layout: post
title: "A Structured Self-Attentive Sentence Embedding"
date: 2019-03-01 13:47:35 +0900
categories: NLP

tag: NLP

2017 ICLR Conference에서 소개된 논문 중 IBM Watson의 Attention mechanism을 사용해 Sentence embedding을 하는 A Structured Self-Attentive Sentence Embedding 논문에 대해서 알아보도록 한다. 해당 모델은 sentence embedding을 위한 self-attention mechanism과 정규화를 위해 새로운 regularization term을 소개한다. 뿐만 아니라 추가적으로 visualizing을 쉽게 할 수 있도록 설계되어 있어 간단하게 visualizing을 할 수 있도록 한다. 해당 모델의 성능을 측정하기 위해서 3개의 task(author profiling, sentiment classification, textual entailment)에서 실험했다.

1 Introduction

Word embedding 기법, 즉 개별 단어들에 대해 유의미한 distributed representation을 학습하는 기법들을 계속해서 많은 발전을 이뤄왔다. 반면 아직 phrase나 sentence의 representation을 만드는 데는 word에 비해 아직은 부족한 상황이다. 보통 이와 같이 phrase나 sentence를 representation하는 방법은 두가지로 나뉜다. 첫 번째는 unsupervised 학습을 사용해 universal sentence representation을 만드는 방법이다.(*SkipThought vector, ParagraphVector, recursive auto-encoders, Sequential Denoising Autoencoder, FastSent*, etc)

또 다른 방법은 특정 task를 위해 특별하게 학습하는 방법이다. 이러한 방법은 보통 supervised 학습하고, downstream application과 합쳐져서 사용된다. 그리고 몇몇 모델의 경우에는 일반적인 단어 임베딩을 사용하고 중간에 *recurrent networks, recursive networks, convolutional networks* 등을 사용함으로써 sentence representation을 얻어 다양한 task에 적용되었다.

Attention mechanism을 CNN 혹은 LSTM 네트워크 상단에 적용함으로써 추가적인 정보를 통해 sentence embedding을 추출하는 모델이 몇가지 task에서 제안되었다. 하지만 sentiment analysis 같은 단일 문장이 입력으로 들어가는 경우에 추가적인 정보로 활용할 문장이 없기 때문에 attention mechanism을 적용할 수 없다.

