2020hci Explanation of Program

1. Introduction

My research focuses on naming recognition techniques in the field of natural language processing. The number of net-words that appear in social networking service is increasing day by day. In this research, we have developed a Japanese corpus of net-words from Twitter, and use embeddings to detect net terms with high accuracy by deep learning using the corpus.

1. Japanese Net-words Corpus
   1. Select a total of 40 net-words
   2. Collect 30 Tweets include those net-words
   3. Collect 30 Sentences whose key words are used in normal meaning
   4. Tokenizer the data by Mecab correctly
   5. Annotation the corpus
   6. Spilt the dataset in to Train set(80%), Valid set(10%) and Test set(10%) randomly
2. Word Embedding
   1. Word2vec

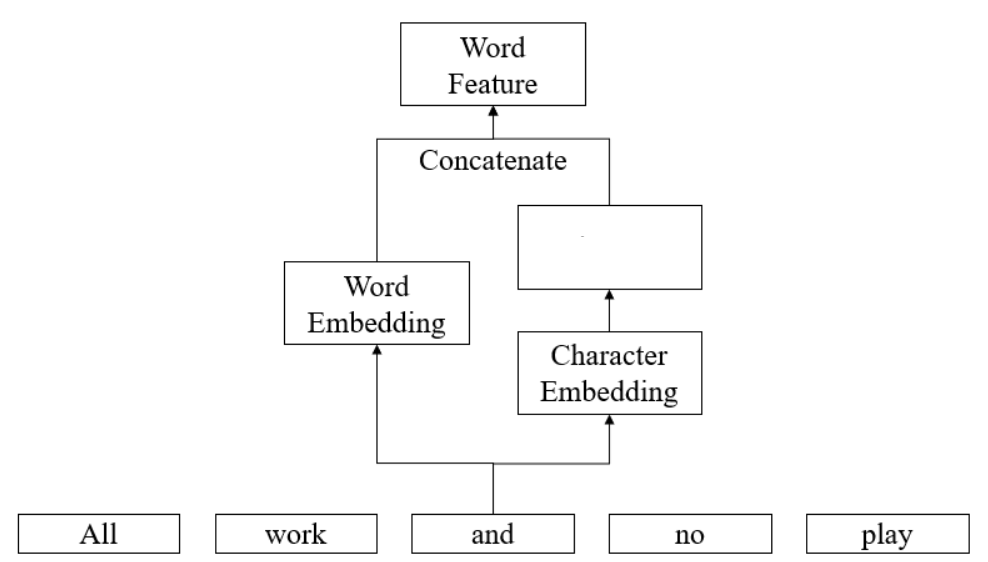
Word2vec model can obtain vectors representing the meaning of words from their surrounding contexts, using models such as Skip-gram and CBOW. It has been found that the word vectors learned by these methods have properties that are especially useful in calculating meaning, such as "words with similar meanings are close to each other" and "semantic relationships can be expressed by operations between vectors".

Japanese Wikipedia dataset as a data source which containing enough information to learn vectors of words and entities represented from the text of all Wikipedia articles. Therefore, I used Tohoku University Wikipedia Entity Vectors1 in a 200-dimension as the input of the model.

1. BiLSTM

As a branch of neural networks, RNN architectures like LSTM and BiLSTM are used in occasions where the learning problem is sequential. LSTMs and BiLSTM are popular because they have tried to learn how and when to forget and when not to using gates in their architecture. In previous RNN architectures, vanishing gradients was a big problem and caused those nets not to learn so much.

Using BiLSTM, you feed the learning algorithm with the original data once from beginning to the end and once from end to beginning. There are debates here but it usually learns faster than one-directional approach although it depends on the task.



LSTMM

**天気**

**が**

**良い**

**です**

**ね**

Figure1.The word embeddings and its character embeddings were concatenated by LSTM to represent the words.

1. ELMo

ELMo is a novel way to represent words in vectors or embeddings. These word embeddings are helpful in achieving state-of-the-art (SOTA) results in several NLP tasks. In this study, I choose AllenNLP to generate Elmo embeddings, which ELMo model roughly divided into five parts.

