텍스트 분류 삽질기

조용래 / dreamgonfly
2018.10.27 @ 모두의연구소

발표자 소개

조용래 / dreamgonfly

(前) Nexon Korea 인텔리전스랩스
데이터 분석가

(現) Naver Company.AI
머신러닝 엔지니어

Experiences

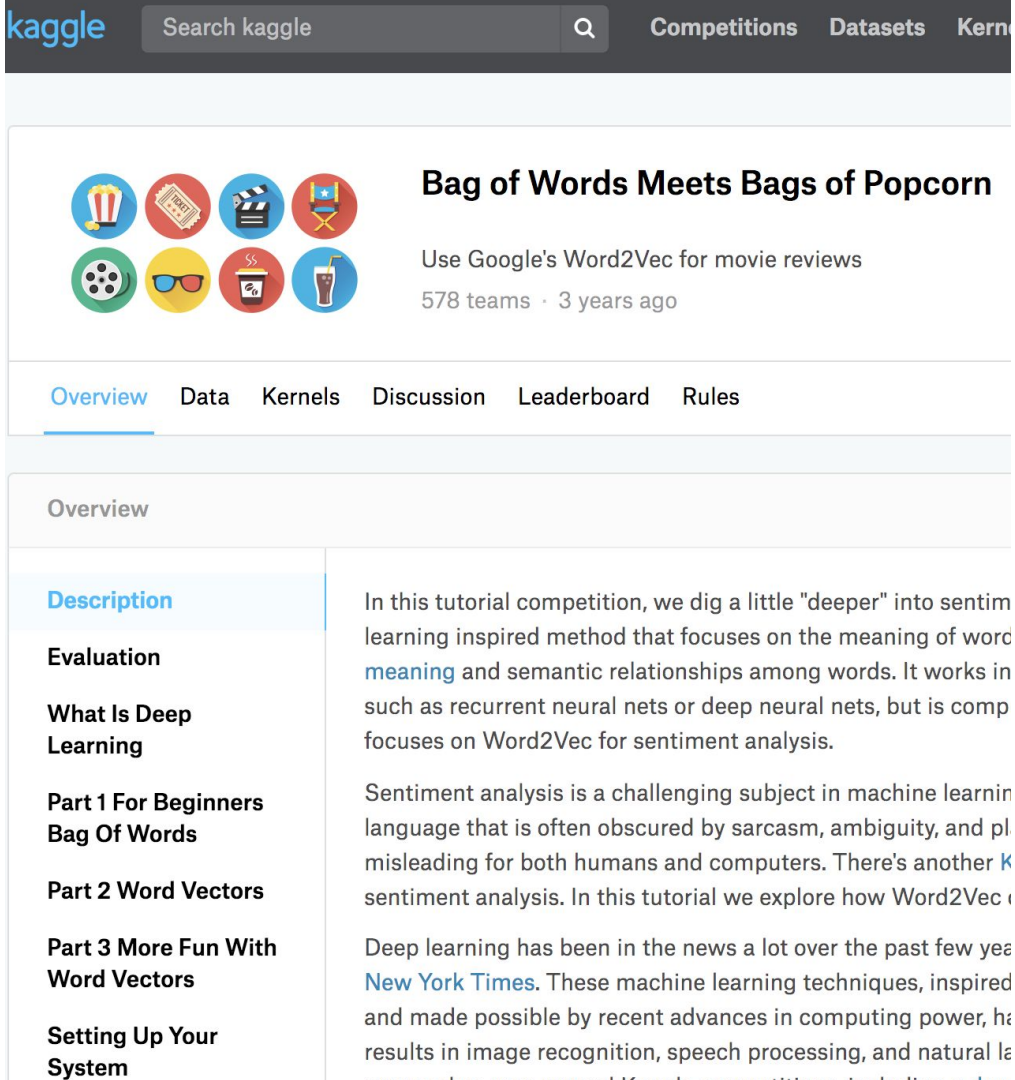
욕설 탐지

https://www.youtube.com/watch?v=K4nU7yXy7R8

감정 분류, 노래 가사 장르 분류 등

어쩌다 보니 텍스트 분류의 노예...

텍스트 분류 : 자연어처리 101



빠르고 성능 좋은 텍스트 분류기

fastText

Library for efficient text classification and representation learning

GET STARTED | DOWNLOAD MODELS

Our first classifier

We are now ready to train our first classifier:

```
>> ./fasttext supervised -input cooking.train -output model_cooking
Read 0M words
Number of words:  14598
Number of labels: 734
Progress: 100.0%  words/sec/thread: 75109  lr: 0.000000  loss: 5.708354  eta: 0h0m
```

텍스트 분류는 이미 끝난 문제 아닌가요?

Challenges of text classification

Multi-class

Positive / negative가 아니라 5개로 나누어야 한다면?
수십가지로 나누어야 한다면?

Multi-label

감정이 여러개가 동시에 나올 수 있다면?

Imbalanced dataset

데이터가 한쪽 클래스로 편중되어 있다면?

Challenges of text classification

Multi-class

Positive / negative가 아니라 5개로 나누어야 한다면?
수십가지로 나누어야 한다면?
Softmax

Multi-label

감정이 여러개가 동시에 나올 수 있다면?
Sigmoid

Imbalanced dataset

데이터가 한쪽 클래스로 편중되어 있다면?
Oversampling or undersampling

Challenges of text classification

모델의 정확도를 높이려면?

- 모델의 구조를 바꿔볼까?
- 다른 문제의 데이터셋을 활용해볼까?
- 레이블이 없는 데이터를 활용해볼까?

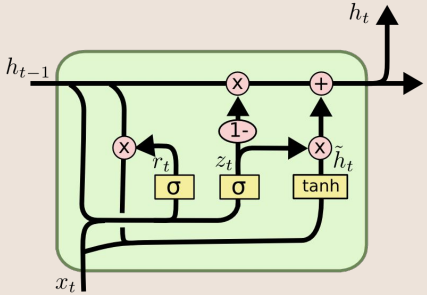
모델이 '왜' 그렇게 예측했는지 설명하고 싶다면?

물론 데이터가 많으면 대부분 풀리는 문제

하지만 레이블링은 시간 & 돈

이 이야기는
부족한 데이터를 극복하려는
삽질기

모델 구조를 바꿔볼까?
좀 더 멋진(?) 아키텍쳐로?



