CS224N - Final Report

Couplet Scoring and Prediction using ALBERT

Ju Huo
huoju@stanford.edu

Abstract

The Chinese couplet has more than 1000 years of history. It becomes a fun activity today to give the first line and ask about the second line. This paper would talk about a NLP model based on ALBERT to score the second line. And based on that model, we generate the candidates, score them and make predictions.

Keyword

NLP | Couplet | Machine Learning | ALBERT | Fine Tune

Introduction

What's Couplet?

The Chinese couplet refers to two complementary poetic lines adhering to certain rules, often written on red paper or carved on wooden uprights for appreciation.

The couplet has to follow a sort of certains rules. It has two equal-length lines, however the number of characters in each line can be from four to seven or more. Seven is most common, then five. The first line and the second line have inverse (or identical) tone patterns (each Chinese character is spoken as a syllable with one of four tones) — usually not closely followed. Corresponding characters must have the same lexical category (noun-noun, verb-verb, etc.).

Many don't follow this very closely. The last character of the first line is of an oblique tone, and its counterpart in the second line of a level tone.

One of the famous couplet is the couplet for the God of the kitchen (灶王爷). In the ancient myth, God of the kitchen will go back to heaven for an end year job review. The couplet for him goes like this:
First: 上天言好事
Second: 回宫降吉祥

The first line of the couplet means, say good words for me when back to heave. The second line means, bring luck to me when I go back to my kitchen.

There are one-to-one mappings like:

上(go to)	->	回 (go back)
天(heaven)	->	宫 (temple, normally represent your kitchen)
言(say)	->	降(bring)
好事 (good word)	->	吉祥 (luck)

The mapping can be either similar meaning words(好事->吉祥), same category of words (言->降) or opposite meaning words(上->回).

There is not only one correct answer for the second line if given the first line. Any sentence that matches the mappings can be a good answer for the second line. For example, if the first line is 上天言好事, 回宫降吉祥 is the most popular one. And 下界降平安 is also the well-known one. Or you can give your own one.

There is a traditional chinese fun quiz coming from couplet. People come up with the first line, and ask other people to give a second line. Beside the basic rules, people may add some fancy pattern as well, here are some complex example:

1. 上海自来水来自海上
It's a palindromic couplet.

2. 陈道明道明寺明道
It has 3 levels of meanings. First, "陈道明" "道明寺" "明道" are 3 famous actors. The second, 道明寺明道 is a palindromic. The third, it can also mean a man named "chen dao ming" understood Tao in a template named "dao ming".

3. 湖海潮落
All the characters in the first line have the same part.

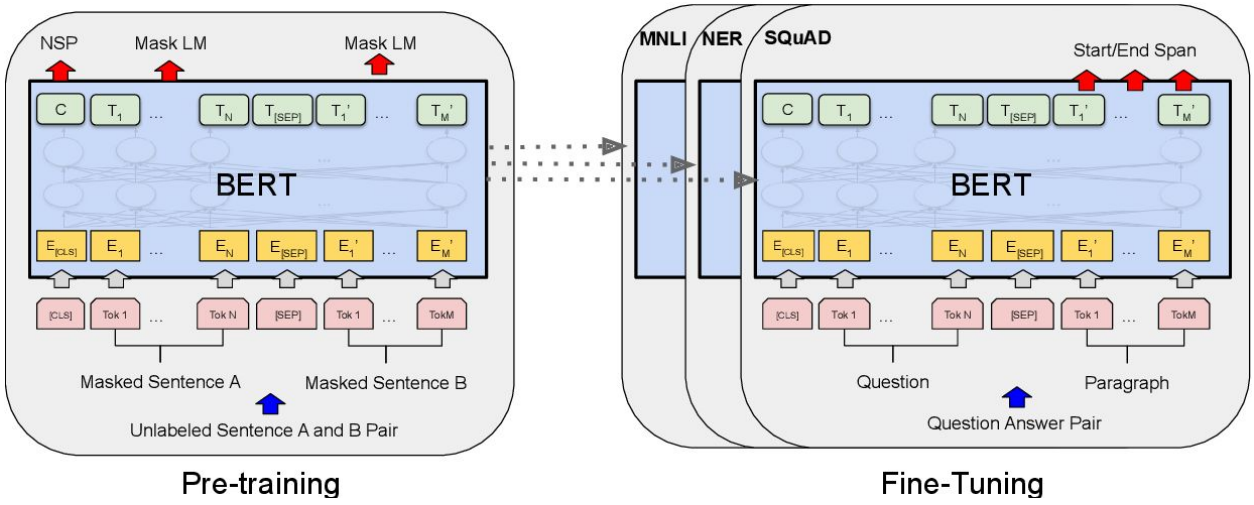
4. 霖临邻林霖麟鳞
All the characters have the same pronunciation.

