Intent Analysis on Social Media Using Convolutional Neural Networks and Word2Vec

Thet Naing Tun

University Of Computer Studies,Yangon

Frontend Developer@Bindez

🐦 @thetnaingchen

About Thesis

Analysing user intention based on the user feedback(comments) posted on social media.

Telecommunication is the target domain.

User feedbacks are classified into five classes(Application, Auto Subscription, Bill, Customer Service and internet).The proposed system is sentence(comment) level classification.

The proposed system uses Word2Vec to transform word into vector and the resulting vectors are fed into Convolutional Neural Network to classify the sentences into pre-defined classes.

Pre-Processing(Cont'd)

1. **Font Converting**

   Myanmar language have font issue(Zawgyi, Unicode).Most of the comments posted on social media are using zawgyi font. Collected comments are converted to Unicode by using python implementation of rabbit library.

2. **Correction of Spelling Errors**

   User generated content from social media are informal such as the mix of Myanmar word to English word(ပြီးp), repeat word(အားးးး) and so on. All of the informal text are corrected manually.

3. **Word Segmentation**

   In the proposed system, word level training data are used. Sentences are segmented by using Myanmar Word Segmenter from UCSY-NLP Lab.

Statistics of Training Data

Total Number of Comments : 13869

Class Name	Number of Comments
App	180
Auto Subscription	246
Bill	631
Customer Service	1290
Internet	411

Classes and Example Feedbacks

Application

My telenor app ကို sign in ဝင်မရလို့ပါ

Auto Subscription

သုံးစွဲသူရဲ့ခွင့်ပြုချက်မရှိပဲ သူ့အလိုလိုဂိမ်းတွေထဲ Subscription လုပ်ပြီး ဘေဖြတ်နေတာရပ်ပေးပါ ခုလုပ်ရပ်က အတော်အောက်တန်းကျနေပြီ

Bill

ဘယ်လိုခိုးအုံးမှာလဲ telenorရယ်

Customer Service

09773205814 အပိတ်ခံထားရတယ် ဘာကြောင့်လဲ ကူညီပါအုံ

Internet

အင်တာနက်လိုင်းက ဒီတစ်သက် ကောင်းဦးမှာလား စုတ်ပြတ်နေတာပဲ

Word2Vec

Word2Vec is used to transform word into vector.

Word2Vec learns the meaning of words just by processing the large corpus of

unlabeled text.

There are two architecture for Word2Vec. Skip-Gram and Continuous bag-of-words.

Continuous bag-of-words(CBOW) predicts the target word from the nearby words.

Skip-Gram approach predicts the context of words from a target word.

The proposed model used Skip-Gram approach.

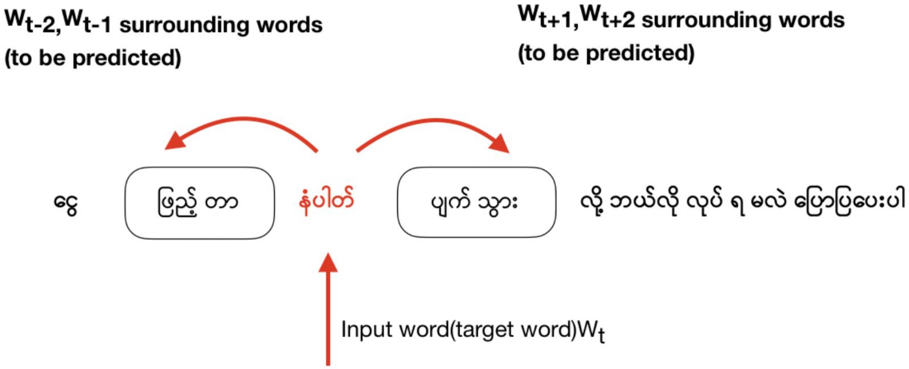
Word2Vec(Cont’d)
(Skip-Gram)

Skip-Gram model train on simple neural network with single hidden layer.

Skip-Gram predicts the nearby context words from the continuous sequence of training words.