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

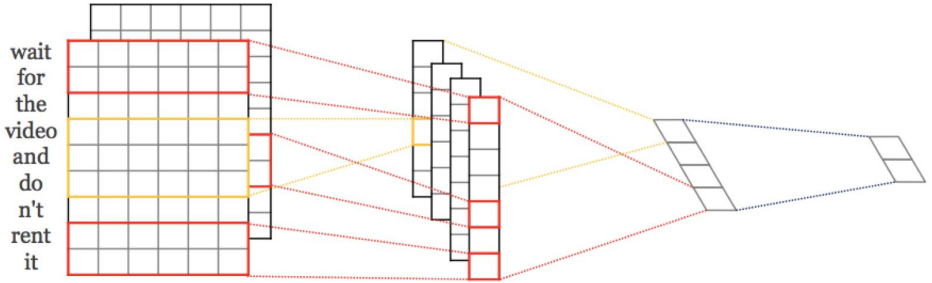
$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

Word CNN

1D CNN for text classification

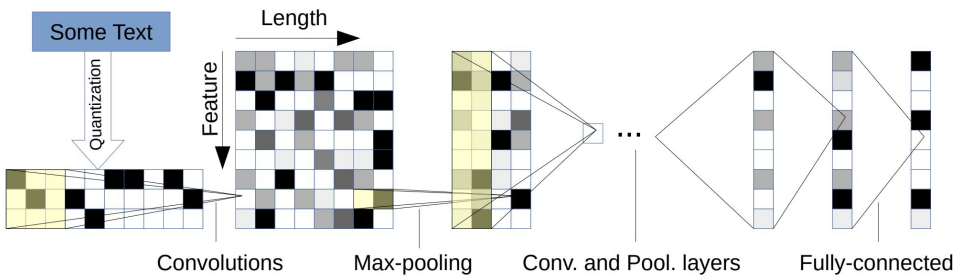
1. 입력 2. Embedding 3. Convolution 4. Pooling 5. 출력 레이어



