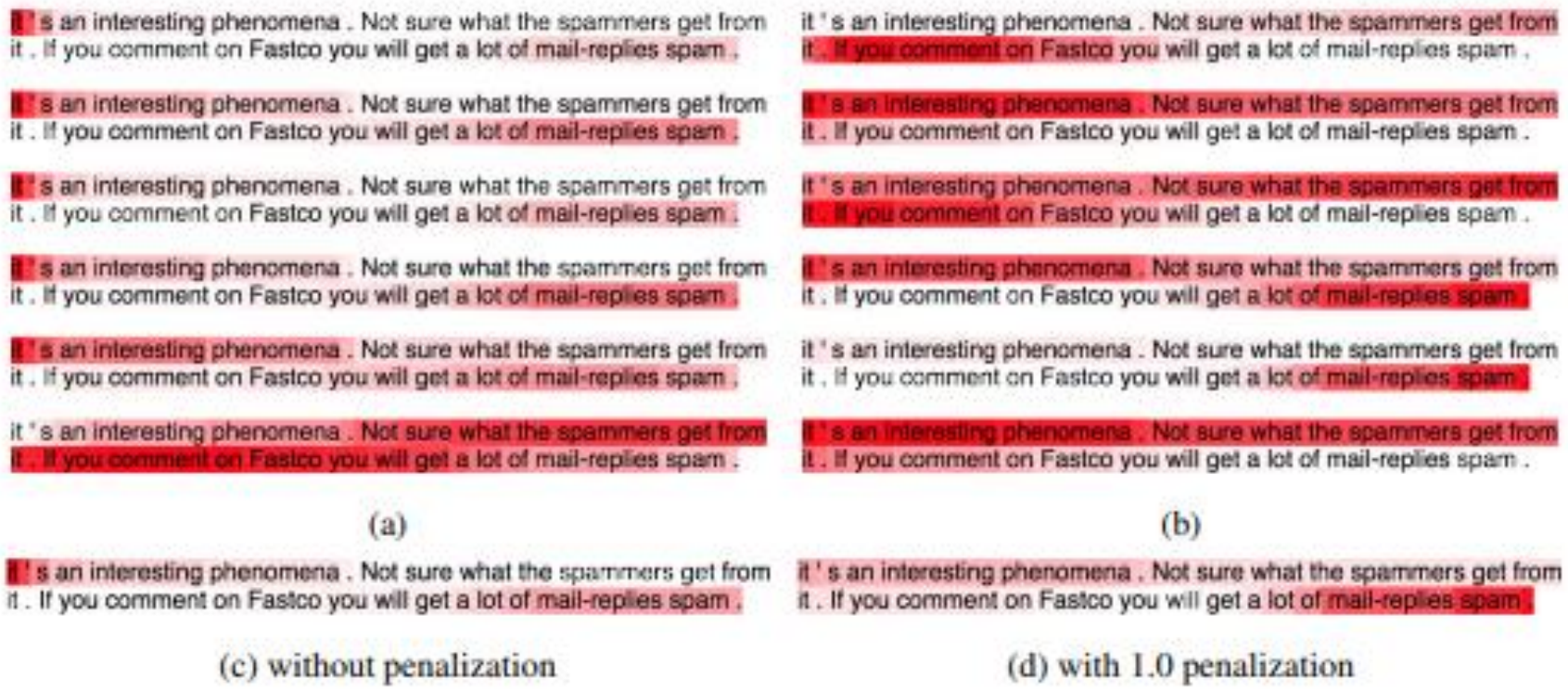
A STRUCTURED SELF-ATTENTIVE SENTENCE EMBEDDING