The word “**nearby**” mean that there is a window size parameter to the algorithm.The model predict the words within the window.For the window size of two, the model predicts the two words behind and two words ahead of target word.

Word2Vec(Cont’d)
(Skip-Gram)



W_{t-2},W_{t-1} surrounding words
(to be predicted)

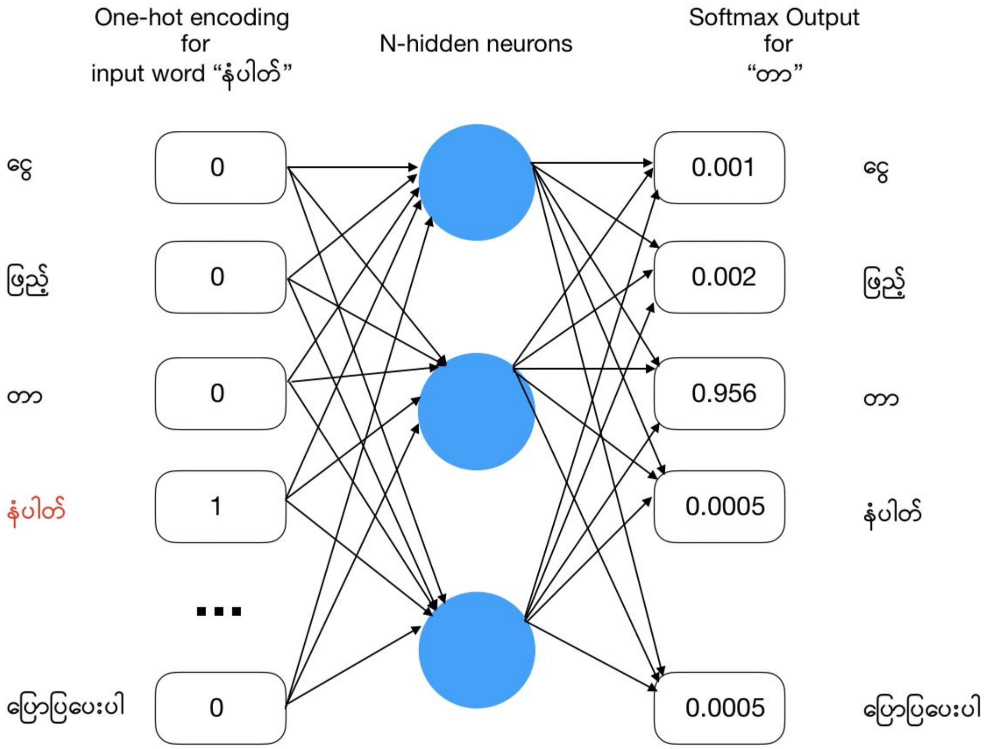
W_{t+1},W_{t+2} surrounding words
(to be predicted)

ငွေ ဖြည့် တာ နုံပါတ် ပျက် သွား လို့ ဘယ်လို လုပ် ရ မလဲ ပြောပြပေးပါ

Input word(target word)W_t

Word2Vec model using skip-gram window size of two words
(considering two words before and after of each target word)

Word2Vec(Cont'd)
(Network Architecture of Skip-Gram Model)



One-hot encoding
for
input word “နံပါတ်”

N-hidden neurons

Softmax Output
for
“တာ”

ငွေ	0	0.001	ငွေ
ဖြည့်	0	0.002	ဖြည့်
တာ	0	0.956	တာ
နံပါတ်	1	0.0005	နံပါတ်
	...		
ပြောပြပေးပါ	0	0.0005	ပြောပြပေးပါ

Word2Vec(Cont'd)
(Retrieve Word Vectors)

One-hot vector in vocabulary of four words

0	**1**	**0**	**0**

×

3 Neurons Weight Vectors

0.23	0.11	0.43
0.33	0.22	0.76
0.83	0.21	0.22
0.19	0.98	0.51

=

0.33	**0.22**	**0.76**

Convolutional Neural Network (CNN)

