Tech Review: Statistical Language Model

Xipeng Song, xipengs2@illinois.edu

Introduction

In the past decades, researchers developed many different statistical language models, and over time, the models become more and more complex and powerful. Therefore, this paper is designed to summarize different types of statistical language models and analyze the pros and cons of each.

Overview

The goal of statistical language modeling is to estimate the natural languages as accurately as possible. A statistical language model is a probability distribution over strings, which could be words in a sentence, characters, or paragraphs and documents. There are many applications of statistical language models. The most important is speech recognition, and others are machine translation, sentiment analysis, text suggestions, and so on. Thus, statistical language models are widely used in the lives of people.

N-gram Language Models

N-gram language models are the simplest statistical language models. N-gram can be defined as the contiguous sequence of n items from a given string input. An N-gram language model can predict the probability of a given N-gram within any sequence of words in a language, and an advanced N-gram language model could even predict the next word in the sentence. Here is an example equation of n-gram Language models.

$$P(w_n|w_{1:n-1}) \approx P(w_n|w_{n-N+1:n-1})$$

Figure 1. The equation for N-gram language models

N-gram models have many disadvantages but are still widely used for three main reasons. First, they train quickly. Second, they do not require manual annotation. Last but not least, they are incremental.

Class Language Models

The class-based language model is based on the N-gram language model, and it has a new class map with associated word counts or probabilities within classes allowing the word-given-class probability to be evaluated. This model is often used when the data is insufficient to generate a model from the training data. Here is part of the equation of class language models.

$$\Pr(c_1 c_2) = \frac{C(c_1 c_2)}{T} \times \frac{C(c_1)}{\sum_c C(c_1 c)}.$$

Figure 2. The equation for class language models

The advantages of class language models are that they are useful for small and medium size input, and the words will be automatically clustered. However, they do not work well when dealing with fixed phrases and multi-word expressions.

Topic Language Models

In contrast to class language models, topic language models work with large input, for example, a large collection of documents, instead of small and medium size input. This model is widely used in information retrieval, and it is useful for domain adaptation. Here is an equation for an example topic language model.

$$P(w|h) \quad = \quad \sum_t P(w|t)P(t|h)\,.$$

Figure 3. The equation for topic language models

However, there are disadvantages to topic language models. First, they are slow and they do not scale up well. Second, they need to be used with other language models since they do not gather local information.

Exponential Language Models

The exponential language models are based on the principle of entropy, which states that probability distribution with the most entropy is the best choice. This model uses the combination of n-gram models with other features. Here is an equation for the exponential language model.

$$P(w_i|h_i) = \frac{1}{Z(h_i)} \exp\left(\sum_j \lambda_j f_j(h_i, w_i)\right)$$

Figure 4. The equation for exponential language models

Conclusion

In this paper, some currently used statistical language models are introduced and each of them has a different optimal working situation. There is no best statistical language model so far and

the users should choose the one that suits the working environment so that the model would achieve the best results.

Reference

1. https://jon.dehdari.org/tutorials/lm_overview.pdf
2. http://www.seas.ucla.edu/spapl/weichu/htkbook/node220_mn.html
3. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=541121
4. https://www.cs.cmu.edu/~roni/11761/PreviousYearsHandouts/classlm.pdf
5. https://www1.icsi.berkeley.edu/ftp/global/global/pub/speech/papers/euro99-emlm.pdf