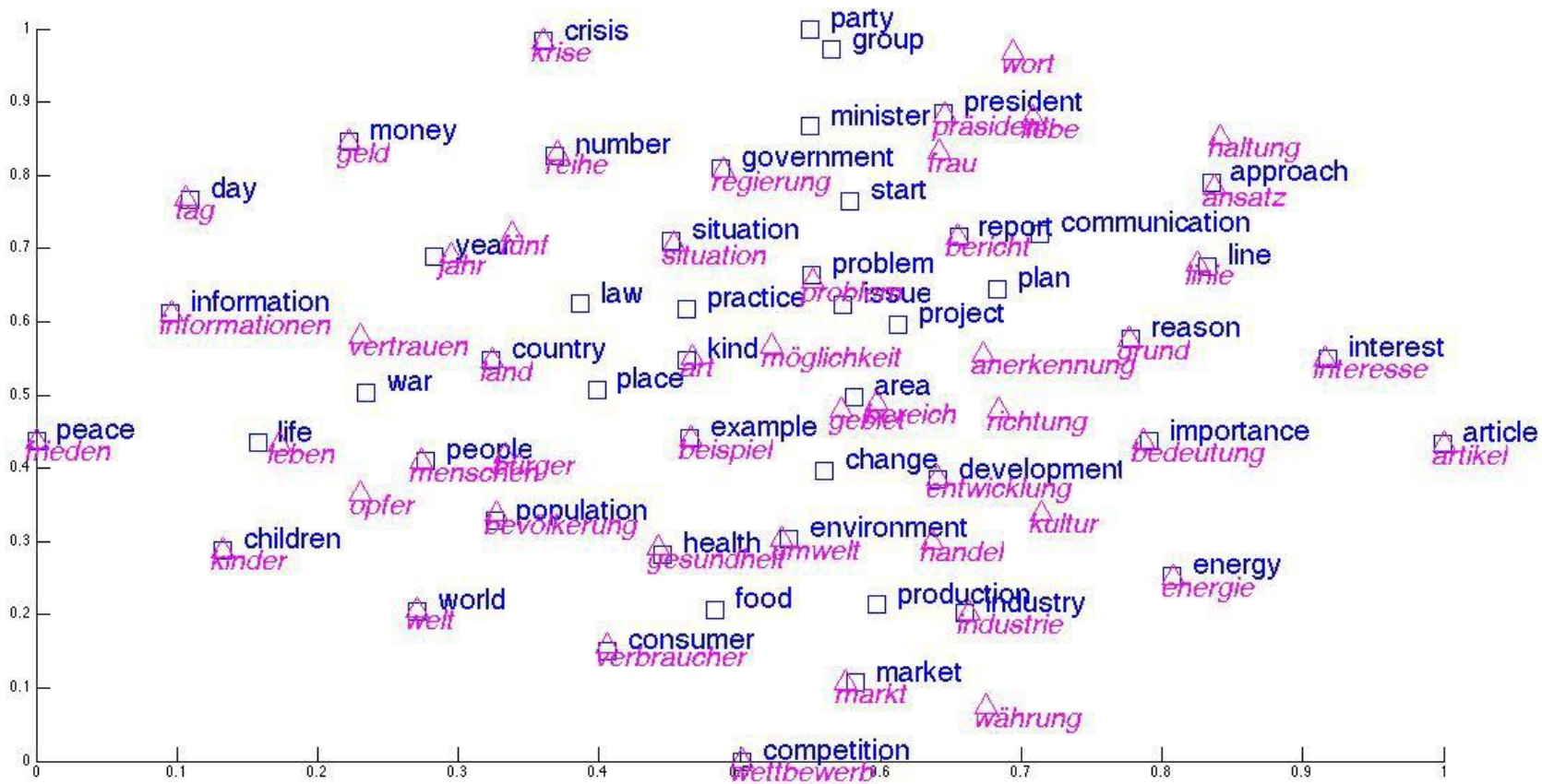
한양대학교 AILAB 석사과정 엄희송

ABSTRACT

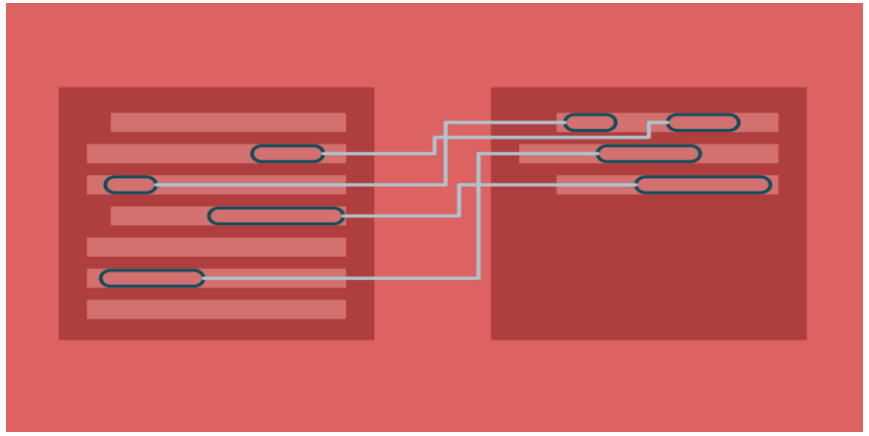
Self-attention을 도입해 sentence 단위의 embedding을 효과적으로 할 수 있도록 함.
이를 위해 기존의 vector를 이용하는 방식에서, 각 row가 문장의 다른 부분들을 attending하는 2-D matrix를 사용해 embedding을 표현함.