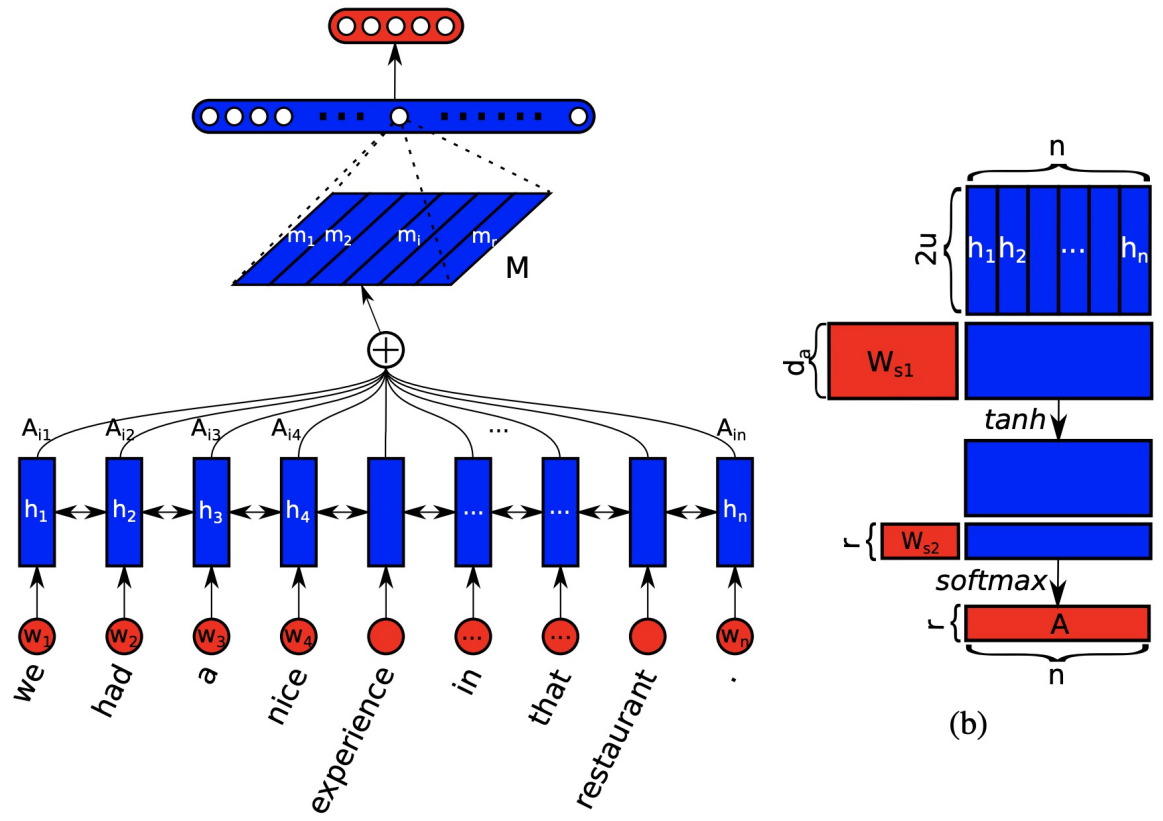
따라서 대부분의 경우에는 max or average pooling 기법을 적용하거나 RNN의 마지막 hidden vector를 선택해서 사용하는데, 해당 모델에서는 self-attention 기법을 통해서 기존의 방법들을 대체한다. self-attention의 경우에는 추가적인 입력값이 없는 하나의 문장에 대해서도 적용할 수 있고, 긴 문장에 대해서도 좋은 성능을 낸다. 이후 section 2.1 에서 self-attentive sentence embedding 모델을 소개하고 2.2에서 모델에서 사용한 정규화 방법에 대해서 소개한다. 마지막으로 2.3 에서는 효과적으로 해당 기법을 시각화 할

수 있는 방법에 대해서 소개할 것이다.

2 Approach

2.1 Model

Sentence embedding 모델은 크게 두개의 part로 구성되어 있다. 첫 번째 part는 bidirectional LSTM 을 사용하는 부분이고 다음은 self-attention을 적용하는 방법이다. 두 번째 part에서 나오는 값들을 사용해 LSTM의 hidden state값을 weighted sum 하게 되고 이 값이 입력 문장에 대한 embedding vector 로 사용된다. 그리고 이 값을 활용해서 각각의 task에 맞게 추가적인 networks를 모델 상단에 적용시킬 수 있다. 예를 들면 sentence embedding vector에 multi-layer perceptron을 적용시켜서 sentiment analysis task에 적용할 수 있다. 아래의 그림은 해당 예시를 도식화한 그림이다.



모델의 세부 과정에 대해서 자세히 알아보도록 하자. 우선 아래와 같이 n개의 token을 가지는 입력 문장이 있다고 하자. 입력 문장은 아래와 같이 각 단어들의 vector들이 모여서 matrix가 된다.

$$
S = (\mathbf{w_1}, \mathbf{w_2}, \dots ,\mathbf{w_n})
$$

여기서 $\mathbf{w_i}$ 는 i번째 단어의 d-dimensional vector이다. 입력 문장 S 는 (n,d) 형태가 된다. 해당 입력 문장에 bidirectional LSTM을 적용시켜 두 개의 u-dimensional hidden vector

값을 구한다.

$$
\begin{matrix}
\vec{h} = \overset{\longrightarrow}{\text{LSTM}}(w_t, \overset{\longrightarrow}{h_{t-1}})\\
\overset{\leftarrow}{h} = \overset{\longleftarrow}{\text{LSTM}}(w_t, \overset{\longleftarrow}{h_{t+1}})
\end{matrix}
$$

Bidirectional LSTM hidden state인 $\overset{\rightarrow}{h_t}$와 $\overset{\leftarrow}{h_t}$를 concatenate한 결과인 h_t를 사용한다. 전체 길이 n에 대해 다음과 같이 n개의 hidden state값이 나오게 된다. 이 값들을 모아서 하나의 matrix로 만들면$(n,2u)$의 size를 가지게 된다.