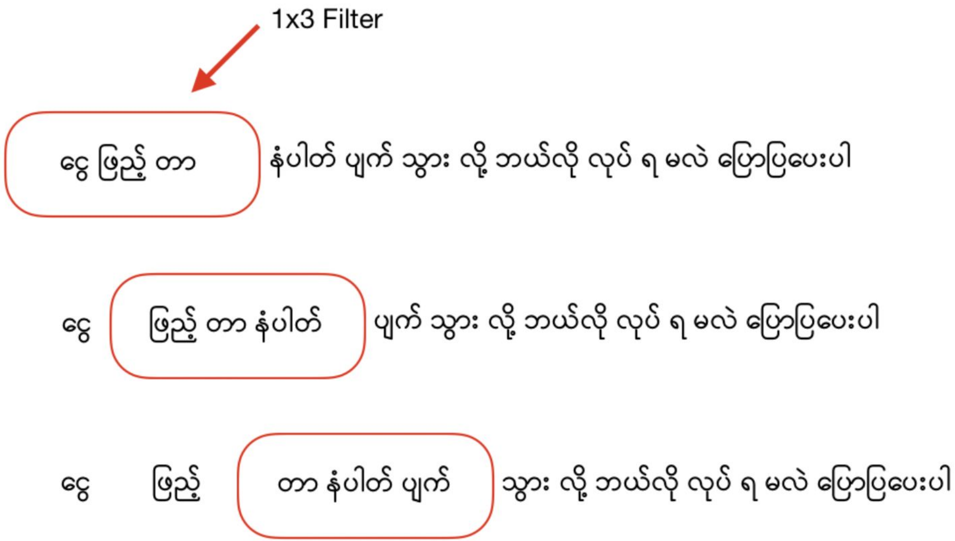
Generally Used in Computer Vision.However they have good result when applied to NLP.

The most natural fit for CNN is classification tasks such as Sentiment Classification, Topic Categorization and Spam Detection.

There are two major components in Convolutional Neural Network.Convolutional Layer and Pooling Layer.

Convolutional Layer is used to detect features in text and Pooling Layer is used for dimensionality reduction.

Convolutional Neural Network(Cont’d)
(1D Convolution)