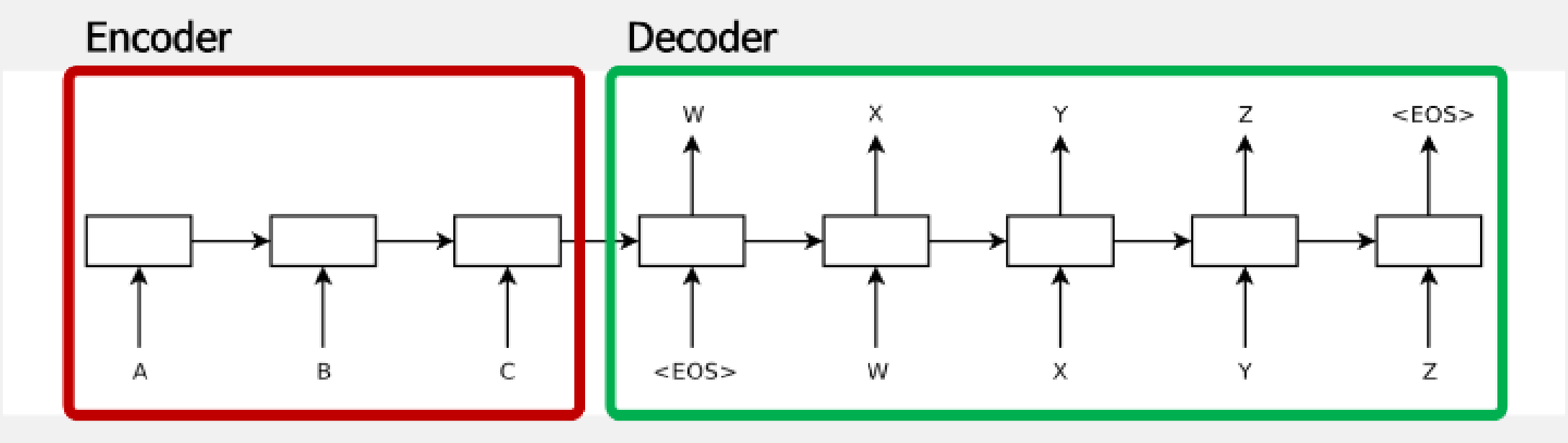
Word embedding을 통해 semantically meaningful distributed representations of words를 학습할 수 있음.



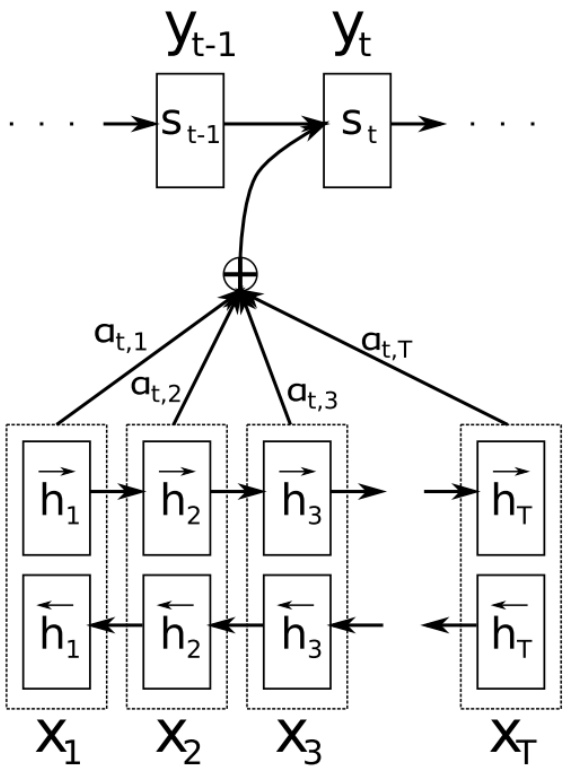
Phrases and sentences를 표현하는 방법은 연구해야 할 부분.
Unsupervised learning ➔ universal sentence embeddings (recursive AE, sequential denoising AE, …)
Supervised learning ➔ specifically trained sentence embeddings (RNNs, CNNs, …)



문장을 embedding하기 위해 기존에 사용했던 방법으로는 문장을 표현하는 간단한 벡터를 만드는 것.
문장이 RNN을 통과한 마지막 hidden state, or RNN hidden state들의 max(또는 average) pooling, or convolved n-grams를 사용.