It's a very interesting topic to score the second line to see if it matches the first line. And i

Model

I used Google Albert[1] + pretrained chinese model with fine tuning. ALBERT is A Lite BERT for Self-supervised Learning of Language Representations. It used transformers and self-attention techniques. And it has a great performance on the popular NLP challenge, and becomes the top 1 on leaderboards of Squad 1.0/2.0, CLUE, RACE, etc.

[1] "ALBERT: A Lite BERT for Self-supervised Learning of ...." https://arxiv.org/abs/1909.11942.



Fine tune example[2]

To fine tune the ALBERT model to score the couplet, The model will do

1. Data processing
2. ALBERT Model
3. Text Classification

In this section, I'll talk about the details about how I fine tune for the couplet scoring model.

Data processing

I got the train data from a open source couplet dataset[3]. It provides more than 70k+ couplets which are fetched from weibo.com. But all the couplets from weibo.com are positive cases, I need to generate the negative cases.

I generated negative cases by the following ways:

Positive case:
Label: 1
first: ABCDE
Second: 12345

A. Reorder the second line:
Label: 0
first: ABCDE

---

[2] "When Not to Choose the Best NLP Model - FloydHub Blog." 6 Aug. 2019, https://blog.floydhub.com/when-the-best-nlp-model-is-not-the-best-choice/. Accessed 14 Mar. 2020.
[3] https://github.com/wb14123/couplet-dataset.git

Second: 14523

B. Random word Replacement
Label: 0
first: ABCDE
Second: 12378

C. Different length by adding words
Label: 0
first: ABCDE
Second: 123456

D. Different length by removing words
Label: 0
first: ABCDE
Second: 1234

E. Random second line with same length
Label: 0
first: ABCDE
Second: 56789

F. Random second line with different length
Label: 0
first: ABCDE
Second: 94763902847

For each case, I generated 10 negative cases (A*2, B*2,C*2,D*1,E*1,F*2). So the total train set will be 770k cases.

