Statistical Models for Language Representation

*Rubén Dorado.

Fecha de recepción: 17 de abril de 2013
Fecha de aprobación: 2 de mayo de 2013
Pag. 28 a 39

* University of Tokyo, Master in Computer Science, Undergradute Systems Engineering, Universidad Nacional de Colombia.

Abstract

This paper discuses several models for the computational representation of language. First, some n-gram models that are based on Markov models are introduced. Second, a family of models known as the exponential models is taken into account. This family in particular allows the incorporation of several features to model. Third, a recent current of research, the probabilistic Bayesian approach, is discussed. In this kind of models, language is modeled as a probabilistic distribution. Several distributions and probabilistic processes, such as the Dirichlet distribution and the Pitman-Yor process, are used to approximate the linguistic phenomena. Finally, the problem of sparseness of the language and its common solution known as smoothing is discussed.

Keywords

Natural language processing, Statistical models, Language modeling, Computer Science.

Modelos estadísticos para la representación del lenguaje

Resumen

Este documento discute varios modelos para la representación computacional del lenguaje. En primer lugar, se introducen los modelos de n-gramas que son basados en los modelos Markov. Luego, se toma en cuenta una familia de modelos conocido como el modelo exponencial. Esta familia en particular permite la incorporación de varias funciones para modelar. Como tercer punto, se discute una corriente reciente de la investigación, el enfoque probabilístico Bayesiano. En este tipo de modelos, el lenguaje es modelado como una distribución probabilística. Se utilizan varias distribuciones y procesos probabilísticos para aproximar los fenómenos lingüísticos, tales como la distribución de Dirichlet y el proceso de Pitman-Yor. Finalmente, se discute el problema de la escasez del lenguaje y su solución más común conocida como smoothing o redistribución..

Palabras clave

Modelo estadistico
Informática
Manejo del lenguaje
Procesamiento natural del lenguaje

1. Introduction

This article presents a review of the most successful statistical models used to represent natural language. The purpose of a statistical language model is to estimate a distribution of different natural language phenomena to be used in language applications and technologies such as translation machines, speech recognition tools, language generation tools, and topic assignment, among many others. The phenomena to be represented can be of different sort of types depending on the nature of the problem. Commonly, the models are used to represent words in sentences or documents that not necessarily are in a sequence.

More formally, given a sequence of n words $s=w_1w_2w_3...w_n$, a statistical language model estimates a probability distribution $p(s)$. The sequence of words s can represent a sentence, a document, spoken utterances or other linguistic phenomena. Linguistic information can be added to the model, in such case we can write p(s,Θ), where Θ represents a set of features such as linguistic information in the form of part-of-speech, morphology or even structural information like preceding words, following words or syntax.

Estimating the density $p(s,\Theta)$ is extremely useful since once a model has been trained or learned, it can give information to determine the probability of previously unseen sentences. For example, given a trained model M, it is possible to determine the probability of two different sentences: S_1= {“The”, “weather”, “is”, “rainy”} and S_2= {“The”, “weather”, “is”, “salty”} to select the more plausible. Some models allow the inclusion of more information; consider for example taking Θ as topic information where each sentence can be related to a specific category C_1= “sports” and C_2= “politics”, and we want to estimate the probability of the topic and the sentence $p(s,c)$ to relate the sentence or article to the most plausible topic.

The rest of this article is organized as follows. Different language models are introduced in section 2. The main idea of each model is explained along with its principal characteristics such as the methods for acquiring the parameters or training. In this chapter, three of the main families of models are introduced: n-gram models, exponential models and Bayesian models. In section 3, one of the main problems of statistical modeling of language, the great sparseness of language, is taken into account. Finally in chapter 4, the conclusions are presented.

2. Language models

A statistical language model estimates a probability distribution of a set of words over all possible sentences. This set of words does not necessarily have to be in a sequence, it depends on the model and the purpose of its use. Thus, a language model is a probability distribution over sentences $w_1w_2w_3...w_n$:

$$p(w_1w_2w_3...w_n;\Theta) \qquad (1)$$

where Ɵ represents additional information given to or required by the model such as part-of-speech, categories, syntax, context, semantic information, distributional parameters or hyper-parameters, etc.

Three different statistical models are stated in the next subsections. First, n-gram models are explained.