https://arxiv.org/abs/1408.5882

Character-level Convolutional Networks for Text Classification

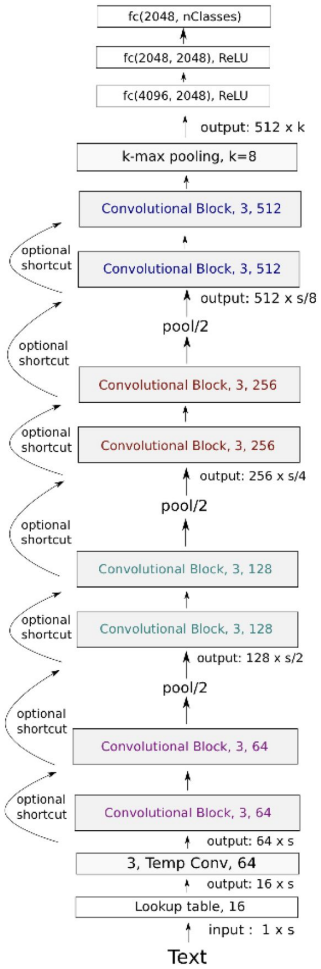
단어가 아니라 character 단위로 해보자



https://arxiv.org/abs/1509.01626

VDCNN : Very Deep Convolutional Networks for Text Classification

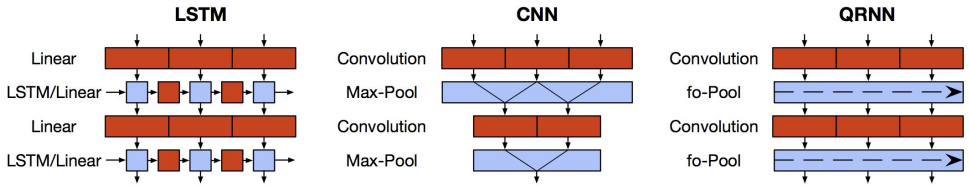
깊게 깊게 쌓아보자



https://arxiv.org/abs/1606.01781

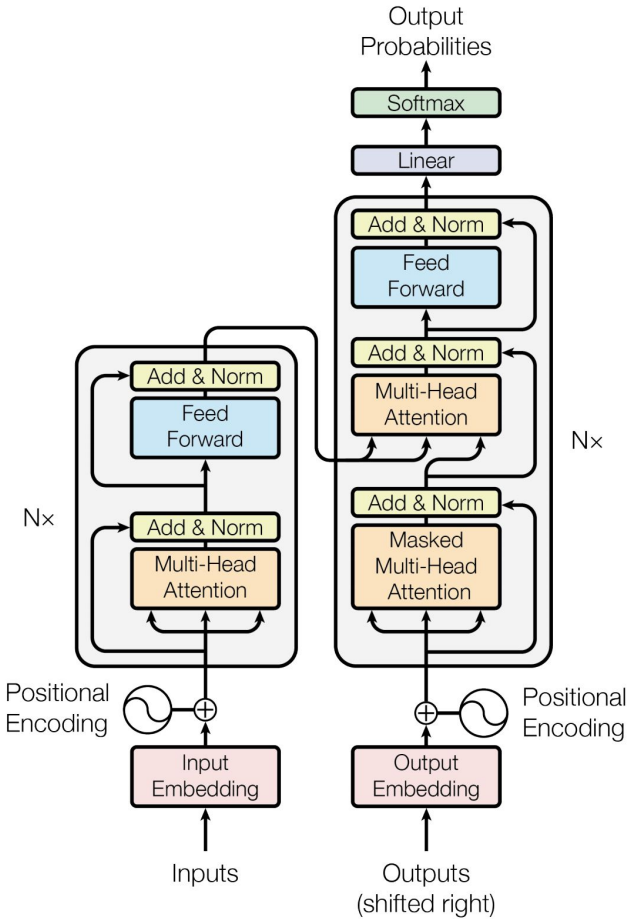
Quasi-Recurrent Neural Networks

RNN을 좀 더 병렬화하자



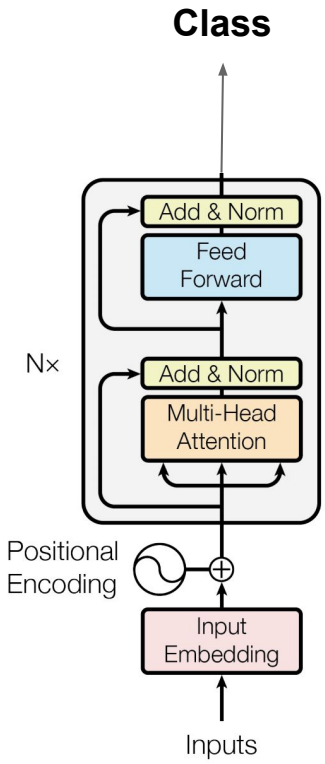
https://arxiv.org/abs/1611.01576

Transformer

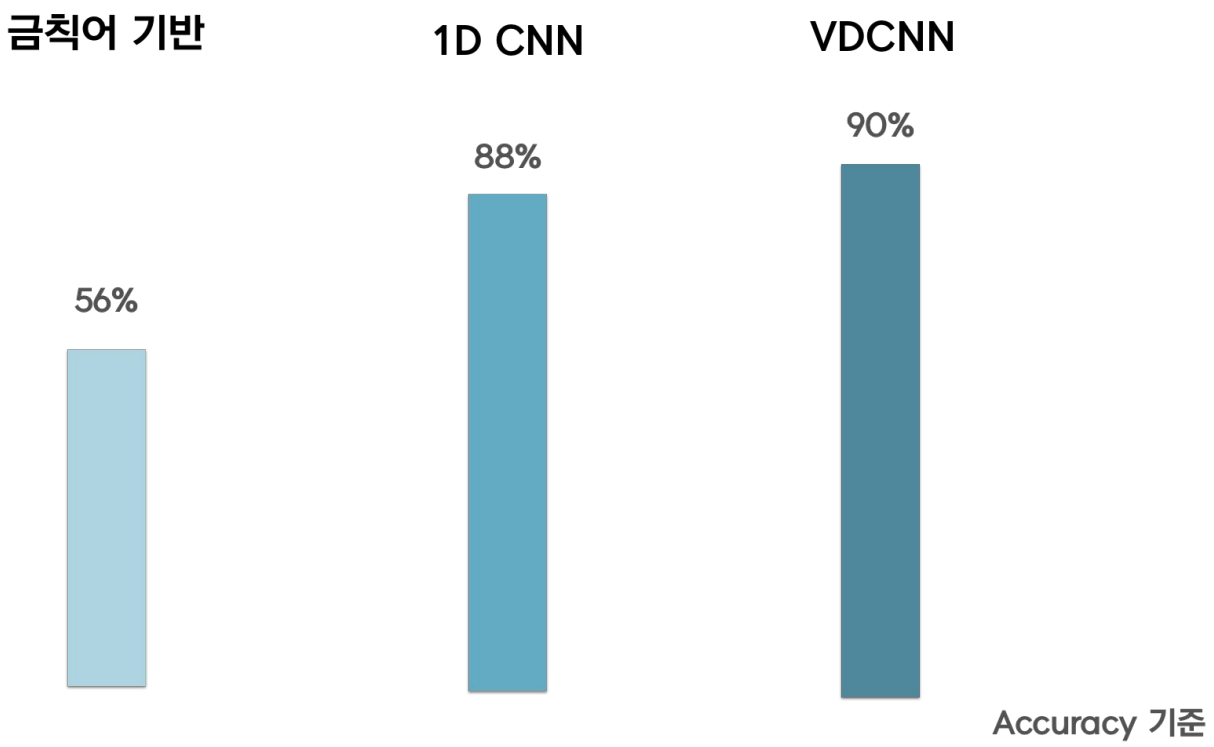


https://arxiv.org/abs/1706.03762

Transformer



욕설 탐지기 예시



삽질 후기

Just use LSTM or Word CNN

QRNN is for training efficiency

VDCNN is strong, but takes too much time to train -> hard to tuning

Transformer encoder has too many parameters for just classification

코드

deep-text-classification-pytorch

https://github.com/dreamgonfly/deep-text-classification-pytorch

Transformer-pytorch

https://github.com/dreamgonfly/deep-text-classification-pytorch

다른 문제의 데이터셋을 활용해볼까?

Multi-task learning

다른 문제의 데이터셋?

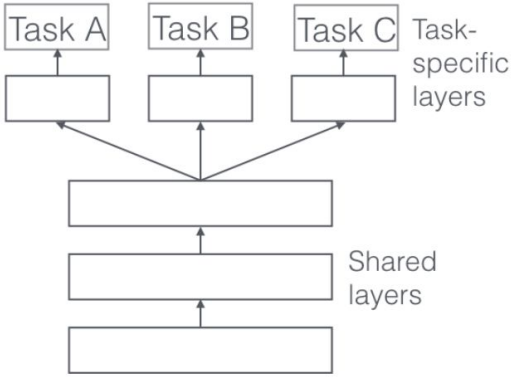
Sentiment analysis

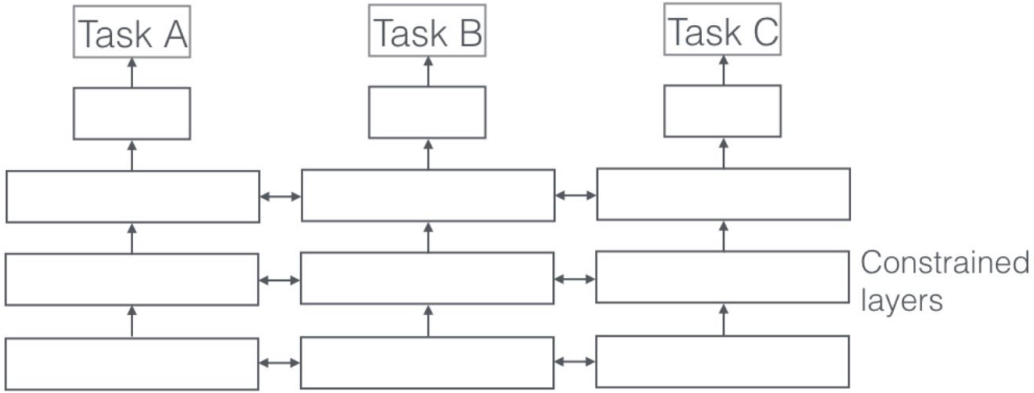
- Movie reviews positive / negative
- Movie reviews 1-10
- Amazon smartphone reviews 1~5점

