Title

Tal Levi, Omri Maoz

Tel Aviv University

Abstract

We were to test the hypothesis that a task of classifying contextualize phrases can be better execute with common NLP tools when feeding them with their word to word corresponding tagging phrases. The key observation when examine a contextual phrase is that there’s a common semantic between some words in phrases that are belong to the same class. For example, a positive review of a movie contains word that hold a positive semantic meaning and this semantic is common in words in other positive reviews. We believe that these semantic might be express with some POS tagging pattern in the corresponding POS tag phrase. Moreover, the POS tagging sentences significantly reduce the vocabulary size, therefore we assume two assumption to test in this project:

1. It possible to learn POS tagging pattern to achieve better result on NLP classification task.
2. An equivalate results can be achieve with fewer data.
3. Introduction

Classification of commercialize text phrases is a task with many application that we use and affect our daily lives. For an example is a technology that scan comments and reviews of users in social network, searching for violation of the terms and rules set by the network In order to remove and block those “bad” contents.

A lot of work and study has been done in this subject and found that deep learning tools achieve good result. Models such as CNN-GRU and BiLSTM for examples achive around 80% accuracy on Movie Reviews dataset with more then 10,000 reviews divided to 2 classes.

Recent work have shown that a task of Question classification in Thai language

Abstract

Abstract

We were to test the hypothesis that a task of classifying contextualize phrases can be better execute with common NLP tools when feeding them with their word-to-word corresponding tagging phrases. The key observation when examine a contextual phrase is that there’s a common semantic between some words in phrases that are belong to the same class. For example, a positive review of a movie contains word that hold a positive semantic meaning and this semantic is common in words in other positive reviews. We believe that these semantic might be express with some POS tagging pattern in the corresponding POS tag phrase. Moreover, the POS tagging sentences significantly reduce the vocabulary size, therefore we assume two assumptions to test in this project:

1. It possible to learn POS tagging pattern to achieve better result on NLP classification task.
2. An equivalate results can be achieve with fewer data.
3. Introduction

Classification of commercialize text phrases is a task with many applications that we use and affect our daily lives. An example is a technology that scan comments and reviews of users in social network, searching for violation of the terms and rules set by the network in order to remove and block those “bad” contents.

A lot of work and study has been done in this subject and found that deep learning tools achieve good result. Models such as CNN-GRU and BiLSTM for examples achieve around 80% accuracy on Movie Reviews dataset with more then 10,000 reviews divided to 2 classes.

Add link or image from the site (a work not a study)- <https://towardsdatascience.com/deep-learning-techniques-for-text-classification-78d9dc40bf7c>

Improving results for those NLP models is not an easy task and there’s many variables and parameters to consider alongside many approaches to test and experiment. One approach, which supported by recent research shows that preprocessing techniques of the data manage to improve NLP models results.

Part-of-Speech (POS) tagging is a popular NLP process which refers to categorizing words in a text in correspondence with a particular part of speech, depending on the definition of the word and its context.

Recent study on Question classification in Thai language have shown that text classification can benefit from Part-of-Speech (POS) tagging as part of feature selection method to achieve better result. In Thai language some words have different meaning when considered alone than when joining other words because of a meaning word based on ordering the sequence of words and context. Therefore, considering a syntactic feature for the obvious classification of Thai sentences is necessary. One feature the research editors manage to find by analyzing frequency of POS tag for each question type in the packages “PyThaiNLP” and “Stanford CoreNLP”, is that each question type was sensitive to some POS tagging ratios.

Add link to the study.

In order to test our experiment hypothesis, we are focusing on 3 NLP models:

* DAN – Deep Averaging Network.
* BiLSTM – Bidirectional Long-Short Term Memory
* Transformer.

We will feed those models with 2 datasets:

* “IMDB reviews” dataset, divided into 2 classes (positive, negative), with more then 15k reviews.
* “News headline” dataset, divided to many topics as classes, but we will focus on the 6 most common, with more than 3000 headlines per category.

One more motivation to use POS tagging as general English words categories is that this preprocessing methodology is general and not depends on pre known knowledge so it can be applied to any given dataset.

We organize the rest of the paper as follow. פסקה על המשך המאמר – ךהשלים בסוף

1. Related Work

[Text classification](https://www.projectpro.io/project-use-case/nlp-text-processing-classification-python) is the process of classifying or categorizing the raw texts into predefined groups. There’s many studies and works developing and testing numerous approaches and methods to tackle this problem and with great results. From fundamental machine learning algorithm such as linear regression and SVM to state-of-the-art deep learning models.

Today, commercial hi-tech companies integrate text classification technologies in their products. From filtering spam emails, to analyzing politician’s speeches, abuse content marking and removal and much more.

