Sentence Classification for News Titles

Chien-Wei Lin
University of Michigan
chienwli@umich.edu

I-An Huang
University of Michigan
huangian@umich.edu

Hung-Wei Wu
University of Michigan
hungwei@umich.edu

1 Project description

Classifying the semantic content is one of the critical problems in natural language processing. There are many cases where only a small number of words are provided to interpret the meaning or intent such as keyword searches. However, the performance of short text classification is limited due to shortness of sentences, which causes sparse vector representations if we use word occurrence to represent sentences, and lack of context. On the other hand, news titles, though consisting of short sentences, provide rich information of the semantic content in a concise way. Because of this property, we believe that news title classification will be a good start point for our sentence classification task.

With news classification in mind, we start with a representative dataset News Aggregator Data Set. In this way, we are free from the hassle of data preprocessing annotating date, removing undesired value, and compensating for missing fields of a sample. Besides, its easier to find related work that also does evaluation on this dataset, which makes comparison with our model easier or possible. Moreover, since there are only 4 news categories in this dataset, the task should be relatively easy and makes a great starting point for our project because of the potential difficulty described in the previous paragraph. Our plan is to increase the complexity gradually so that we would have a good idea of what is going on and be in control of the machine learning model. With regard to datasets, we would be searching for datasets that have short summaries like news headlines and a larger number of categories.

With respect to the algorithm to solve the problem, we are going to adopt machine learning techniques. Machine learning is shining and promising in tackling hard natural language processing problems in these days. Our plan for choosing from different machine learning models is to follow the best practice, which starts with simple models and then more complex ones. In this way, we could avoid overestimating the difficulty of our problem, which is unfamiliar to us. Specifically, the simplest model in our consideration is logistic regression because it retains much power without sacrificing simplicity. When things go complex and beyond the capability of simple models, our ultimate approach is to devise a non-trivial neural network to cope with upcoming difficulty. Since there are already many prior work done in NLP, we should be able to find some great foundation on which our work could be based on.

When it comes to evaluation, we want to compare different combinations of factors. Beside different choices of machine learning algorithms mentioned above, the choice of representation is also of interest to us. As far as news title classification is concerned, we mostly deal with sentence representation. Because of its simplicity, we would like to use the bag-of-words model as our baseline representation. Beyond that, we are also considering transforming word2vec representations of words in a headline or sent2vec embeddings. Lastly, we would like to compare similar works either directly or by implementing their model and evaluating in the same setting of our work. Due to insufficient results at hand, we leave the discussion of different neural network architectures to the future.

To briefly summarize, we define our problem

as the following: given news titles from various sources of data, we want to classify them into predefined news categories such as sports, politics, or technology.

As of our current status, we have already implemented bag-of-words model for our dataset and incorporate pretrained word2vec from Google News separately. Both have good performance using logistic regression, with the bag-of-word approach being much superior. Notably, we find out that linear SVM only makes little difference from logistic regression in terms of accuracy but requires much more training time in our task. Consequently, we probably will exclude SVM from our candidate list and return to it when there is free time. We will keep investigating and try these embeddings with modern CNN from related work. We plan to start considering changing machine learning model when our scope of project has been expanded to such an extent that the accuracy falls below 90%. Currently, we decide to expand the scope by using datasets of a larger domain or having more categories to predict.

2 Datasets

2.1 Data Collection and Annotation

1. *News Aggregator Data Set* is provided in UCI machine learning repository with titles, timestamps, publisher information. It contains more than 420000 annotated topics. The dataset is already annotated with 4 categories. (b = business, t = science and technology, e = entertainment, m = health)

2. *News API (or similar services)* has provided API with an option to query with a news category and responds with the title of articles in the category. We can utilize the service to create massive data for our classifier. The service has provided at least 9 possible categories, which are business, entertainment, gaming, general, music, politics, science-and-nature, sport, and technology. We can query with different categories evenly and assign corresponding labels.

2.2 Interesting Data Samples

Take some interesting mis-classified samples by the bag-of-words logistic regression model for example.

- *Colon cancer rates drop 30 percent, report says*: The true class is health, while our bag-of-words model would predict a business class. It might because the word *rates* is often used in business titles.

- *Controllers work exhausting schedules*: The true class is health, while the bag-of-words model would predict an entertainment class. It might because the word *controllers* usually appears in entertainment topics, such as gaming controllers.

- *Seattle & Portland among most walkable cities*: The gold class is business, while the bag-of-words model would predict an entertainment class. This is a more interesting and challenging case, since only if the model understands that this kind of titles are often written to attract workers to working in the city (walkable city implies comfortable environment), the model could classify it as business but entertainment.

3 Related Works

Yoon Kim, Convolutional Neural Networks for Sentence Classification (Kim, 2014) The paper aims to classify sentence using convolutional neural network(CNN). The model would use word2vec as embedding method and parts of sentences as input of CNN. It has been trained in multiple sentence classification dataset including IMDB movie reviews.