Lyrics classification

- genres
- Tags

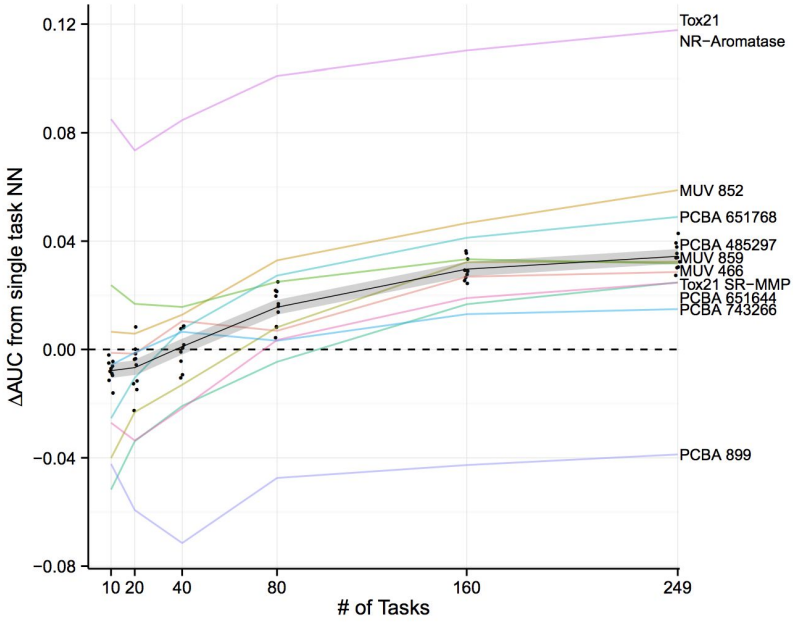
What is multi-task learning?





http://ruder.io/multi-task/index.html

Multi-task learning helps!



https://arxiv.org/abs/1502.02072

삽질 후기

Multi-task learning is powerful!

Information from multiple sources is always useful.

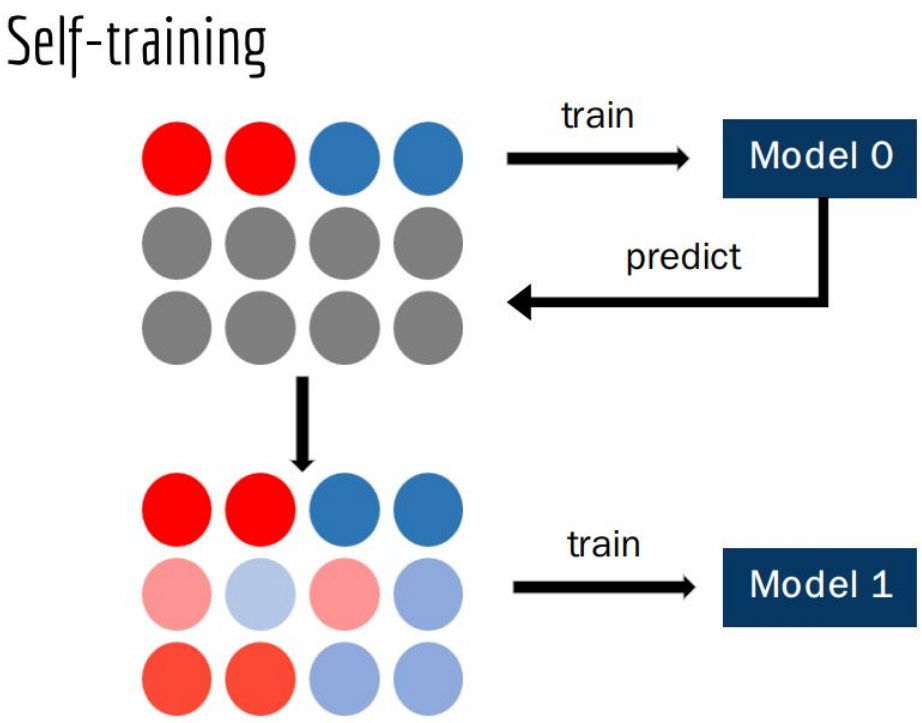
레이블이 없는 데이터를 써먹어볼까?

Semi-supervised learning

Unlabeled data >> labeled data

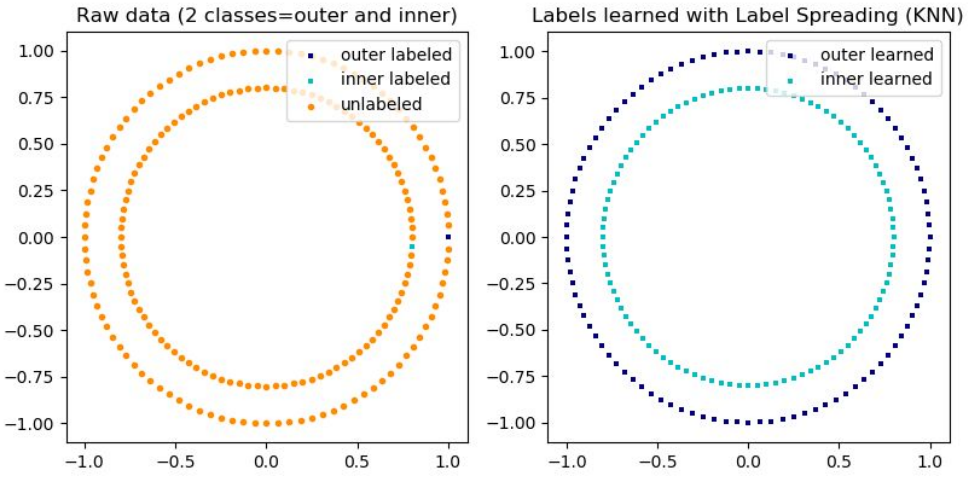
https://www.slideshare.net/lucypark/02-ssl-introduction20091123v11

self-training



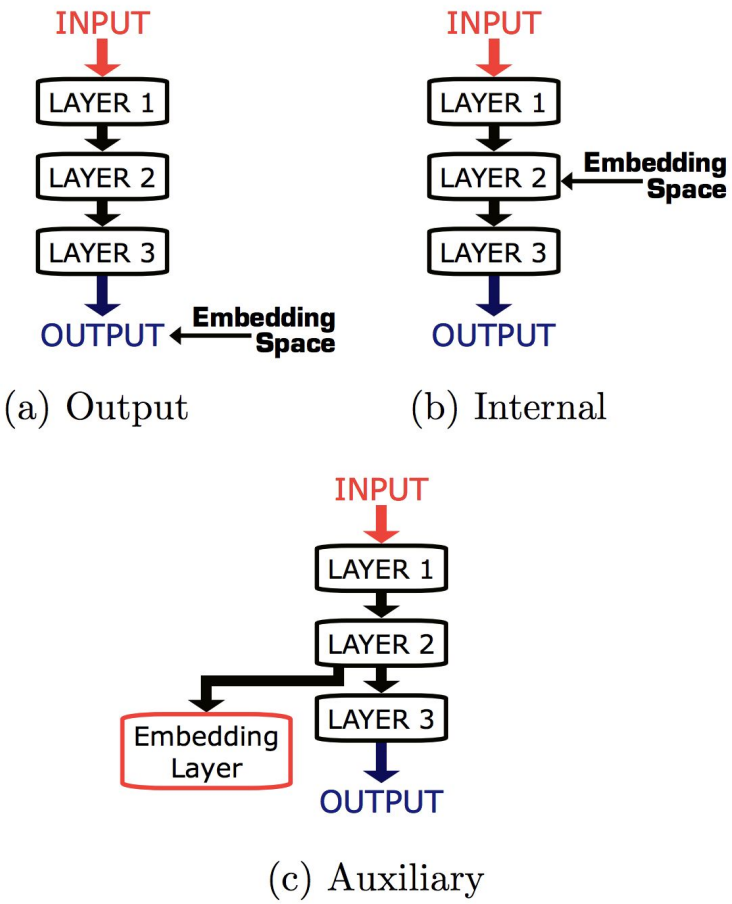
https://sunprinces.github.io/ML-Assignment3/bonus.html