1. Methodology
   1. Datasets

For performance evaluation of the hypothesis suggested we used two datasets to test binary classification and multi-class classification. The IMDB reviews dataset contain various length reviews dividing to positive and negative reviews. This dataset is interesting to extract feature from because positive and negative writing have many language forms (cynicism, exaggeration, straight-forward, etc.). The second dataset is News headlines dataset, which contain 40 different subjects (As classes). In contrast to the first dataset, this dataset has approximately close headlines length, but may contain some interesting tagging patterns. We chose 7 subjects to focus on (Politics, Entertainment, Wellness, Travel, Sport, Busyness and Crime), each with more than 3k headlines. We will use different sizes of subset from the datasets, with decrease order in order to test the 2 research assumption, split validation factor will be 0.2 for evaluation, and we will preserve uniform distribution between classes.

* 1. Data preprocessing
     1. Tagging and Tokenization

We used Stanza (Python NLP package) to create for each sentence from the dataset a corresponding POS tagging sentences by calculate each word it’s tag.

To tokenize words, we used spaCy (Python NLP Library) with ‘en\_core\_web\_sm’ trained pipeline.

* + 1. Feature selection

The goal is to improve text classification with the 3 tested NLP models. In order to do that we calculate 3 preprocess methods to produce features which in turn will be fed to the models and will try to improve result accomplish by them.

The methods:

## Tag Filtering – when overviewing the text in the dataset we notice that some tags are more capable to form a combination with other tags in order to express a tag pattern and semantic relation. POS tags such PUNCT (punctuation), PRON (pronoun), etc. are less likely to belong to a key sub-sentence which point to the class semantic of the sentence. On the other hand, NOUN, ADV (adverb), ADJ (adjective) and VERB are more likely to do so. For example, a ‘Crime’ headline from the News headline dataset: Add figure from below

## With “Tag Filtering” we filter all non NOUN, ADV, ADJ and VERB tags from the dataset.

1. Tag Extend – most of NLP deep-learning models work better with big vocabulary. By looking on POS tag corresponding phrases we greatly reduce the vocabulary size. In order to compensate this effect, we will extend the 4 “interesting tags” (NOUN, ADV, ADJ and VERB) to subcategories.
   1. NOUN and VERB – python NLP package call NLTK with WordNet lexical database by Princeton are providing with subcategories for nouns and verbs. We use Synset, a tool from NLTK to look for semantic relation words in WordNet, to distribute words to the subcategories.
   2. ADJ and ADV – We took ADJ and ADV subcategories from (add-links), add suffix as subcategories and used Spacy similarity feature tool to distribute words with these subcategories.
2. Bigram Interpolation – In order to highlight tag patterns in text we used pairing 2 close words in the text in a most common appearance in the dataset, with a decrease order. Doing that, we are emphasize common tag pairs over lesser common tag pairs.

Lastly, we will use different combination of the methods above as feature selection to test our models with.

* 1. Classification (models) – In order to test our hypothesis a comparison need to be done and for that we need relevant NLP models. DAN, BiLSTM and Transformer models are a good choice because while differ in their architecture, they are taking into calculation information between neighbor words in the text.
     1. DAN – Deep Averaging Network: *fill*
     2. BiLSTM – Bidirectional Long Short-Term Memory - we constructed embedding layer of size 300 follow by 3 recurrent layers with 128 BiLSTM cells, a 50% dropout layer, a linear layer from 128 to 32 neurons with Relu activation function, and a Linear layer with Softmax activation function from 32 to the dataset classes number neurons.
     3. Transformer – Might be not relevant as it a seq-to-seq models.

1. Evaluation – Metrics

For evaluate results we chose accuracy and F1-score metrics. Accuracy is the measure of all the correctly identified cases. At the end, the task is classification, so we interest with classified correctly. F1-score is the harmonic mean of Precision and Recall, therefore gives a better measure of the incorrectly classified cases than the Accuracy metric. We will use F1-score to determine how balance our models preform.

1. Experiments

“Joe Arpaio Is **Still** **Guilty** Despite

PROPN PROPN AUX **ADV** **ADJ** ADP

**Presidential** **Pardon** , **Judge** **Says**”

**ADJ** **NOUN** PUNCT **NOUN** **VERB**

1. Introduction – general talk, dataset, modeling, what we attend to do
2. Related work - <https://www.sciencedirect.com/science/article/pii/S2405844021023197>

<https://towardsdatascience.com/deep-learning-techniques-for-text-classification-78d9dc40bf7c>

https://medium.com/analytics-vidhya/accuracy-vs-f1-score-6258237beca2

גם לקחת השראה מכאן

1. Methodology – models, data preprocessing, tokenization, embbeding, feature selection
2. Evaluation – metrics
3. Experiment result
4. Conclusions