2.1 N-gram models

N-gram models are the simplest method to represent language statistically. N-gram models use the *n-1* preceding words to construct a probabilistic Markov model. This way, equation (1) can be transformed into a more convenient way using the probability theory to represent the joint probability as the combination of the conditional probability of words by making use of Bayes' rule:

$$p(w_1w_2w_3...w_n) = p(w_1)p(w_2|w_1)$$
$$p(w_3|w_1,w_{2-1}) \dots p(w_n|w_1,w_2,..., w_{n-1}) \quad (2)$$

The main problem of this model is that although this representation is extremely simple, the number of conditional probabilities to be calculated is too high. The number of parameters required by the previous model with a corpus that contains a vocabulary of $V=400$ words. In order to calculate the probabilities for the bigram model, it is required to store $V^2=16000$ counts. It can be clearly seen that this number is exponential and it would be practically impossible to store such number of counts for full sentences.

The simplest approach to deal with this problem is to simplify this representation using a fixed value for n. For example, unigram models use $n=1$. This means unigram models do not take into account the history or preceding words. In such case, equation (2) becomes:

$$p(w_1w_2w_3...w_n) = p(w_1)p(w_2)p(w_3) \ldots p(w_n)$$

Bigram models use $n=2$ and the model turns into:

$$p(w_1w_2w_3...w_n) = p(w_1)p(w_2|w_1)p(w_3|w_2) \ldots p(w_n|w_{n-1})$$

The acquisition of the probabilities for n-gram models is quite straightforward. There is, however, a problem when calculating the probability of a sentence. N-gram models suffer of sparseness of the data. In other words many of the possible n-grams will not appear in the body since the combination of possibilities is too high. This is cumbersome because when calculating the probability of a particular sentence $p(s)$, it is possible that an n-gram does not appear, and in such case, the probability of the n-gram is 0, so it is the probability of the whole sentence.

2.2 Exponential models

Exponential models or Maximum Entropy was first introduced to NLP area in. Exponential models allow including other features into a statistical model such as morphology, preceding words, following words, part-of-speech among other possible options. This is possible by defining functions in the form $f_t(w;y)$ where y can be any linguistic feature such as a part-of-speech tag, a word or a morpheme. The function $f_t(\cdot)$ should return to 1 if the pair $w;y$ exists, or 0 if it does not.

An exponential model of the following form can be used:

$$p(w|y) - \frac{1}{Z(h)} * \exp\left(\sum_i \Theta_i f_i(w;y)\right)$$

where Θ_t are the parameters of the model and $Z(h)$ is the normalizing term.

There are several options in order to train these models. One of them is to use a maximum likelihood approach. Another option related to ML is to use a maximum entropy setting.

2.3 Bayesian models

A great amount of current research in statistical modeling of language has been done in this area. Most of the models use informative or non-informative priors and hyper-priors to represent language complexity. In the field of topic acquisition, one of the most successful works is the LDA. Other studies are focused on the Pitman-Yor process, both explained as follows.

2.3.1 Latent Dirichlet Allocation

The motivation of the LDA model is to represent a set of topics that are related to the words that appear in the content of a specific document. According to the model, each word has a probability to appear in the document and is related to a topic. In other words, the probability of a word changes according to the topic. In the LDA model, each document is considered as the result of a generative process that consists of the following steps:

1. Choose $N \sim Poisson(x)$, where N is the number of words in the document.
2. Choose $\Theta{\sim}Dir(\infty)$, where ∞ is a hyperparameter for the Dirichlet distribution and Θ is a k-multidimentional variable that determines the proportion of the k topics in the document.
3. For each of the N words w_n:
   a. Choose a topic $Z_n{\sim}Mult(\Theta)$, where the topic Z_n is selected from the multinomial distribution with parameter Θ.
   b. Choose a word W_n from $p(w_n|zn,\beta)$, a multinomial probability conditioned on the topic Zn and a hyperparameter β.

The LDA model is related to the mixture of unigrams in two different ways. First, it extends the topic distribution over the document allowing it to have multiple topics per document through a k-mixture of topics Z that have an associated probability distribution Θ. Second, there are assumed Dirichlet priors for the mixture of topics controlled by parameter /alpha and for the per-topic word distribution mixture with a parameter β.