* + - Char Embedding
    - Char Convolution
    - Highway Net
    - Projection
    - Bi-LSTM

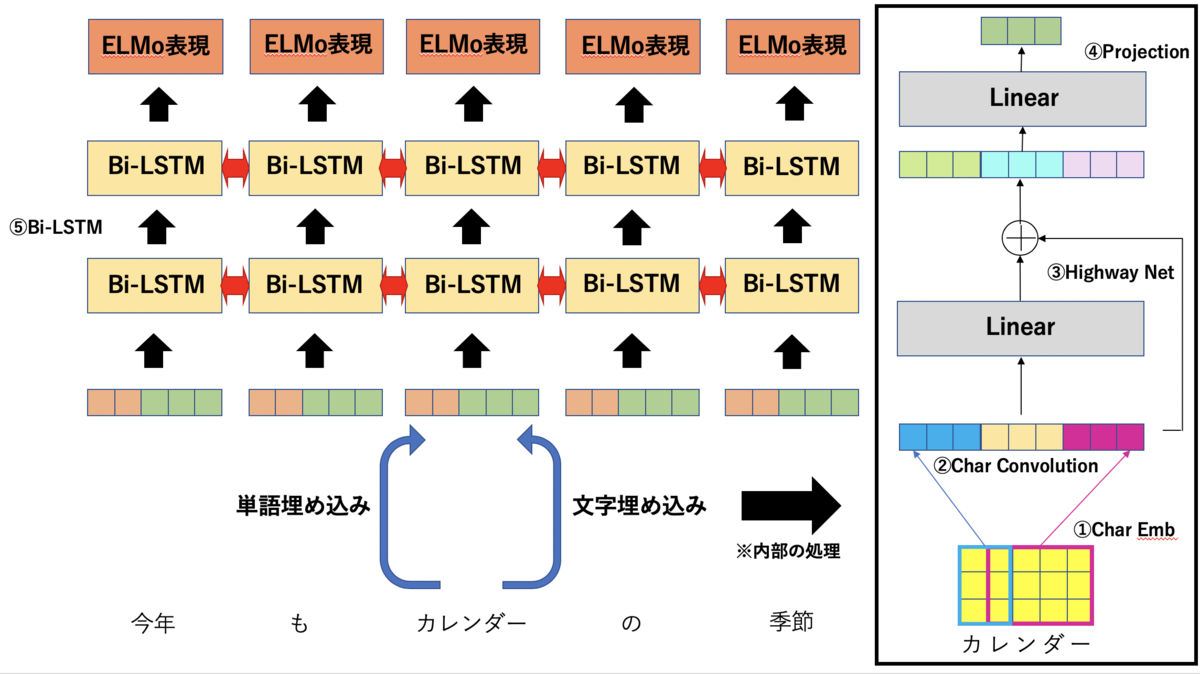


Figure2. The structure of ELMo

1. CRF

CRFs are a class of statistical modeling method often applied in pattern recognition and machine learning and used for structured prediction. Whereas a classifier predicts a label for a single sample without considering "neighboring" samples, a CRF can take context into account. As a result, I used CRF to predict the label of the words.

1. Result

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Loss | Accuracy | Precision | Recall | F1-score |
| elmo | 9.241 | 98.440% | 64.151% | 69.106% | 66.536% |
| lstm | 18.68 | 97.932% | 51.931% | 98.374% | 67.978% |

Table1. The results of the Comparison Experiments

1. Conclusion

Both models preform well, and ELMo converges quickly to the model with small losses during the 10 epochs. Although both models can analyze words with reference to context, ELMo can analyze not only location information but also semantic information compared to word2vec, which can only take location information into account, and thus its results are more closely balanced between accuracy and recall results.

1. Reference

[1] 東北大学: 日本語 Wikipedia エンティティベクトル.(2018)

http://www.cl.ecei.tohoku.ac.jp/~m-suzuki/jawiki\_vector/

[2] Yongyu Wen. (2018)

[https://github.com/yongyuwen/PyTorch-Elmo-BiLSTMCRF.git](https://github.com/yongyuwen/PyTorch-Elmo-BiLSTMCRF.git.(2018))

[3] ymym3412: 学習済みELMoをAllenNLPで読み込む -りたーんず！.(2019)

https://ymym3412.hatenablog.com/entry/2019/12/15/042818#ELMoの変換