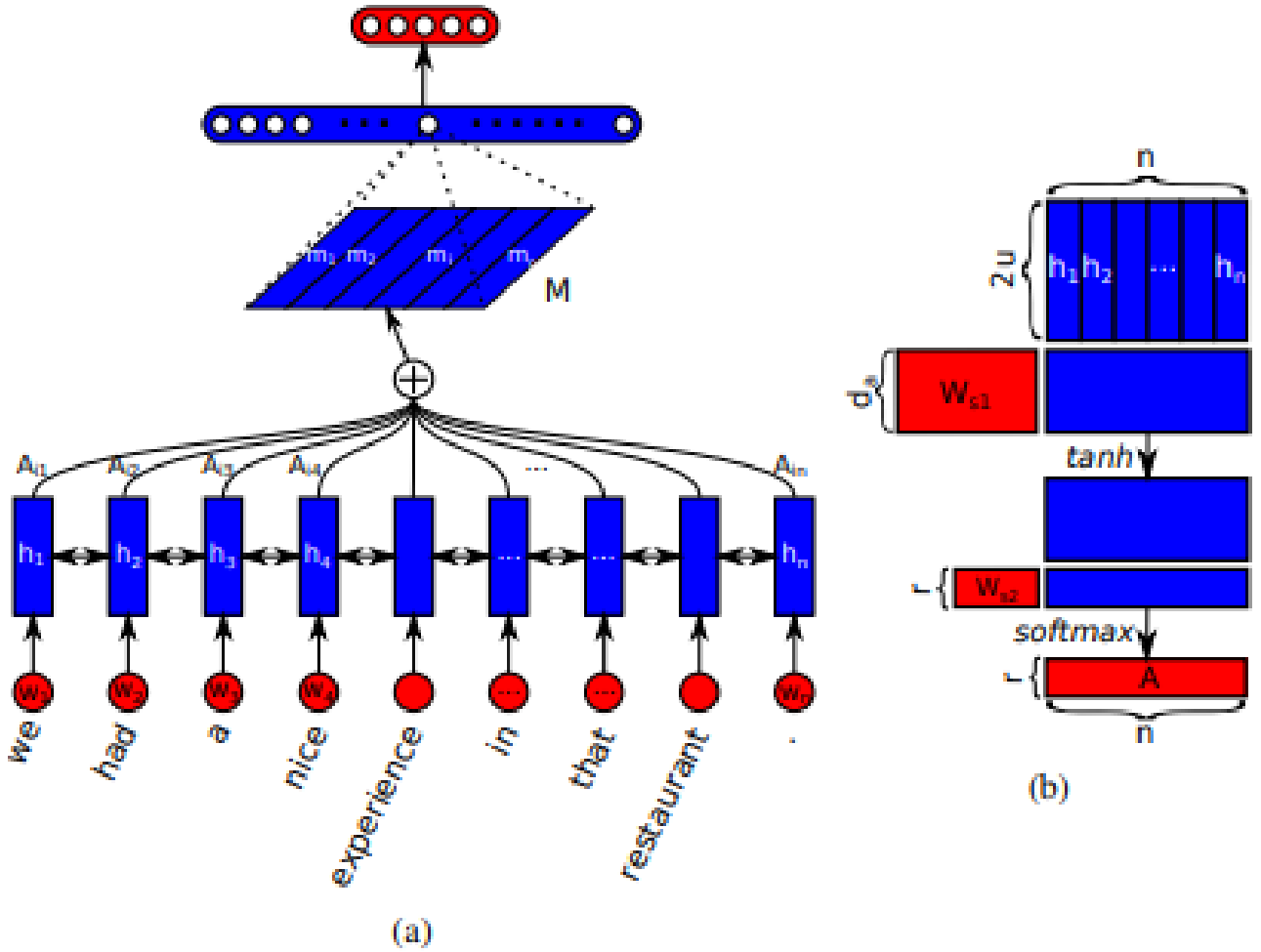
Attention mechanism을 CNN이나 LSTM 모델에 사용해서 sentence embedding을 추출하는 것에 도움을 줄 수 있는 extra source of information을 제공하는 것을 제안함.



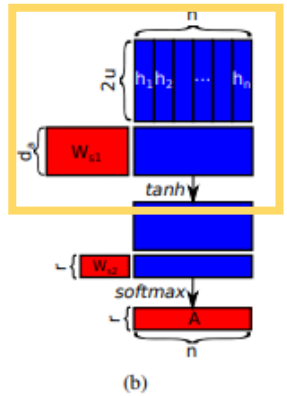
sentiment classification과 같은 몇몇 task에서는 attention을 적용하기 위해
문장이 RNN을 통과한 마지막 hidden state, or RNN hidden state들의 Max(또는 average) pooling을 사용함.

이 논문에서는 extra inputs 없이 하나의 sentence를 multiple vector representation으로 만드는 self-attention을 제안함. 이를 통해 pooling을 대체함. (아주 긴 문장의 경우 pooling만으로는 정보를 갖기 어려운 부분)

Approach
기존 Attention과 구성은 같음.



Approach
기존 Attention과 구성은 같음.



(b)

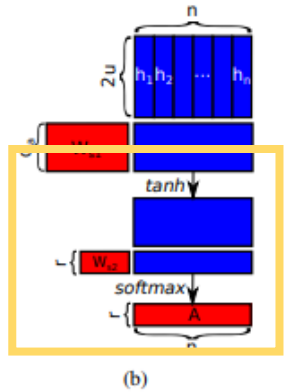
T
E
X
T

n x d Embedding matrix

n x 2u Hidden state matrix

tanh(d_a x 2u Weight matrix 2u x n (Hidden state matrix)^T)

Approach
기존 Attention과 구성은 같음.

