COMP9318 Project
Session 1, 2017
Project: Predict Stress in English Words

REPORT

Student Name:	Boshen Hu
ZID:	z5034054
Lecturer:	Prof. Wei Wang

Features and their significance

I may not try a feature if I think it is not important enough or its effect could be covered by another existed features, there are only 9 features.

1. **part of speech (using pos_tag())**
   Significance: Part of speech may highly affect where is the stress in a word, for example, two-syllable verbs are usually stressed on the second syllable, such as ac-`cpet, while two-syllable nouns are usually stressed on the first, like `ac-cent (Zhang and Cercone, 1999).

2. **the sequence marks between vowel and consonant phonemes in a word**
   For example, DESTINATIONS:D EH2 S T AH0 N EY1 SH AH0 N Z, the pronunciation would be marked as "CVCCVCVCVCC", where a "C" means a consonant phoneme, and a "V" means a vowel phoneme. However, "CVCCVCVCVCC" in this sample would be transferred to "CVCVCVCVC" in order to reduce the feature number.
   Significance: This feature gives the computer macroscopical hints about words:
      i) how many syllables in the word
      ii) how vowel and consonant phonemes cross

3. **the consonant phoneme before the first vowel phoneme**
   Significance: It helps the first vowel to make a syllable, although it may not enough to make a syllable, it may work partially, hence it is considered to be important.

4. **all the vowel phonemes separately,** if there are no 4 vowel phonemes, just leave the place with a mark (Therefore, they are 4 features.)
   Significance: I think they are significant, simply for the reason that they are the targets, and pick them out may contribute to some potential rules which the computer is trying to find. What is more, these features can also show how many vowels in the word, which could be another significant information.

5. **the phoneme before the last vowel** phoneme (and it is not necessary a consonant phoneme)

   Significance: It is the similar reasons for the 3. . The last syllable could be as significant as the first one. Therefore, pick another phoneme out may work.

6. **the last phoneme in the word**

   Significance: This may be a repetitive job against the last vowel phoneme, but from the performance score I got, it does work. Therefore, I leave it.

Experiment and improvement on the classifier

Classifier: I choose Logistic Regression as classifier. Regarding the categorical features, I encode them by using OneHotEncoder.

Experiment: I first test the tiny test, after I got the right output i.e. [1, 1, 2, 1], I try another large test. I do not implement a K-fold cross validation, which I meant to do, for the reason that before I try it I got another list of data from the COMP9318 piazza forum (URL: https://piazza.com/class/izgb2o37nvc40x?cid=111), and I believe that the huge data may provide more difficulty to achieve a score than a K-fold cross validation. In addition, our submission system uses the whole training data to train the classifier, not (K-1)/K part of data, therefore I choose to test the classifier basing on the huge data set called "test_word.txt". I make a function "test1" to arrange the test data into features and labels, just like the training data, and compute the f1 score between the training data and test data.

Improving the classifier: The tools I used to refine the classifier are:

i) Add new features: such as the last phoneme, which helps me to improve my f1 score in the local computer from about 0.65 to 0.7

ii) Reduce features: once I am ambitious that I want to get all the vowel and consonant phonemes as features, then the computer just cannot stop from operating the script. Then, I cut unnecessary consonant phonemes out.

iii) Refine the parameters of the Logistic Regression classifier.

1) Since I do not get a problem of overfitting (still better performance in C>1), hence I make a C = 20 instead of C=1.
2) Change the solver among 'sag', 'lbfgs' and 'sag', and 'newton-cg' gives the best for me.
3) Multi_class: from 'ovr' to 'multinomial', 'ovr' is better.
4) Class_weight: Though I thought a 'balanced' method could give me a good performance, I got a better one with a manual setting class_weight = {1:0.7,2:0.5,3:1,4:2}, as I thought there could be few sample for the stress at the fourth vowel and it could be more chance at the first two vowels.

Reference:

ZHANG, JIANNA JIAN, and NICK J. CERCONE. *LEARNING ENGLISH STRESS R ULES USING A MACHINE LEARNING APPROACH*. 1st ed. Pacific Association for Computational Linguistics, 1999. Web. 12 May 2017.