1x3 Filter

ငွေ ဖြည့် တာ နံပါတ် ပျက် သွား လို့ ဘယ်လို လုပ် ရ မလဲ ပြောပြပေးပါ

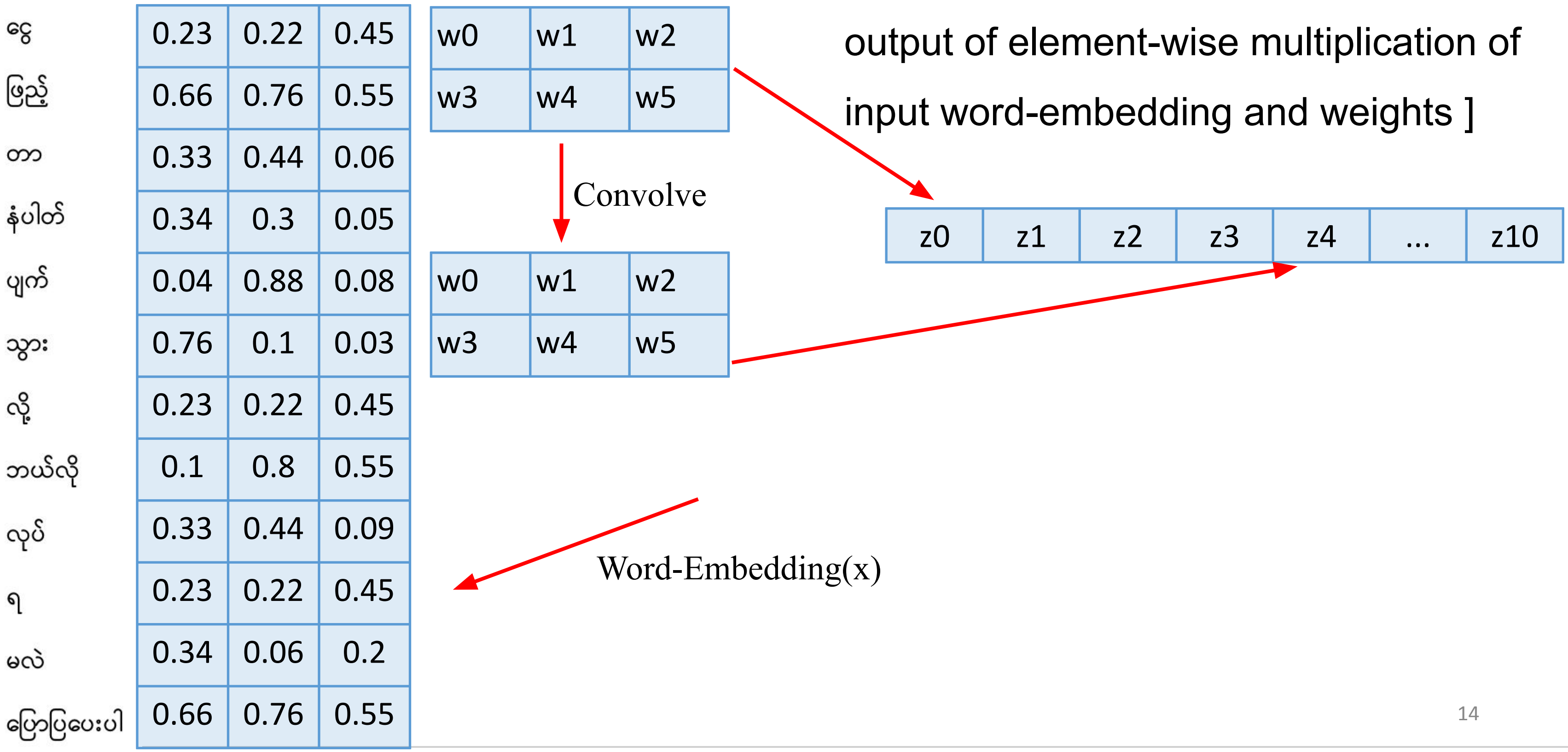
ငွေ ဖြည့် တာ နံပါတ် ပျက် သွား လို့ ဘယ်လို လုပ် ရ မလဲ ပြောပြပေးပါ

ငွေ ဖြည့် တာ နံပါတ် ပျက် သွား လို့ ဘယ်လို လုပ် ရ မလဲ ပြောပြပေးပါ

One-dimensional filters convolve over one-dimensional input (sentence)

Convolutional Neural Network(Cont'd) (1D Convolution)

z0 = max(x*w,0)[ReLu activation function output of element-wise multiplication of input word-embedding and weights]



ငွေ	0.23	0.22	0.45
ဖြည့်	0.66	0.76	0.55
တာ	0.33	0.44	0.06
နံပါတ်	0.34	0.3	0.05
ပျက်	0.04	0.88	0.08
သွား	0.76	0.1	0.03
လို့	0.23	0.22	0.45
ဘယ်လို	0.1	0.8	0.55
လုပ်	0.33	0.44	0.09
ရ	0.23	0.22	0.45
မလဲ	0.34	0.06	0.2
ပြောပြပေးပါ	0.66	0.76	0.55

w0	w1	w2
w3	w4	w5

Convolve

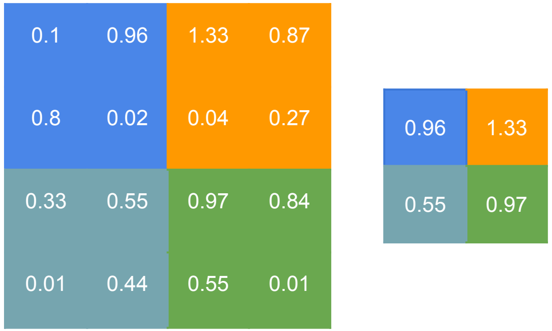
w0	w1	w2
w3	w4	w5

z0	z1	z2	z3	z4	...	z10

Word-Embedding(x)

Convolutional Neural Network(Cont’d) (CNN)

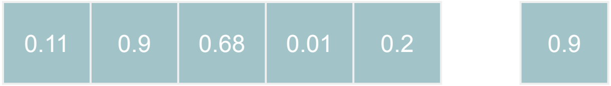
There are two types of pooling.Max Pooling and Average Pooling.