The joint distribution of a topic mixture Θ, a set of topics z , and a set of words w given the parameters ∞ and β is represented by:

$$p(\Theta,z,w|\infty,\beta)=p(\Theta|\infty)\prod_{n=1}^{N} p(zn|\Theta)$$

$$p(wn|zn,\beta$$

This distribution can be marginalized over w to obtain the probability of a specific document by integrating over the mixture of topics and summing over z:

$$p(w|\infty,\beta)=\int p(\Theta|\infty)(\prod_{n=1}^{N}\sum_{z_n} p(z_n|\Theta)p(w_n|z_n,\beta))d\Theta$$

The first use of this model is the automatic acquisition of a set of words that are topic related. Although this is the main purpose, it can be easily extended in several ways. For example, it can be extended to try to model relation between topics and the words that compose them.

2.3.2 Pitman-Yor Processes

Other Bayesian model that has been strongly studied in recent years is the Hierarchical Pitman-Yor Process Language Model (HPYLM). Is a Bayesian language model based on the Pitman-Yor Process, a stochastic process which sample path is a probability distribution.

This process is more suitable than any other prior distribution for language modeling because of its ability to generate power-law distributions. Power law distribution is appropriate for modeling words since the probability distribution of words resembles it. Few words occur with a high probability while most of the words have a very low probability to be selected.

The Pitman-Yor process, noted by $PY(d,\Theta,Gb)$, is a distribution over distributions, where d is a discount parameter, Θ a strength parameter, and Gb a base distribution that can be understood as a mean of draws from $PY(d,\Theta,Gb)$.

The generative process to obtain a particular $G \sim PY(d,\Theta,Gb)$ is usually compared to the Chinese restaurant metaphor. This consists of imagining a Chinese restaurant with an infinite number of tables, each with an infinite number of seats. Clients start to enter and sit where they want according to the number of clients on the other tables, but the more people the better. Specifically, each client sits at an occupied table with probability proportional to $c_k - d$, or at an unoccupied table with probability proportional to $\Theta - dt$, where c_k is the number of clients at table k and t is the current number of used tables.

In the context of language modeling, it can be considered a vocabulary V with $|V|$ word types. Let $G0(w)$ be the unigram probability of a particular word w, and

$G_0 = [G_0(w)]_{wew}$

3. Smoothing

One common problem when modeling language is the large sparseness of the data. Even though a large train set is used, some words occur few times and the model cannot learn enough parameters. In the case of a bi-gram model, suppose a word w_t only appears one time. This means the model only can learn one bigram related to w_t. Any other combination would result in a zero probability for other possible combinations. The second problem is called OOV words or out of vocabulary words. There are utterances that do not appear in the training data and therefore they are not going to be included in the vocabulary. When a new word appears, then the model calculates a zero probability.

Smoothing methods deals with these two problems by trying to assign a non-zero probability to both events. A large number of smoothing methods have been proposed. One of the simplest methods is to use Laplace smoothing, also known as additive smoothing. In the case of a bigram model we can calculate $p(w_t|w_{t-1})$ with Laplace smoothing as:

$$p(w_i|w_{i-1}) = \frac{1 + c\ (w_i|w_{i-1})}{\Sigma_{w1}[1 + c\ (w_i|w_{i-1})]}$$

It can be seen that one for each word is added in the vocabulary, avoiding a zero probability for zero count of bigrams. Other method of smoothing includes Good-Turing [3], Backoff and Interpolation.

4. Conclusions

In this paper, several statistical models to represent language are presented. The model to choose should depend on its propose. If a simple representation of the structure is wanted, the most suitable model is the n-gram model. Exponential models are the best choice when adding other kind of features such as part-of-speech or morphology. In the case that a more powerful model is required, Bayesian models should work fine. However, since some of them are currently under research, it is kind of difficult to find developed tools or the implemented models.

Some models such as the Pitman-Yor based models, avoid the sparseness. However, other models cannot deal with it and require some method to avoid zero probabilities such as smoothing.

5. References

Berger, A. Della Pietra, S. y Della Pietra, V. A *Maximum-Entropy Approach to Natural Language Processing.* 1996. Computational Linguistics.

Blei, A. Y. Ng, y M. I. Jordan. *Latent Dirichlet allocation.* Journal of Machine Learning Research, 3:993–1022, 2003.

Church K. y Gale, W. (1991) *A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams*. Computer Speech and Language.

Teh, Y.W. (2006) *A hierarchical Bayesian language model based on Pitman-Yor processes*. Proc. 21st Int. Conf. Comput. Linguistics.

Teh, Y.W. Jordan, M.I. Beal, M.J. et al. (2006). *Hierarchical Dirichlet processes*. Journal of the American Statistical Association, 101.