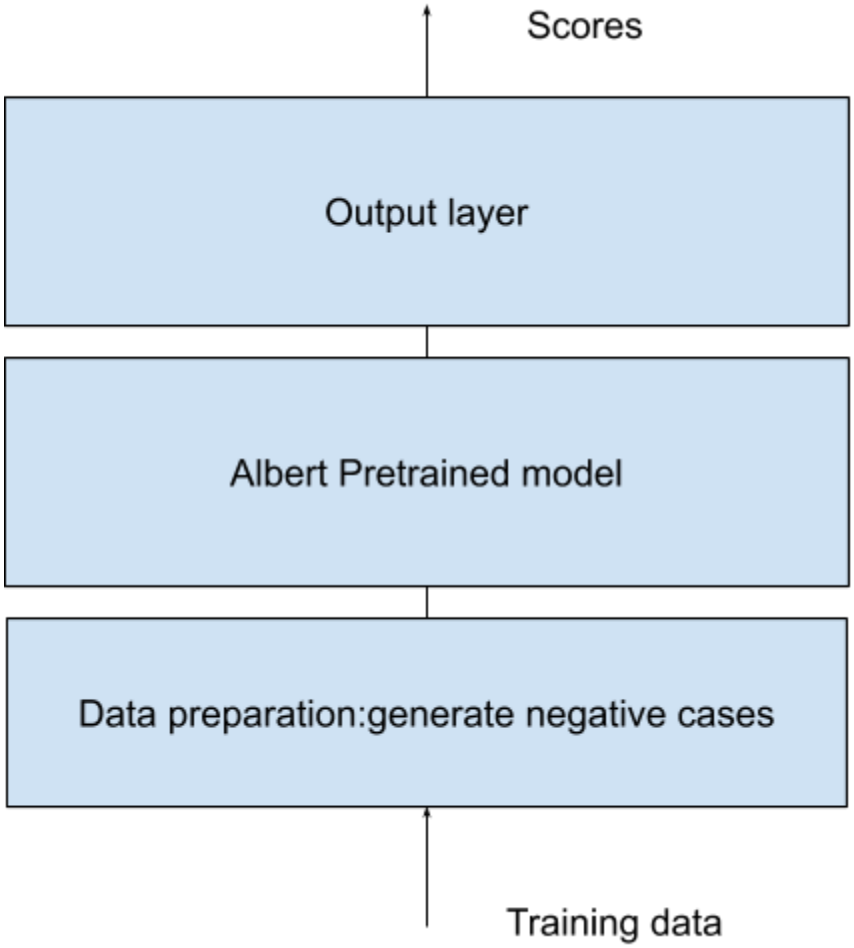
Tokenization

Since Chinese couplets have a long history, people normally use mid-ancient chinese grammar to write couplets. In mid-ancient chinese grammar, single characters have a high chance to be a word alone. So in this model, I just tokenize the couplets by each character.

Output layer

I added text classifier layer as output layer[4] on top of the pretrained chinese model using a non-linear sigmoid function as activation function. The output will be a possibility if the second line matches the first line. It will be a float from 0 to 1, and it can be thought of as a score of the second line.

The whole model is shown in the following diagram.



Prediction

The model can be applied for predicting the second line of a couplet by the following steps.

[4] "How to Fine-Tune BERT for Text Classification?." https://arxiv.org/abs/1905.05583. Accessed 14 Mar. 2020.

1. Candidates generation
2. Scoring Candidates

Candidates generation

Candidates generation is to generate the possible candidates for the second line. First we need to follow the pattern from the first line by doing word break.
For example:
The first line is: 上天言好事
It will be breaded into several words [“上”, ‘’天“, “言”, “好事”]. And we will find the most similar words for each word and replace them with random similar words.
For example:
上 ---------- 下 左 右 去 来 回
天 ---------- 地 空 界 宫
言 ---------- 说 曰 降
好事 ---------- 吉祥 如意 福气

So the candidates can be any combination of the similar words. For example:
下地说吉祥
左空曰如意
来空说福气
…

Scoring Candidates

After generating the candidates, I will pass them into the model to get a score. And the top scored one will be returned as the second line.

下地说吉祥 ～ 0.634
左空曰如意 ～ 0.154
来空说福气 ～ 0.218
-> 回宫降吉祥 ～ 0.937
…

Discuss

Evaluation

The output of the model will be a possibility, so here I will use the accuracy with different confidences to evaluate the model. The result show as below:

Confidence	Accuracy(Train set)	Accuracy(Dev set)	Accuracy(Test set)
95%	73.12%	70.94%	71.40%
90%	78.94%	75.13%	74.27%
80%	89.24%	87.31%	87.11%
70%	96.94%	95.48%	95.43%
60%	97.34%	97.23%	97.61%

Error Analysis

The model can understand the normal couplet but failed with some complex cases.
For example:
Example 1:
First: 江河湖海
Second: 魑魅魍魉

The output for this case is only 12.4%. I think it might because of the word embedding. ALBERT word embedding is a meaning based word embedding. But some of the couplets are following other patterns than meaning. In this case, all the characters in first line have a “水”, and all the characters in the second line have a “鬼”. Because of the word embedding, the model failed to learn this pattern.

Example 2:
First: 妈妈骑马马慢妈妈骂马
Second: 妞妞轰牛牛拗妞妞拧牛

The output for this case is 34.25%. In the first half “妈” “骂” “马” are all pronounced as “ma”, and second half “妞” “牛” ”拗“ are all pronounced as “niu”. Due to the work embedding, the model cannot learn the pattern based on pronunciation.

To improve the model, maybe we can develop other embedding methods. English has a character-level embedding and word embedding. Probably, chinese can have a stroke(笔画) -level embedding and pronunciation-level embedding. Some researchers have already explored some ways to embed chinese words. One of them is Visual Character-Enhanced Word Embeddings[5]. It used a convolutional neural network to embed chinese characters from an image of chinese characters.

Applying model on prediction

Example
First:春暖花开

Prediction
夏热草放 - [output : 0.6382]

Other Candidates
夏凉草关 - [0.243]
秋凉树放 - [0.386]
秋热草放 - [0.259]
冬冷草关 - [0.343]
……

The candidate generator is based on the ALBERT Chinese word vector. And we find the close word for each character and combine them randomly.

The generator currently cannot generate a very meaningful candidate. To improve, I think it might be good to try some other NLP model such as LSTM to direct predict the second candidate.

[5] "VCWE: Visual Character-Enhanced Word Embeddings." 23 Feb. 2019, https://arxiv.org/abs/1902.08795. Accessed 14 Mar. 2020.