UDK 541.628

Modeling a modern POS tagger using HMM and Viterbi Algorithm

Shatornaya A., Vorobiev A. nasbka@inbox.ru
JSC "OKEANPRIBOR"

The paper presents a modern way of an improved POS tagger which was modeled by using Hidden Markov Model and Viterbi Algorithm. The algorithm provides ideas for handling unknown words. This issue is caused by such words that have the same spelling but relate to complete different parts of speech and have various meanings. The accuracy of such POS tagger exceeds 96%. Such algorithm's accuracy was analyzed to other existing algorithms and suggestions were provided.
Keywords: HMM, Viterbi Algorithm, POS tagger, part of speech tagging

Моделирование современного определителя частей речи при использовании скрытой Марковской модели и алгоритма Витерби

Анастасия Шаторная, Александр Воробьёв
nasbka@inbox.ru
ОАО «ОКЕАНПРИБОР»

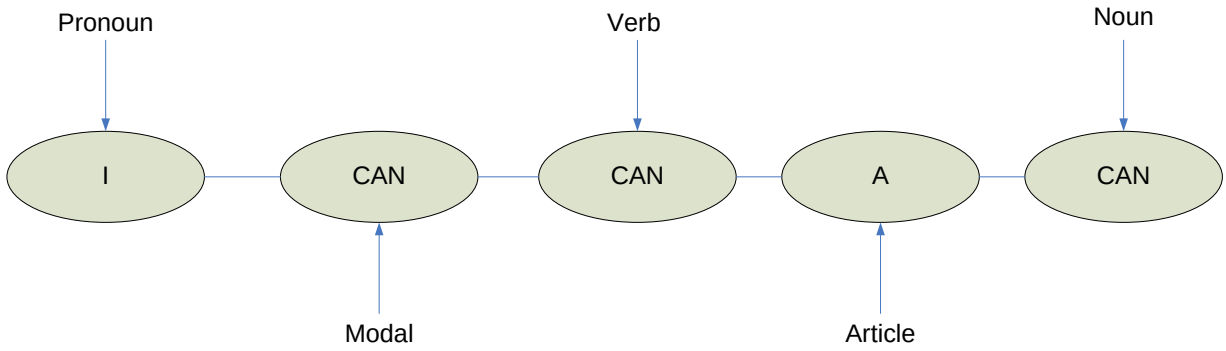
Данная статья описывает разработку улучшенного определителя частей речи за счет использования скрытой Марковской модели и алгоритма Витерби. Данный алгоритм предоставляет идеи того, как неизвестным словам, которые не находятся в банке слов, может быть присвоена определенная часть речи с высокой вероятностью. Данный вопрос возникает в связи с тем, что достаточно большое количество одинаковых слов (по правописанию) имеют разный смысл и относятся к различным частям речи. Точность определения превышает такого алгоритма 96%. Был произведен анализ данного алгоритма с другими имеющимися алгоритмами и также были предложены идея по улучшению производительности алгоритма.
Ключевые слова: скрытая Марковская модель, алгоритм Витерби, определитель частей речи.

In this paper, we present the technology of using Hidden Markov Models (HMM) that we believe is very effective for creating a part-of-speech (POS) tagger. Accuracy of the POS tagger using this technology exceeds 96.3%. We discuss the case of handling unknown words.

1. Introduction

Part of speech tagging is the process of assigning to each word its correct part of the speech, based on both its definition, as well as its context i.e. relationship with adjacent and related words in a phrase, sentence, or paragraph [1].

The biggest difficulty of POS tagging is the fact that a single word can have multi part of speech tags assigned to it depending on the context of its usage. Let's consider the following example:



To overcome this problem, there are two solutions:

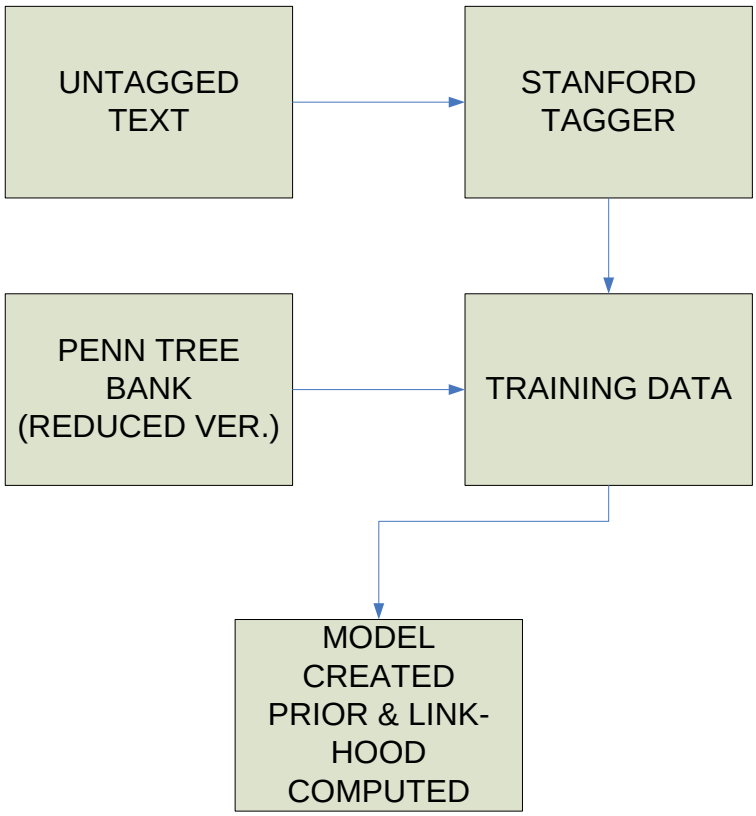
1. Some tag sequences are more likely comparing to other tag sequences. For example, the probability of having a verb is more likely than using an article after a modal [2].
2. For specific word, probability of a certain tag is more comparing to other tags. The word “Dog” is most commonly used as a pronoun other than using as a verb [3].

By combining a probability of these two factors we were able to predict part of speech tags with higher accuracy.

2. Architecture

2.1. Dataset

There have been two datasets used to train the model. First, we used a reduced version of the Penn Treebank that is freely distributed with the NLTK POS tagging library. Moreover, we also tagged a large amount of data spanning many different knowledge domains using the Stanford POS tagger. The Stanford POS tagger was used in order to more accurately train the model onsidering the fact that some of it had to be kept aside as testing data [4].



2.2. Hidden Markov Model

Hidden Markov Models are generative models, in which the joint distribution of observations and hidden states, or equivalently both the prior distribution of hidden states and conditional distribution of observations given states is modeled. HMMs have found a wide range of usage among machine learning applications [5]. These applications also use different techniques including but not limited to Markov Decision Processes, Partially Observable Markov Decision Processes, Markov Chains etc [6].

The four main components of the Hidden Markov Model are the following:

1. States: in this case the tags that are assigned to a word are the states. There are 36 tags used in this model.
2. Initial Distribution: It is the initial distribution of the probability of the observed states which are in our case a set of 36 POS tags.
3. Emission Probability: The conditional distribution of observations given some states is called emission probability. In this model, It is the probability of a particular word given a tag: P(word|tag) [5,7].
4. Transition Probability: The conditional probability distribution of the hidden state is called Transition Probability. In this model, P(tagt|tagt-1) is the transition probability [7].

In this model, sentences as a sequence of tokens (both words and punctuations) were used where the probability of a taken taking on a certain speech depends on what part of speech we assign to previous word making it a 1^{st} order HMM. The most probable sequence of POS tags given a sentence is obtained by using the Viterbi algorithm [8,9].

2.3. Training and model creation

During the staining set, the emission probability of a token and its features as well as the transition probability of a tag given the tag assigned to the previous token must be calculated. The maximum likelihood estimates of all three are provided below. Here, feature ij indicates the j^{th} feature of the i^{th} token. The details regarding the features can be found in the subsequent sections. This model currently contains around 130,000 unique tokens.

