시간이 지났지만 기분이 좋아지고 있습니다.

번역기를 사용해 다른 언어로 번역 후 다시 기존 언어로 번역해서
새로운 데이터를 얻는 증강 방식

Text Surface Transformation

It's awsome

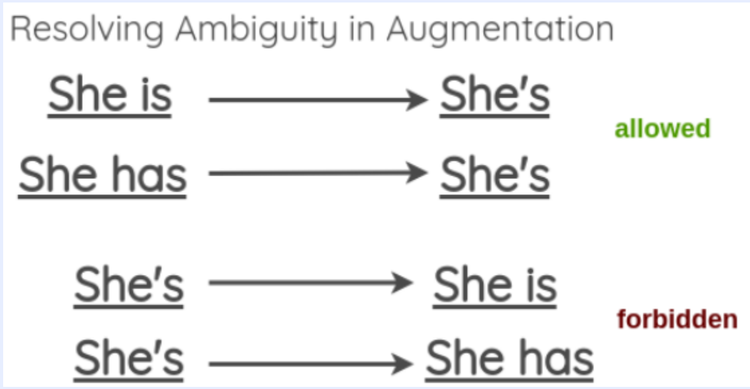
확장

축약

It is awsome

정규식을 이용 contraction과 expansion

모호한 축약은 허용되지만 모호한 확장은 허용되지 않는다.



Random Noise Injection

This is very cool!

Thes is very cool!
This id very cool!
This is bery cool!

The movie was cool. I liked the characters. The length could have be shortened.

I liked the characters. The movie was cool. The length could have be shortened.

모델을 더욱 굳건하게 학습하기 위해 노이즈(오탈자)를 추가

(자주 발생하는 철자 오류 리스트 사용, 쿼티 키보드 노이즈, 문장 순서 변경 등)

RI (Random Insertion)

Random Noise Injection

제가 우울감을 느낀지는 오래됐는데
점점 개선되고 있다고 느껴요

제가 우울감을 점차 느낀지는 오래됐
는데 점점 개선되고 있다고 느껴요

문장 내에서 랜덤하게 선택된 토큰의
동의어를 랜덤한 위치에 추가

RS (Random Swap)

Random Noise Injection

제가 우울감을 느낀지는 오래됐는데
점점 개선되고 있다고 느껴요

제가 우울감을 느낀지는 느껴요
점점 개선되고 있다고 오래됐는데

랜덤하게 선택해 스와핑

RD (Random Deletion)

Random Noise Injection

제가 우울감을 느낀지는 오래됐는데
점점 개선되고 있다고 느껴요

제가 우울감을 느낀지는 점점
개선되고 있다고 느껴요

랜덤하게 선택해 삭제

Instance Crossover Augmentation

아주 재미있는 영화였다.
다음에 꼭 다시 보고싶다.

너무 즐거운 날이었다.
이런 하루가 다시 주어지면
좋겠다.

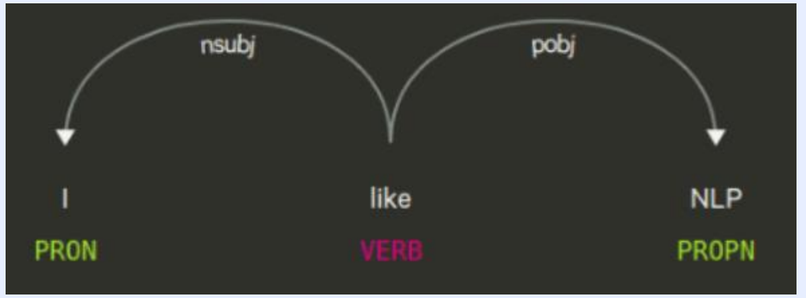
아주 재미있는 영화였다.
이런 하루가 다시 주어지면
좋겠다.

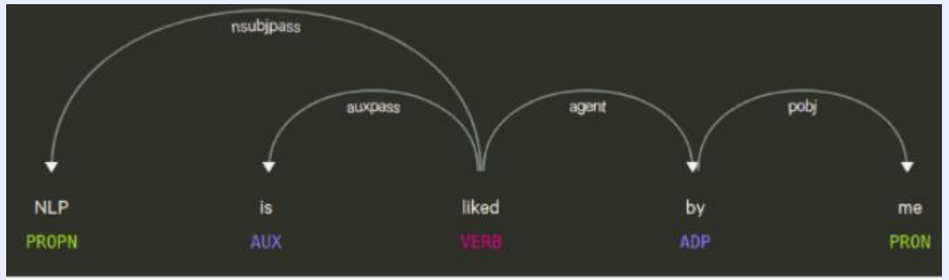
너무 즐거운 날이었다.
다음에 꼭 다시 보고싶다.