Label propagation



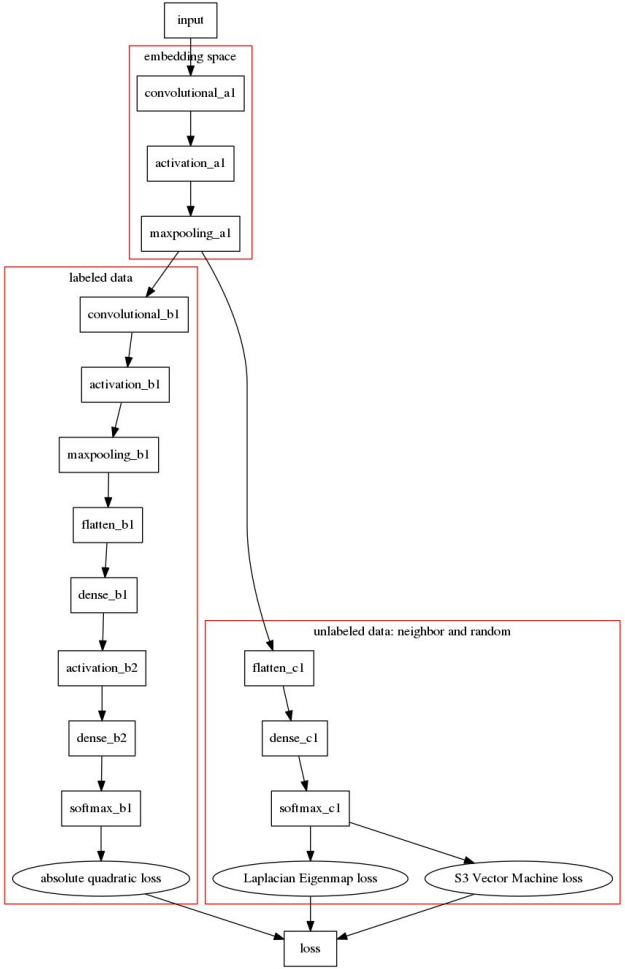
http://scikit-learn.org/stable/modules/label_propagation.html

Semi-supervised embedding (1)



http://www.thespermwhale.com/jaseweston/papers/deep_embed.pdf

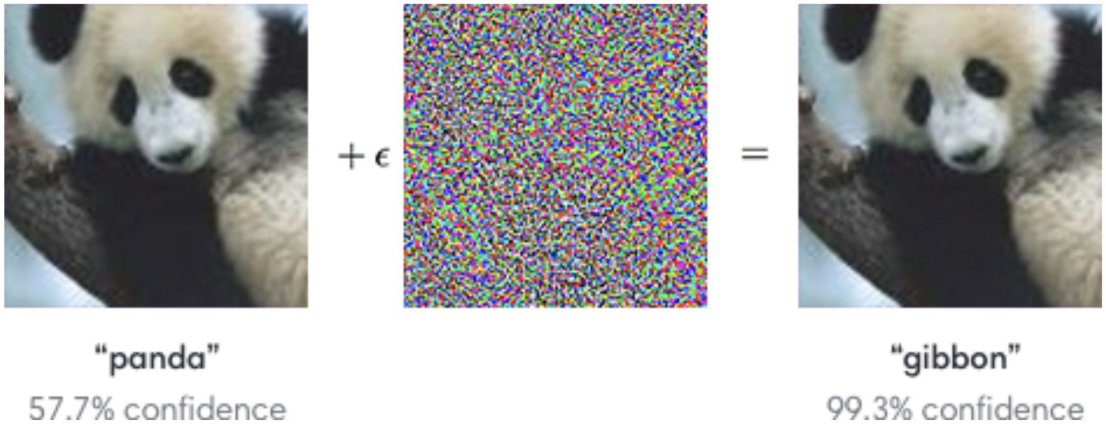
Semi-supervised embedding (2)



https://github.com/yangminz/Semi-supervised_Embedding

Virtual Adversarial training (1)

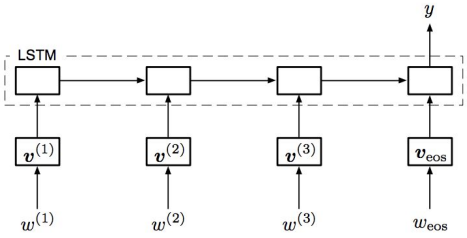
Adversarial example



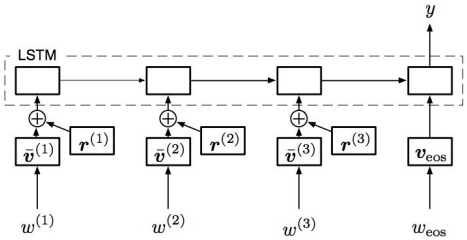
https://blog.openai.com/adversarial-example-research/

Virtual Adversarial training (2)

Adversarial training



(a) LSTM-based text classification model.