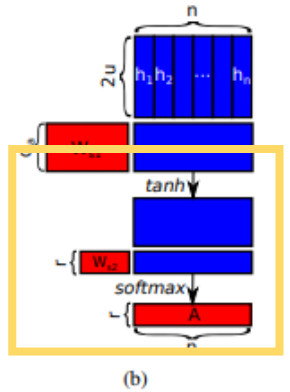


1xd_a weight tanh(d_a x 2u Weight matrix 2u x n (Hidden state matrix)^T)

Softmax

⇨ 1 x n annotation vector

Approach
기존 Attention과 구성은 같음.



$r \times d_a$ weight matrix tanh($d_a \times 2u$ Weight matrix 2u x n (Hidden state matrix)^T)

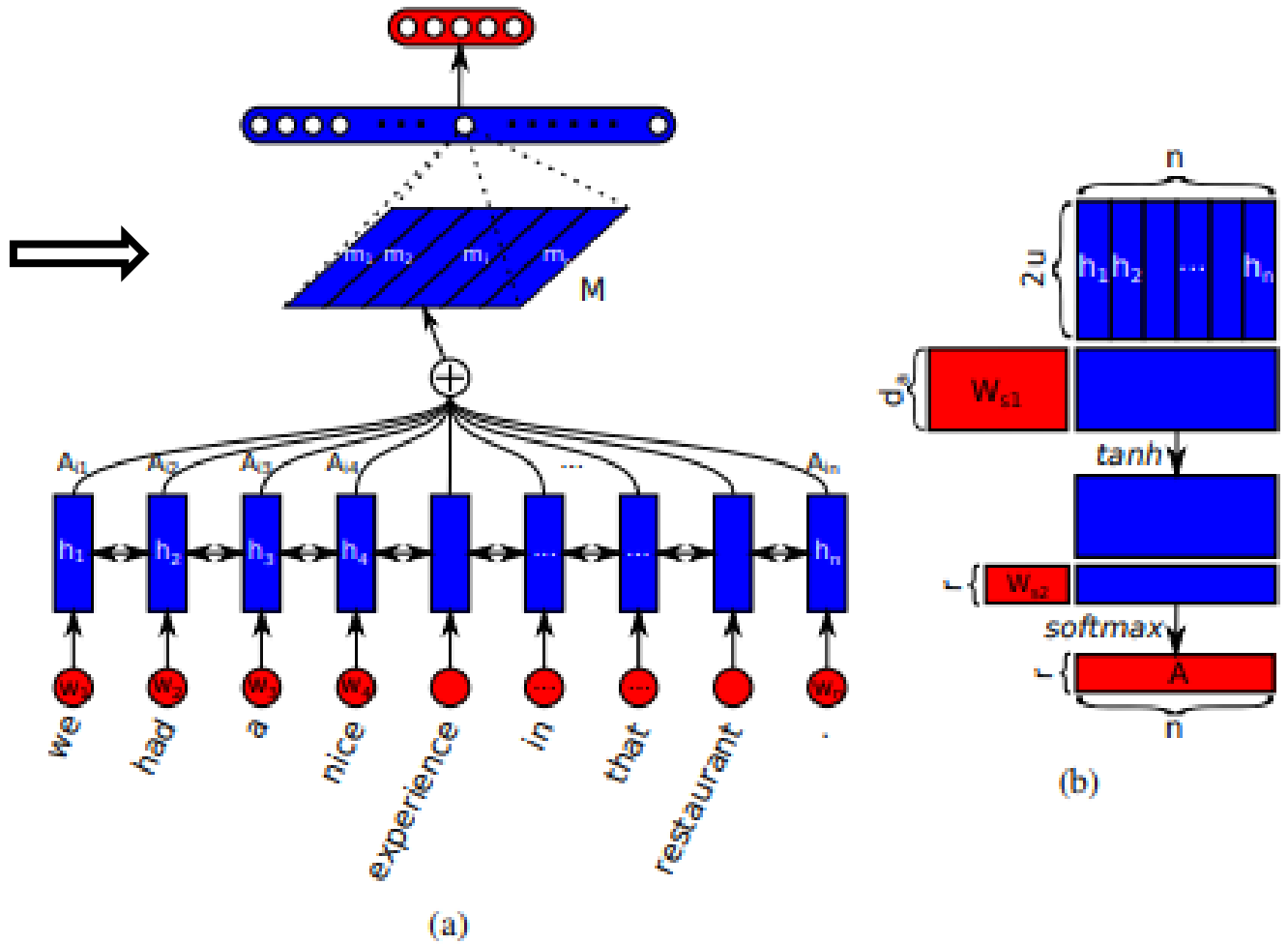
Softmax

⇨ r x n annotation matrix

Approach
기존 Attention과 구성은 같음.