$$
H=(\mathbf{h_1},\mathbf{h_2},…,\mathbf{h_n})
$$

가변 길이의 입력값에 대해서 동일한 크기의 embedding 값을 얻는 것을 목표로 하기 때문에, n개의 LSTM state를 적당한 linear combination을 통해 일정한 크기로 만들어 줘야 한다. 여기서는 self-attention mechanism을 linear combination으로 사용한다. Attention mechanism의 입력으로는 H를 사용하고, weight로 사용되는 output $\mathbf{a}$가 나오게 된다.

$$
\mathbf{a} =softmax(\mathbf{w_{s2}}\tanh(W_{s1}H^T))
$$

여기서 W_{s1}은 $(d_a, 2u)$ 크기의 가중치 행렬이고, $\mathbf{w_{s2}}$는 (d_a) dimension의 가중치 벡터이다. 최종 output인 $\mathbf{a}$는 n dimension의 벡터가 나오게 된다. 해당 값은 각 token에 대해 얼마나 반영할지를 확률값으로 표현되어있다. 이 값을 사용해 H의 가중 합을 구하게 된다.

$$
\mathbf{m} = \text{sum}(\mathbf{a}\odot H)
$$

이 값은 한 문장에 대해서 하나의 semantic 정보를 담고있다. 하지만 일반적으로 문장의 경우 여러개의 의미를 담는경우가 많이 있다. 예를 들면 'and'로 연결되어 있는 문장의 경우 한문장이더라도 여러개의 의미를 담고 있다. 따라서 이러한 전체적인 의미를 담은 represent하기 위해서 multiple $\mathbf{m}$을 필요로 한다. 따라서 multiple hops of attention을 사용한다. 문장에서 r개의 각각 다른 부분의 의미를 추출하기 위해서 기존의 $\mathbf{w_{s2}}$를 (r,d_a)크기의 가중치 행렬로 확장시켜서 다음과 같이 attention matrix를 구하게 된다.

$$
A = softmax(W_{s2}\tanh(W_{s1}H^T))
$$

이후 최종 output은 위의 attention matrix A와 H를 행렬곱해서 얻게된다.

$$
M = AH
$$

2.2 Penalization Term

앞서 구한 M은 r개의 정보를 담아야 하는데 만약 비슷한 값들만을 갖게 된다면 정확한 정보를 전달하기 어려워지는데 이러한 문제를 해결하기 위해 penalization term을 통해 다양한 정보를 각각의 attention hop이 가질 수 있도록 만들어 준다.

다양성을 평가하는 가장 좋은 방법은 Kullback Leibler divergence를 측정하는 것이다. 하지만 해당 모델에서 KL-divergence를 사용한 경우에 unstable하기 때문에 해당 모델에서는 다른 regularization term을 사용해서 Regularization을 한다. 뿐만 아니라 여기서 제시하는 penalization term의 경우 KL-divergence와 비교해서 연산량이 1/3로 cost 측면에서도 효율적이다.

해당 term은 아래와 같이 계산한다.

$$
P = \Vert (AA^T-I)\Vert_F^2
$$

여기서 사용한 $\Vert\cdot\Vert_F$은 Frobenius norm이다. L2 regularization term과 비슷하게 해당 term은 coefficient를 곱한 후 loss와 함께 최소화하게 된다.

2.3 Visualization term

해당 모델에서 sentence embedding을 interpretation하는 것은 매우 간단하게 annotation matrix A를 사용함으로써 매우 간단히 해결할 수 있다. embedding matrix M의 각 row에 대해 각각 상응하는 annotation vector $\mathbf{a}$를 가진다. 각 element는 각 position의 token이 얼마나 contribution을 한지 확인 할 수 있다. 이 값을 사용해 Visualization을 쉽게 할 수 있다. Visualization결과는 다음과 같이 나타난다.