(b) The model with perturbed embeddings.

https://arxiv.org/abs/1605.07725

Virtual Adversarial training (3)

Adversarial training

$$\mathrm{KL}[p(\cdot \mid \boldsymbol{x}; \hat{\boldsymbol{\theta}}) || p(\cdot \mid \boldsymbol{x} + \boldsymbol{r}_{\text{v-adv}}; \boldsymbol{\theta})]$$

$$\text{where } \boldsymbol{r}_{\text{v-adv}} = \underset{\boldsymbol{r}, \|\boldsymbol{r}\| \leq \epsilon}{\arg\max}\, \mathrm{KL}[p(\cdot \mid \boldsymbol{x}; \hat{\boldsymbol{\theta}}) || p(\cdot \mid \boldsymbol{x} + \boldsymbol{r}; \hat{\boldsymbol{\theta}})]$$

https://arxiv.org/abs/1605.07725

삽질 후기

Not so effective

Problem structure vs. model assumption

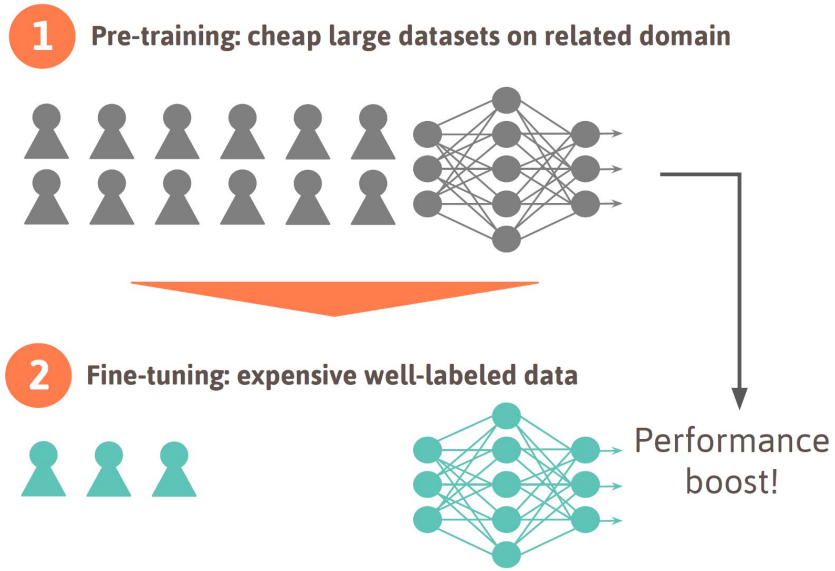
- Error propagation
- ‘좋은’ vs. ‘싫은’ has similar embeddings
- Short distance -> confusion

큰 데이터에 미리 학습시켜 놓은 걸 써먹어볼까?

Transfer learning

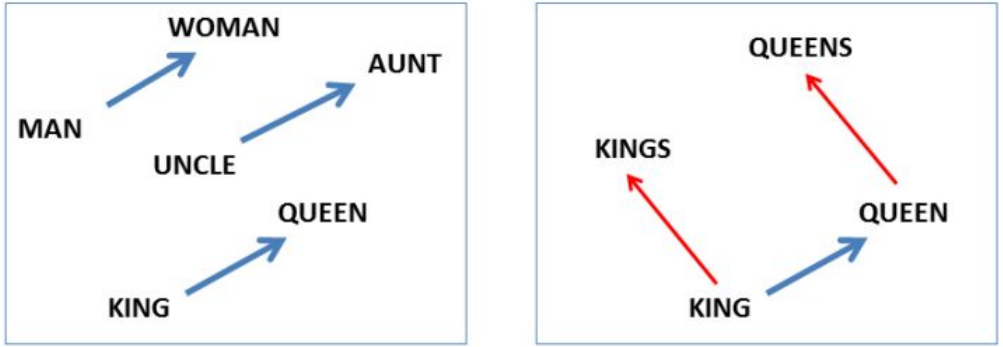
Pretraining

Wikipedia news twitter

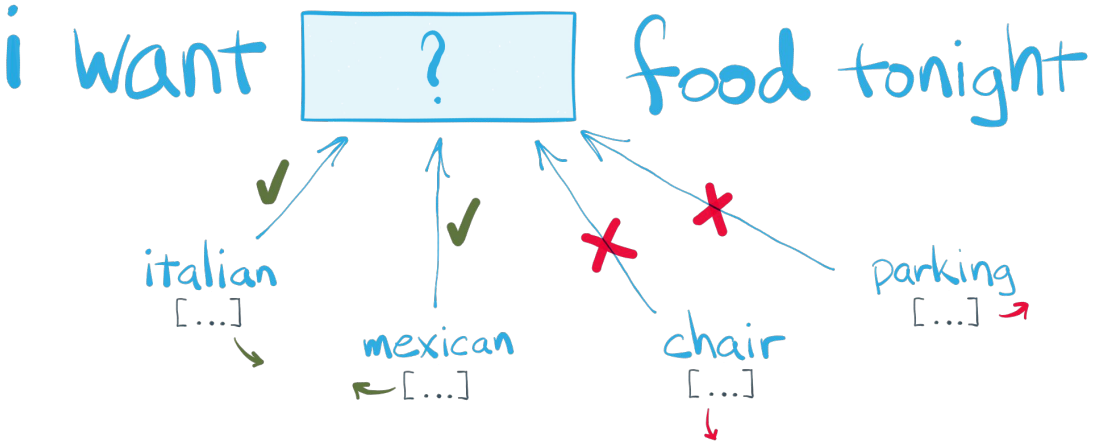


Word embedding

- Word2Vec
- FastText
- Glove
- Swivel



(Mikolov et al., NAACL HLT, 2013)



Word embedding works!

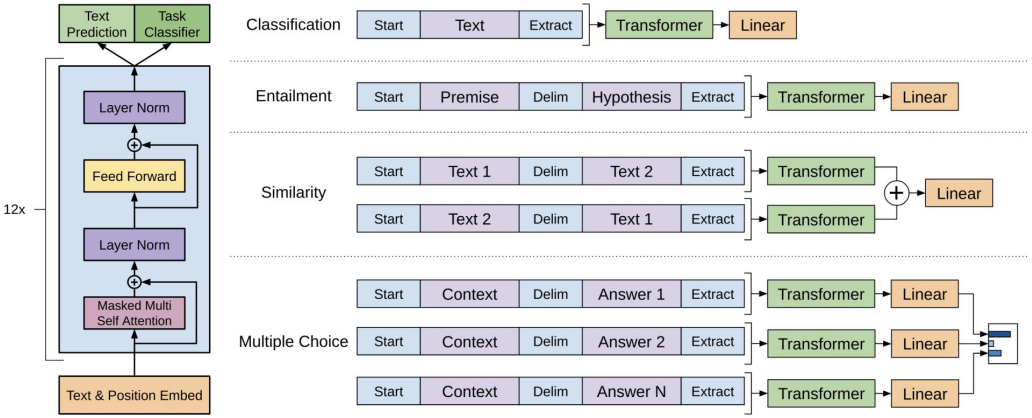
Model	MR	SST-1	SST-2	Subj	TREC	CR	MPQA
CNN-rand	76.1	45.0	82.7	89.6	91.2	79.8	83.4
CNN-static	81.0	45.5	86.8	93.0	92.8	84.7	**89.6**
CNN-non-static	**81.5**	48.0	87.2	93.4	93.6	84.3	89.5
CNN-multichannel	81.1	47.4	**88.1**	93.2	92.2	**85.0**	89.4