r x n
Annotation
matrix
(softmax)

n x 2u
Hidden
state
matrix



Approach

달라진 점: Penalization
Attention matrix와 그 역행렬을 곱하고 단위행렬을 뺀 행렬의 frobenius norm.
이 것을 original loss와 함께 minimize하는 방향으로 학습하면 됨.

$$P = \left\| \left(AA^T - I \right) \right\|_F^2$$

Frobenius Norm

DOWNLOAD
Wolfram Notebook

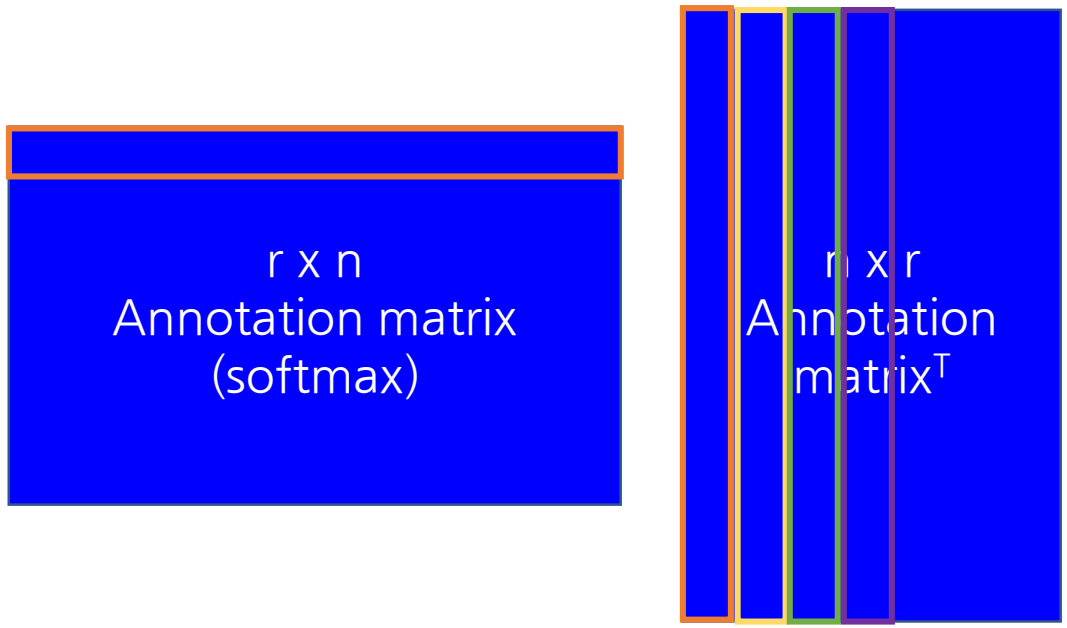
The Frobenius norm, sometimes also called the Euclidean norm (a term unfortunately also used for the vector L^2-norm), is matrix norm of an $m \times n$ matrix $\mathbf{A}$ defined as the square root of the sum of the absolute squares of its elements,

$$\|\mathbf{A}\|_F \equiv \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2}$$

(Golub and van Loan 1996, p. 55).

The Frobenius norm can also be considered as a vector norm.

APPROACH



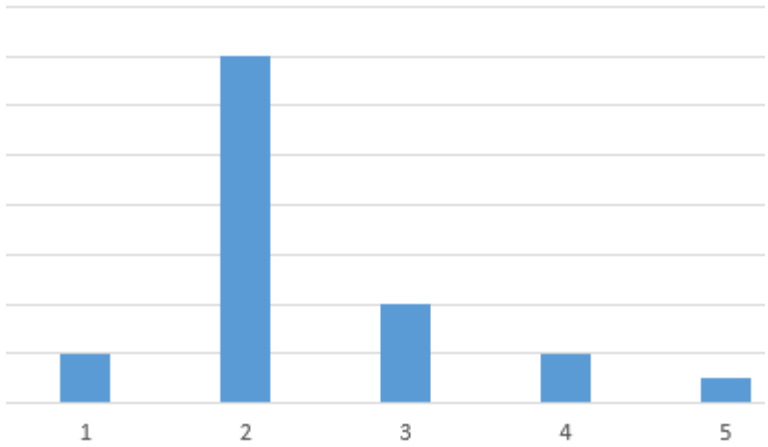
대각선: 나 자신의 attention과의 곱
나머지: 나와 다른 attention과의 곱