감성 분석에서 동일한 감성 라벨을 갖는
데이터를 반으로 나눠 스와핑

Syntax-tree Manipulation

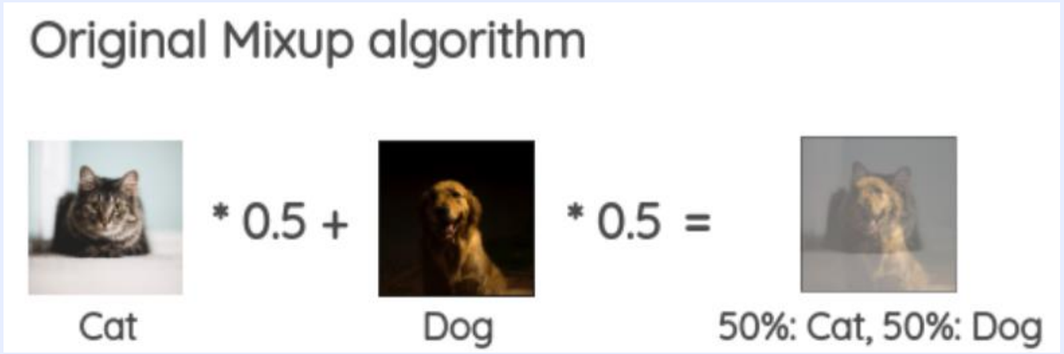
I like NLP → NLP is liked by me

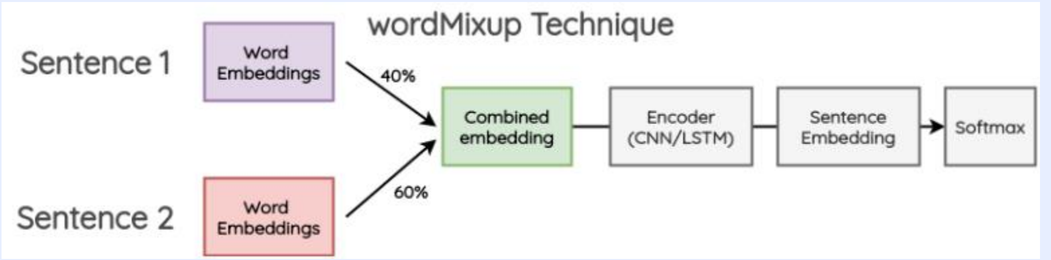


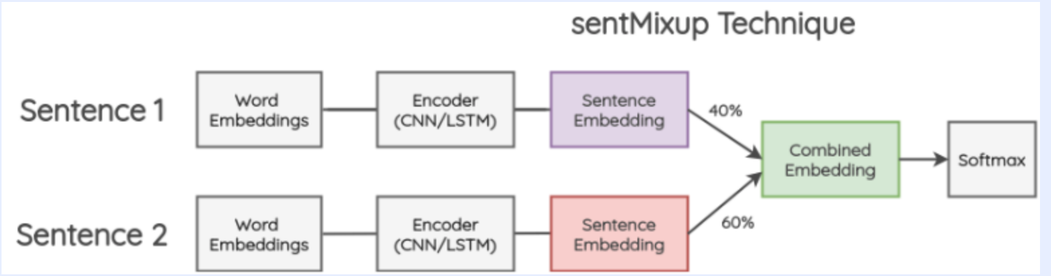


active voice(능동태) <-> passive voice(수동태)

MixUp for Text



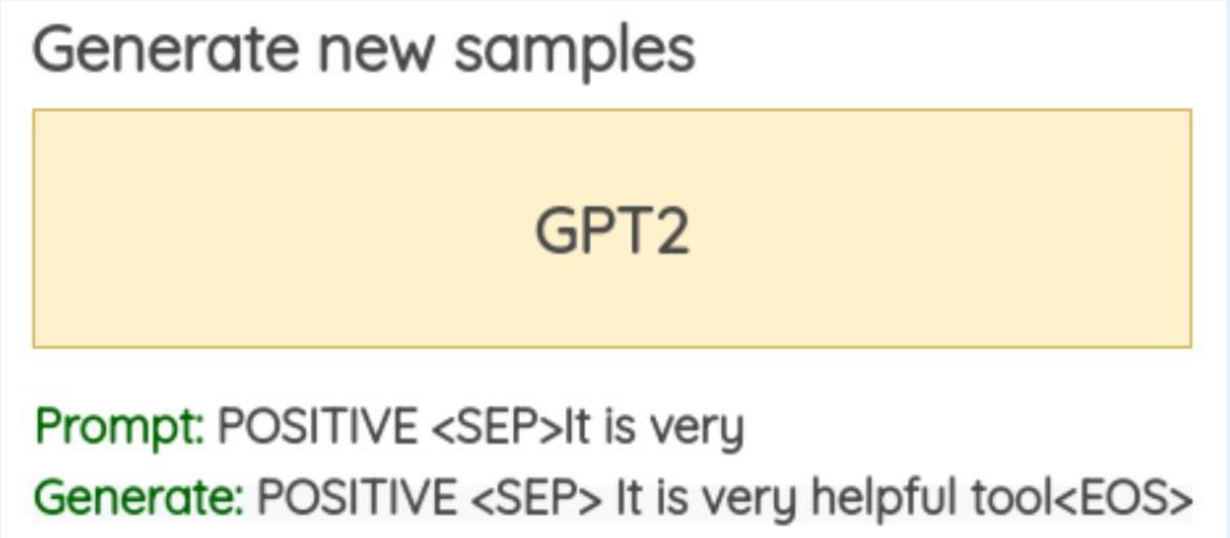




일종의 regularization 역할을 함.

임베딩을 일정 비율로 합쳐 기존 라벨과 비교

Generative Methods



pretrained 모델을 기존 데이터에 fine-tuning 한 후,

initial 단어로 n개의 단어와 라벨을 입력으로 넣어주면

라벨을 보존하며 새로운 데이터 생성

참조

- data augmentation survey: https://amitness.com/2020/05/data-augmentation-for-nlp/
- survey 논문: https://link.springer.com/content/pdf/10.1186/s40537-021-00492-0.pdf
- EDA 논문: https://arxiv.org/pdf/1901.11196.pdf

Thank you