https://arxiv.org/abs/1408.5882

Pretrained Language model

ELMo: Deep contextualized word representation

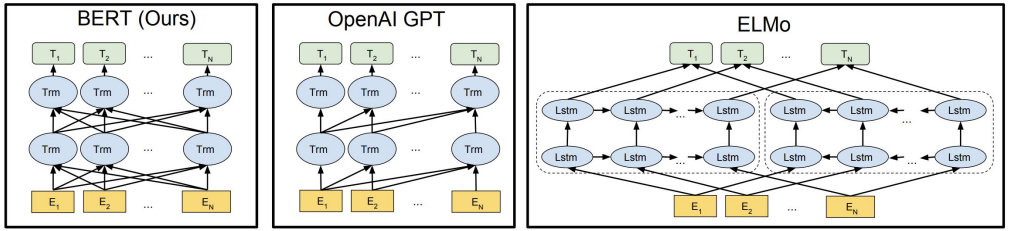
GPT : Generative Pre-Training



https://blog.openai.com/language-unsupervised/

https://arxiv.org/abs/1802.05365

BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding

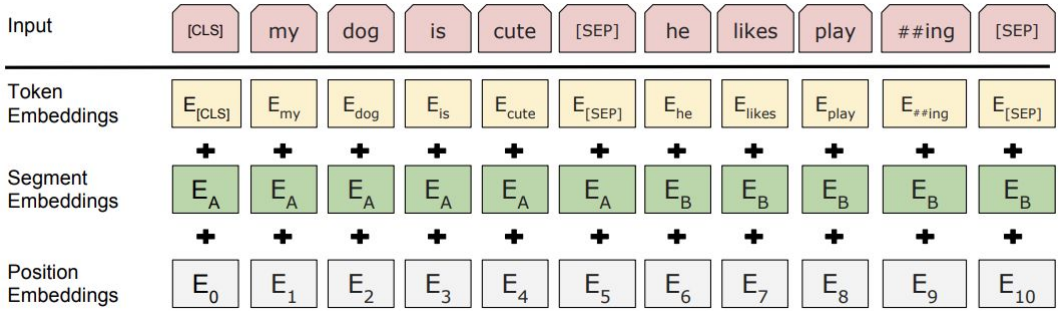


https://arxiv.org/abs/1810.04805

BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding

MLM : Masked Language Model

NSP : Next Sentence Prediction



https://arxiv.org/abs/1810.04805

논문 구현

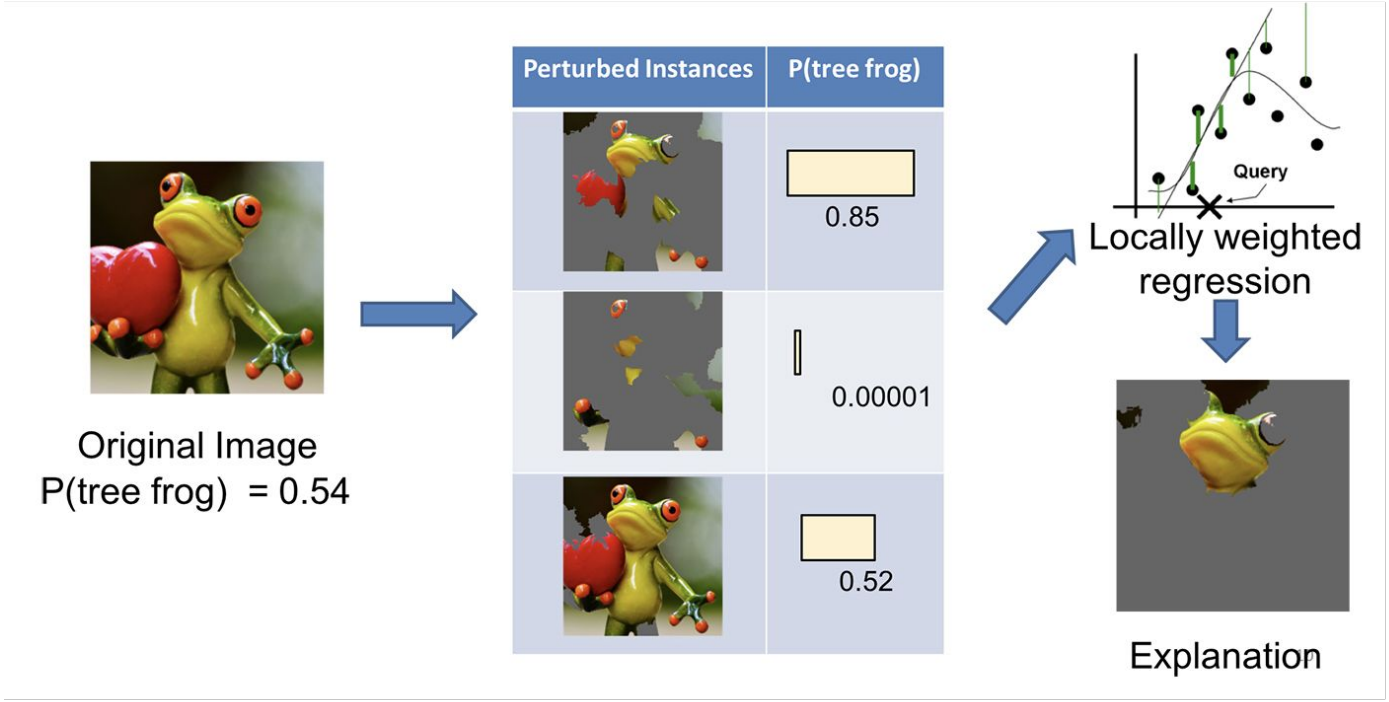
BERT-pytorch

https://github.com/dreamgonfly/BERT-pytorch

모델이 ‘왜' 그렇게 예측했는지 설명하기

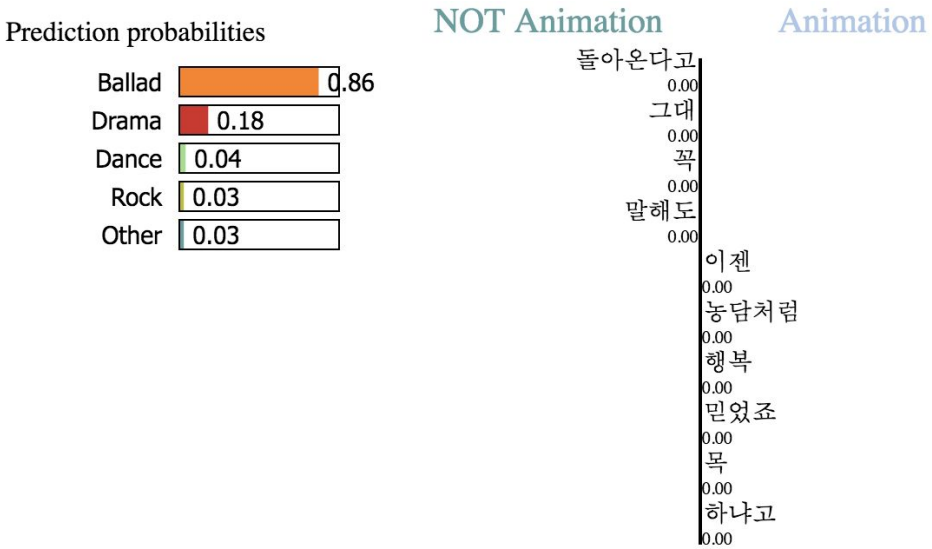
LIME

Local Interpretable Model-agnostic Explanations



Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin.
"Why should i trust you?: Explaining the predictions of any classifier."

발라드는 왜 발라드일까

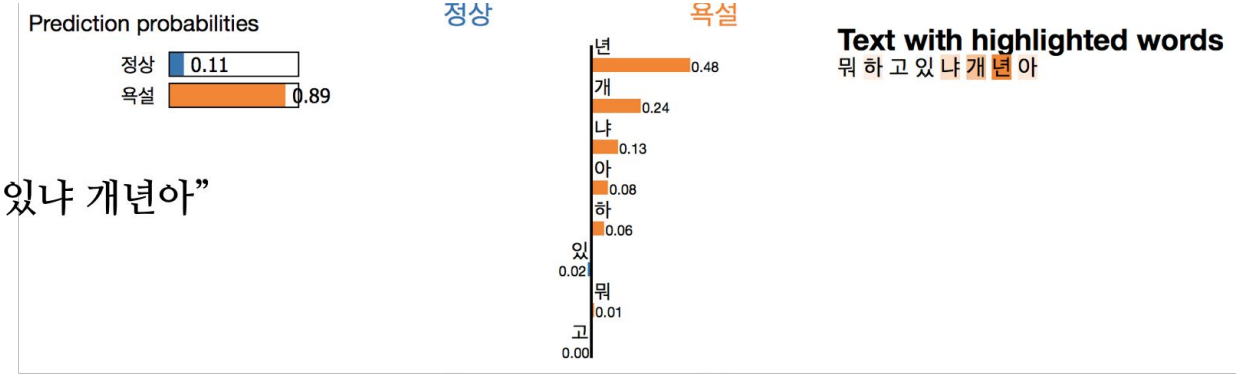


Text with highlighted words

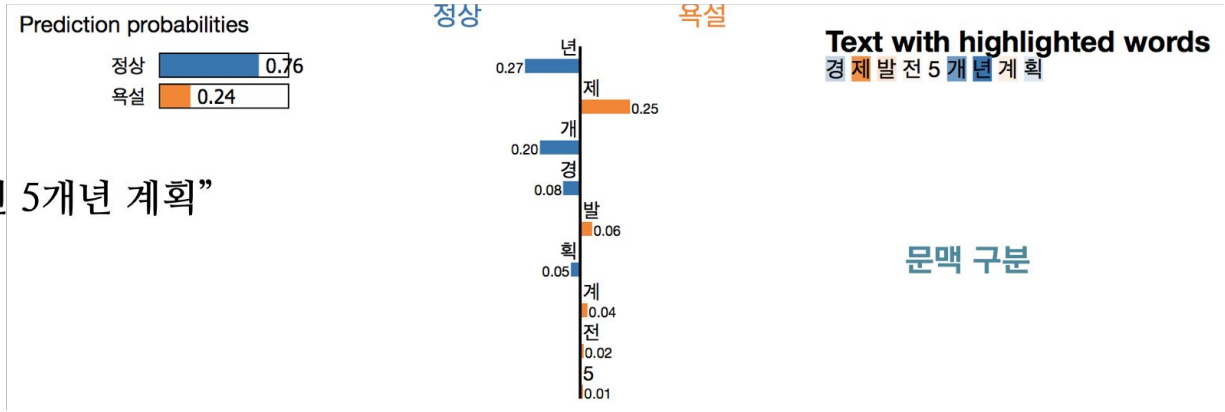