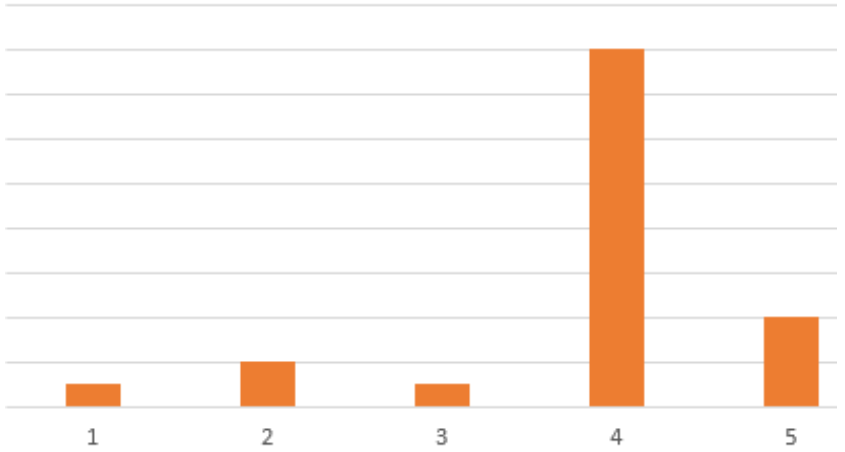
각각의 값들은 0~1 사이의 값을 가짐
(Softmax를 거치기 때문)

Penalization을 하면 좋은 점?

AA^T의 의미: 하나의 annotation vector에서 다른 hop의 vector들을 곱함.
다른 hop의 벡터들이 같은 분포(같은 attention의 모양)이면, 1에 가까운 값을 가지게 될 것.
이것을 error에서 최소화하는 방향(0에 가까이)로 가면 자연스럽게 서로 다른 distribution을 갖게 될 것.
대각선에 1을 빼는 의미: 최대한 좁은 범위의 distribution을 갖게 할 것.

➔ Hop마다 다른 attention을 갖게 하는 annotation vector가 만들어짐.





EXPERIMENTAL RESULTS

Author profiling - Age
Sentiment analysis - Yelp

Table 1: Performance Comparision of Different Models on Yelp and Age Dataset

Models	Yelp	Age
BiLSTM + Max Pooling + MLP	61.99%	77.40%
CNN + Max Pooling + MLP	62.05%	78.15%
Our Model	**64.21%**	**80.45%**

Textual entailment - SNLI

Table 2: Test Set Performance Compared to other Sentence Encoding Based Methods in SNLI Datset

Model	Test Accuracy
300D LSTM encoders (Bowman et al., 2016)	80.6%
600D (300+300) BiLSTM encoders (Liu et al., 2016b)	83.3%
300D Tree-based CNN encoders (Mou et al., 2015a)	82.1%
300D SPINN-PI encoders (Bowman et al., 2016)	83.2%
300D NTI-SLSTM-LSTM encoders (Munkhdalai & Yu, 2016a)	83.4%
1024D GRU encoders with SkipThoughts pre-training (Vendrov et al., 2015)	81.4%
300D NSE encoders (Munkhdalai & Yu, 2016b)	**84.6%**
Our method	84.4%

EFFECT OF PENALIZATION TERM

it ' s an interesting phenomena . Not sure what the spammers get from it . If you comment on Fastco you will get a lot of mail-replies spam .

it ' s an interesting phenomena . Not sure what the spammers get from it . If you comment on Fastco you will get a lot of mail-replies spam .

it ' s an interesting phenomena . Not sure what the spammers get from it . If you comment on Fastco you will get a lot of mail-replies spam .

it ' s an interesting phenomena . Not sure what the spammers get from it . If you comment on Fastco you will get a lot of mail-replies spam .

it ' s an interesting phenomena . Not sure what the spammers get from it . If you comment on Fastco you will get a lot of mail-replies spam .

it ' s an interesting phenomena . Not sure what the spammers get from it . If you comment on Fastco you will get a lot of mail-replies spam .

(a)

it ' s an interesting phenomena . Not sure what the spammers get from it . If you comment on Fastco you will get a lot of mail-replies spam .

it ' s an interesting phenomena . Not sure what the spammers get from it . If you comment on Fastco you will get a lot of mail-replies spam .

it ' s an interesting phenomena . Not sure what the spammers get from it . If you comment on Fastco you will get a lot of mail-replies spam .

it ' s an interesting phenomena . Not sure what the spammers get from it . If you comment on Fastco you will get a lot of mail-replies spam .

it ' s an interesting phenomena . Not sure what the spammers get from it . If you comment on Fastco you will get a lot of mail-replies spam .

it ' s an interesting phenomena . Not sure what the spammers get from it . If you comment on Fastco you will get a lot of mail-replies spam .