- if I can give this restaurant a 0 I will we be just ask our waitress leave because someone with a reservation be wait for our table my father and father-in-law be still finish up their coffee and we have not yet finish our dessert I have never be so humiliated do not go to this restaurant their food be mediocre at best if you want excellent Italian in a small intimate restaurant go to dish on the South Side I will not be go back

- this place suck the food be gross and taste like grease I will never go here again ever sure the entrance look cool and the waiter can be very nice but the food simply be gross taste like cheap 99cent food do not go here the food shot out of me quick then it go in

- everything be pre cook and dry its crazy most Filipino people be used to very cheap ingredient and they do not know quality the food be disgusting I have eat at least 20 different Filipino family home this not even mediocre

- seriously f *** this place disgust food and shitty service ambience be great if you like dine in a hot cellar engulf in stagnate air truly it be over rate over price and they just under deliver forget try order a drink here it will take forever get and when it finally do arrive you will be ready pass out from heat exhaustion and lack of oxygen how be that a head change you do not even have pay for it I will not disgust you with the detailed review of everything I have try here but make it simple it all suck and after you get the bill you will be walk out with a sore ass save your money and spare your self the disappointment

- i be so angry about my horrible experience at Medusa today my previous visit be amaze 5/5 however my go to out of town and I land an appointment with Stephanie I go in with a picture of roughly what I want and come out look absolutely nothing like it my hair be a horrible ashy blonde not anywhere close to the platinum blonde I request she will not do any of the pop of colour I want and even after specifically tell her I do not like blunt cut my hair have lot of straight edge she do not listen to a single thing I want and when I tell her I be unhappy with the colour she basically tell me I be wrong and I have do it this way no no I do not if I can go from Little Mermaid red to golden blonde in 1 sitting that leave my hair fine I shall be able go from golden blonde to a shade of platinum blonde in 1 sitting thanks for ruin my New Year's with 1 the bad hair job I have ever have

(a) 1 star reviews

- i really enjoy Ashley and Ami salon she do a great job be friendly and professional I usually get my hair do when I go to MI because of the quality of the highlight and the price the price be very affordable the highlight fantastic thank Ashley i highly recommend you and ill be back

- love this place it really be my favorite restaurant in Charlotte they use charcoal for their grill and you can taste it steak with chimichurri be always perfect Fried yucca cilantro rice pork sandwich and the good tres lech I have had.The desert be all incredible if you do not like it you be a mutant if you will like diabeetus try the Inca Cola

- this place be so much fun I have never go at night because it seem a little too busy for my taste but that just prove how great this restaurant be they have amazing food and the staff definitely remember us every time we be in town I love when a waitress or waiter come over and ask if you want the cab or the Pinot even when there be a rush and the staff be run around like crazy whenever I grab someone they instantly smile acknowlegde us the food be also killer I love when everyone know the special and can tell you they have try them all and what they pair well with this be a first last stop whenever we be in Charlotte and I highly recommend them

- great food and good service .... what else can you ask for everything that I have ever try here have be great

- first off I hardly remember waiter name because its rare you have an unforgettable experience the day I go I be celebrate my birthday and let me say I leave feel extra special our waiter be the best ever Carlos and the staff as well I be with a party of 4 and we order the potato salad shrimp cocktail lobster amongst other thing and boy be the food great the lobster be the good lobster I have ever eat if you eat a dessert I will recommend the cheese cake that be also the good I have ever have it be expensive but so worth every penny I will definitely be back there go again for the second time in a week and it be even good ..... this place be amazing

(b) 5 star reviews

Conclusion & Discussion

해당 논문에서는 self-attention을 사용해서 고정된 크기의 matrix sentence embedding을 만들었다. 해당 모델을 3개의 task에 실험한 결과는 해당 모델의 다른 sentence embedding 모델에 비해 더 좋은 성능을 보인다는 것을 확인할 수 있다.

LSTM의 결과에 attention mechanism을 적용함으로써 LSTM은 마지막 hidden state에 모든 token의

정보를 담을 필요가 없고 단지 각 token의 정보들만을 담으면 된다. 따라서 해당 모델은 sentence의 길이가 길어지더라도 좋은 성능을 보인다.

그리고 해당 모델은 가변 길이의 문장을 하나의 고정된 길이의 representation으로 나타낼 수 있고, long-term의 경우에도 동일하게 정보를 잠 담는다. 이러한 장점은 모델이 scalability 하다는 것을 나타낸다. 따라서 단순 문장이 아니라 paragraph, articles등 더욱 긴 content에도 적용할 수 있다는 것을 볼 수 있다.