그렇죠 내가 바보였어요 그렇게 그댈 많이 울렸단 걸 까마득히 잊어버리고 그대 내게 다시 온다고 믿었죠 꼭 돌아온다고 맞아요 나는 못된 남자죠 이제와 농담처럼 그댄 말해도 알아요 그대 내 곁에서 혼자 흘린 눈물 이젠 다 내 몫이 됐는 걸 차라리 다시 울어줘요 나를 붙잡고 밀고 때리고 예전처럼 내게 안겨요 어쩌죠 여전히도 나는 못됐나 봐요 그대는 웃고 있는데 다 잊고 행복하단 그대 앞에서 자꾸 눈물이 나요 나 어쩌죠 이제야 그대 사랑에 겨우 난 눈을 떴는데 모르죠 끝내 모르겠죠 그대 떠나고 하루 한 번도 웃어본적 없던 나란 걸 어쩌죠 여전히도 나는 못됐나 봐요 그대는 웃고 있는데 다 잊고 행복하단 그대 앞에서 자꾸 눈물이 나요 나 어쩌죠 이제야 그대 사랑에 겨우 난 눈을 떴는데 아니라고 말해요 나를 못 잊겠다고 너무 그리웠다고 내가 없는 하루가 마치 일년 같아

욕설은 왜 욕설일까

“뭐하고 있냐 개년아”



“경제 발전 5개년 계획”



문맥 구분

남은 문제

모르는 걸 모른다고 하기

Article: Endangered Species Act
Paragraph: "*. . . Other legislation followed, including the Migratory Bird Conservation Act of 1929, a 1937 treaty prohibiting the hunting of right and gray whales, and the Bald Eagle Protection Act of 1940. These later laws had a low cost to society—the species were relatively rare—and little opposition was raised.*"

Question 1: "*Which laws faced significant opposition?*"
Plausible Answer: *later laws*

Question 2: "*What was the name of the 1937 treaty?*"
Plausible Answer: *Bald Eagle Protection Act*

https://arxiv.org/abs/1806.03822

Conclusion : Text Classification Recipe

데이터를 충분히 수집하기

LSTM / Word CNN

Multi-task learning

Pretrained word embeddings

Pretrained language models

Explanations and visualization

Conclusion

Not just data

But how to extract information from data!