(b)

it ' s an interesting phenomena . Not sure what the spammers get from it . If you comment on Fastco you will get a lot of mail-replies spam .

(c) without penalization

it ' s an interesting phenomena . Not sure what the spammers get from it . If you comment on Fastco you will get a lot of mail-replies spam .

(d) with 1.0 penalization

EFFECT OF PENALIZATION TERM

we have a great work dinner here there be about 20 us and the staff do a great job time the course the food be nothing extraordinary I order the New York strip the meat can have use a little more marbling the cornbread we get before the salad be the good thing I eat the whole night 1 annoying thing at this place be the butter be so hard / cold you can not use it on the soft bread get with it

this place be great for lunch / dinner happy hour too the staff be very nice and helpful my new spot

price reasonable staff - helpful attentive portion huge enough for 2 if you get the chimichanga plate food too salty as u know when you cook or add anything with cheese it have it own salt no need add more to the meat ... pls kill the salt and then you can taste the goodness of the food ... ty

(a) Yelp without penalization

we have a great work dinner here there be about 20 us and the staff do a great job time the course the food be nothing extraordinary I order the New York strip the meat can have use a little more marbling the cornbread we get before the salad be the good thing I eat the whole night 1 annoying thing at this place be the butter be so hard / cold you can not use it on the soft bread get with it

this place be great for lunch / dinner happy hour too the staff be very nice and helpful my new spot

price reasonable staff - helpful attentive portion huge enough for 2 if you get the chimichanga plate food too salty as u know when you cook or add anything with cheese it have it own salt no need add more to the meat ... pls kill the salt and then you can taste the goodness of the food ... ty

(b) Yelp with penalization

EFFECT OF MULTIPLE VECTORS

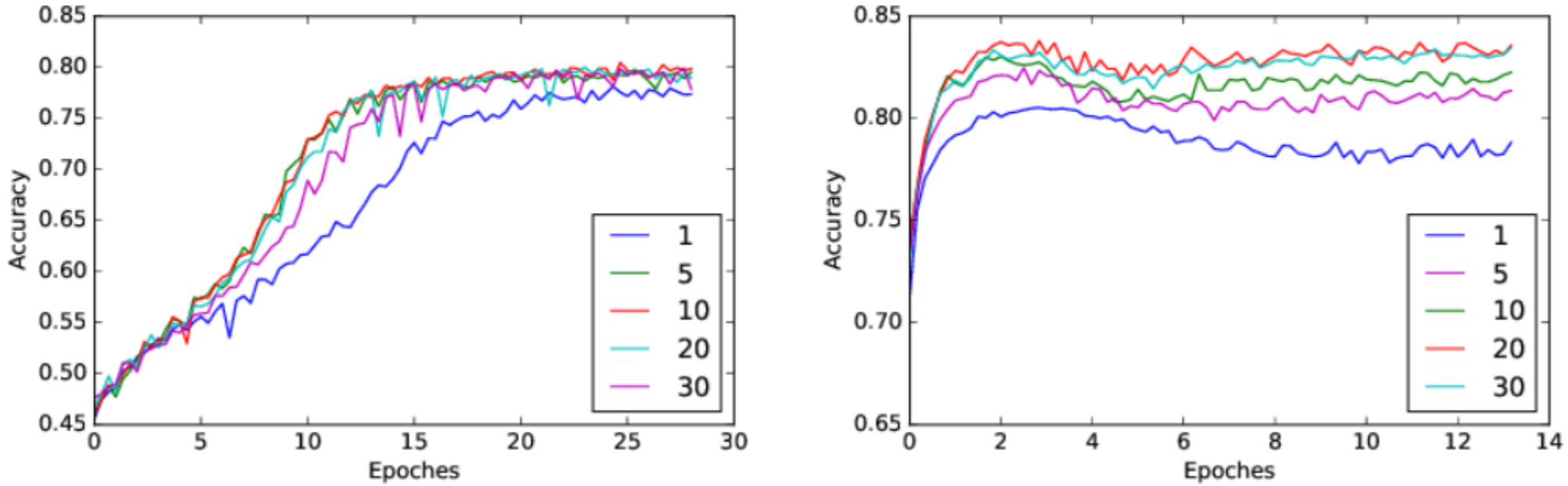


Figure 5: Effect of the number of rows (r) in matrix sentence embedding. The vertical axes indicates test set accuracy and the horizontal axes indicates training epoches. Numbers in the legends stand for the corresponding values of r. (a) is conducted in Age dataset and (b) is conducted in SNLI dataset.

CONCLUSION

Experimental results over 3 different tasks show that the model outperforms other sentence embedding models by a significant margin.

The model is able to encode any sequence with variable length into a fixed size representation, without suffering from long-term dependency problems.

Attention 쓰세요.