Rui Zhao and Kezhi Mao, Topic-Aware Deep Compositional Models for Sentence Classification (Zhao and Mao, 2017) In addition to feeding word embeddings to the convolutional neural network which contains general word information, the model tries to incorporate word level and sentence level information by introducing latent Dirichlet allocation on task-specific corpus. The level information would be concatenated to different network layers. The paper further evaluate their model with/without word/sentence level concatenation and compared the results of several topic classification datasets.

Exploring Newspaper Language : (Hagen, 2012) The book first introduced to classic method of topic classification, such as rule-based and pattern-matching based approaches. Then the article talks

about how to extract features and build datasets. The author further introduced a method using concatenated binary classifiers to classify news topics. The author also provided an evaluation method different to just counting numbers of correct guesses. The evaluation method is extended on how much does a human agree/disagree to the classification made by the model.

4 Model

4.1 Sentence Representation

We are considering two different sentence representations which are bag of words (BoW) and word2vec as the sentence embeddings. With representations as the input, the CNN architecture , which is adapted from Collobert et al (2011), is capable of classifying the categories of the news titles.

BoW is a simple model which generally used in NLP to simplify the representation in classification tasks. The idea is to consider a sentence or a document as a bag which contains the counts of occurrence of words, disregarding grammar and the order. By analyzing how many categories (words) in the documents, BoW records the number of each category appears in a sentence.

Given a document, word2vec maps every word in the document into a unique vector that can be easily computed mathematically. Word2vec learns vector representation from the training text data. Since common words will be mapped to closer region in the vectors space, it has been used widely in natural language processing and machine translation. Specifically, we are using pre-trained 300 dimension vectors for three million words and phrases from Google News database in this project.

4.2 Architecture

Let $x_i \in \mathbb{R}^k$, where k is the dimension of each word vector, be the i-th word in the sentence. Suppose the longest sentence has length n. We pad zeros to those sentences whose length is less than n. After concatenation operation, each sentence has its representation as a matrix $\mathbb{R}^{nk}$.

Convolution operation is applied to a window $w \in \mathbb{R}^{hk}$, where h is the window length. In other words, h is the number of words to produce a new features. Consequently, we will have a new feature map with length $n - h + 1$ for every n length sentence. Let $x_{i:j}$ refers to the concatenation of word $x_i, x_{i+1}, \ldots, x_{j-1}, x_j$. We have

$$c_i = f(w \cdot x_{i:i+h-1} + b) \tag{1}$$

where f is a nonlinear function such as sigmoid or hyperbolic tangent, w as the weights, and b is the bias term. Then we have the feature map $\mathbf{c} = [c_1, c_2, \ldots, c_{n-h+1}]$ from each possible window. For all the windows with different length, we then applied multiple filters on the feature maps to obtain different features form the convolution layer.

The size of the output from the convolution layer depends on the length of the windows applied to the sentence. However, the local features fed to the network should have the fixed size independent of the length of windows. To apply the affine transform on the next layer, weighted average and max operation over the time are used to deal with this type of issue. However, the weighted average operation does not make much sense in our design. Instead, we apply a max-over-time pooling operation (Ronan Collobert, 2011) over the feature maps. Max pooling is a downsampling strategy in CNN. It is a filter on a features that will take the maximum value correspond to the region. This idea is to indicate the most important feature for each feature map and fixed the size to feed to the standard affine network layer.

4.3 Regularization

These features are fed to the fully connected layer with dropout and softmax output. For regularization, we apply dropout to reduce the complex co-adaptation of hidden unit. Randomly removed the hidden neurons from the network with Bernoulli random variable, the output selects p portion of the hidden unit to conduct the forward pass and does not apply any update of the weights on the backward pass. If neurons are randomly dropped out of the networks during training, other neurons will have to step in and handle the representation required to make prediction for missing neuron. Therefore, we have

$$y = w \cdot (z \circ r) + b \tag{2}$$

For each output unit y, instead of

$$y = w \cdot z + b \tag{3}$$

where $r \in \mathbb{R}^m$ is a masking vector of Bernoulli random variables with probability p of being 1. At testing procedure, we scale the weights as $\hat{w} = pw$, and $\hat{w}$ is used (without dropout) to score the unseen data.

5 Evaluation

We are using News Aggregator Data Set and chose the 70% of sentence as training dataset, 15% as the validation dataset, and 15% as the testing dataset. This ratio is followed by the empirical experience. We applied general machine learning classification algorithm such as logistic regression and SVM as our baseline model. In addition, we have interested in two different sentence representations, such as BoW and word2vec for sentence embeddings. Due to its simplicity, we chose the BoW as the baseline representation model. Currently, our objective model is the CNN with the word2vec representations of words in sent2vec embeddings. For the next step of the project, we would like to evaluate our objective model with other combinations of different machine learning algorithms and the different sentence representations.

References

Thomas M. Hagen. 2012. Automatic topic classification of a large newspaper corpus. *Exploring Newspaper Language : Using the Web to Create and Investigate a Large Corpus of Modern Norwegian*, pages 111–130.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

L. Bottou M. Karlen K. Kavukcuglu P. Kuksa Ronan Collobert, J. Weston. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research I2*, pages 2493–2537.

Rui Zhao and Kezhi Mao. 2017. Topic-aware deep compositional models for sentence classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(2), FEB.