0.1	0.96	1.33	0.87
0.8	0.02	0.04	0.27
0.33	0.55	0.97	0.84
0.01	0.44	0.55	0.01

0.96	1.33
0.55	0.97

Two Dimensional Max Pooling



0.11	0.9	0.68	0.01	0.2

0.9

One Dimensional Global Max Pooling

There are also two types of Max Pooling: Ordinary Max Pooling and Global Max Pooling.The proposed model uses one dimensional global max pooling.

Convolutional Neural Network(Cont'd)

(Global-Max Pooling Layer)

1x3 filter

ငွေ ဖြည့် တာ နံပါတ် ပျက် သွား လို့ ဘယ်လို လုပ် ရ မလဲ ပြောပြပေးပါ

Filters are applied to each input sample.

Each filter produce one dimensional vector which is slightly smaller than the original input.

For each filter, take single maximum value from each filter's output.

So the resulting vector is 1 x no_of_filter and this is the higher representation for the corresponding input sample.

Challenges

NLP tools and datasets for Myanmar Language are under resourced.

Only few open sources datasets and tools are available.

User generated contents on social media have a lot of noises(spelling error, mix of Myanmar Words to English words(ပြီးp), informal).

Experiment Result Evaluation and Further Experiments

Experimental results are shown by confusion matrix and reports in precision and recall.

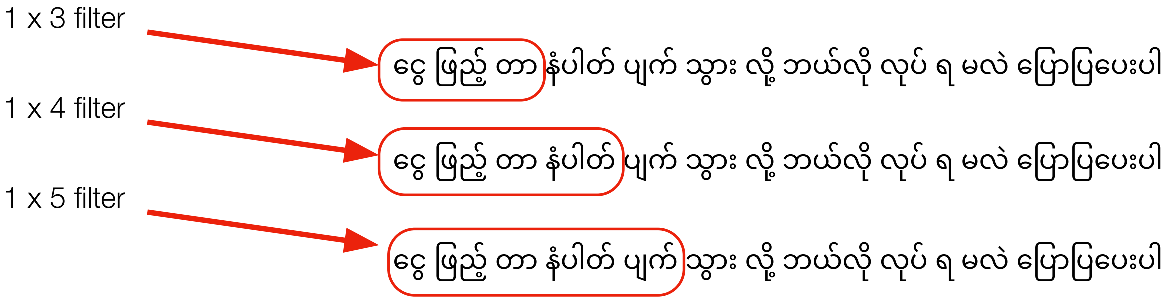
Further Experiment:

CNN model with Multiple Channels(multiple filter width)

(Yoon Kim.“Convolutional Neural Networks for Sentence Classification” 2014)

In this experiment input sentences are detected by different filter width.

Experiment Result Evaluation and Further Experiments(Cont’d)



1 x 3 filter

1 x 4 filter

1 x 5 filter

ငွေ ဖြည့် တာ နံပါတ် ပျက် သွား လို့ ဘယ်လို လုပ် ရ မလဲ ပြောပြပေးပါ

ငွေ ဖြည့် တာ နံပါတ် ပျက် သွား လို့ ဘယ်လို လုပ် ရ မလဲ ပြောပြပေးပါ

ငွေ ဖြည့် တာ နံပါတ် ပျက် သွား လို့ ဘယ်လို လုပ် ရ မလဲ ပြောပြပေးပါ

CNN Model with multiple channel(filter)

Resources

Word2Vec:

Understanding word vectors: A tutorial for "Reading and Writing Electronic Text,".

Word2Vec Tutorial - The Skip-Gram Model.

CNN:
Understanding Convolutional Neural Networks for NLP

Evaluation:

Multi-Class Metrics Made Simple, Part I: Precision and Recall

Thank You!