$$P(w_i|t_i) = \frac{Count(t_i,w_i)}{Count(t_i)}$$

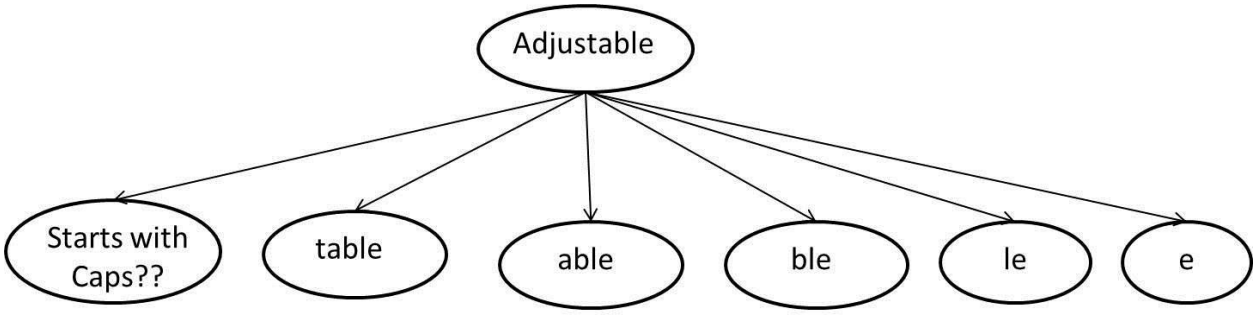
$$P(t_i|t_{i-1}) = \frac{Count(t_i,t_{i-1})}{Count(t_{i-1})}$$

$$P(feature_{ij}|t_i) = \frac{Count(t_i,feature_{ij})}{Count(t_i)}$$

2.4. Handling unknown words

There are a potentially infinite number of character sequences for a single token that the POS tagger can encounter and it is impossible for the training step to account for the emission probabilities for each of them. Handling unknown words is a similar process to detecting whether a certain email should be considered as ham or spam [10]. An email may appear to seem ham, although it was sent by someone who is unknown. As mentioned above, it is a very tough task to determine whether the email follows legitimate purposes or not. When the POS tagger encounters previously unseen tokens, the emission probabilities for them are calculated to be zero, thereby incorrectly tagging the entire sentence. This process is in some way similar to detection of pulsed signals in uniform or non-uniform sampling [11]. When the sampling is unknown, there are different techniques to use to determine the type of sampling.

We therefore estimate the probability of an unseen token by a weighted linear combination of the probabilities of its features. There are currently six such features which include all token suffixes of length 5 to length 1 and a Boolean value indicating whether the token starts with a capital letter or not. The last feature is used to discriminate between nouns and proper nouns. An example of an unknown token being broken up into its features is given below:



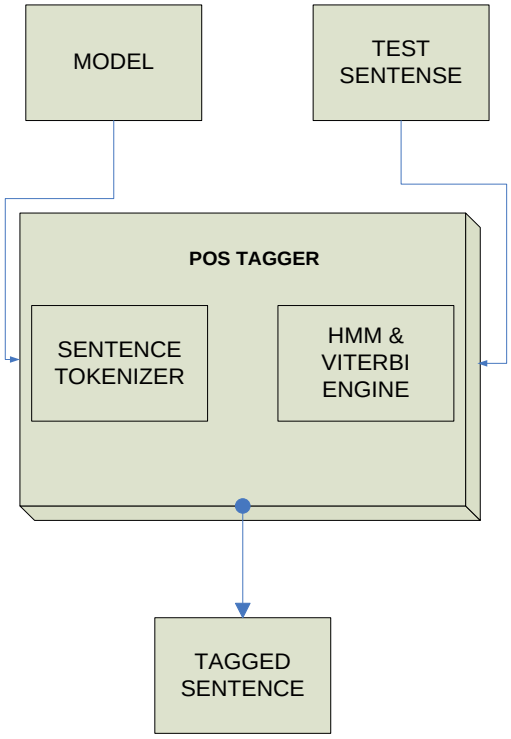
While working on this problem, we found that suffixes are very strong indicator of part of speech tag. For example, words ending with “able” are adjectives with a probability of 98% and nouns with a probability of 2%. Similarly, about 90% of all words at the beginning of a sentence and ending with “ing” were found to be gerund verbs. The equation of the probability estimate of an unknown token, given its features is given below:

$$P(word) = \lambda_1 * P(feature_1) + \lambda_2 * P(feature_2) + \cdots + \lambda_k * P(feature_k)$$

$$\lambda_1 + \lambda_2 + \cdots + \lambda_k = 1$$

2.5. Implementation

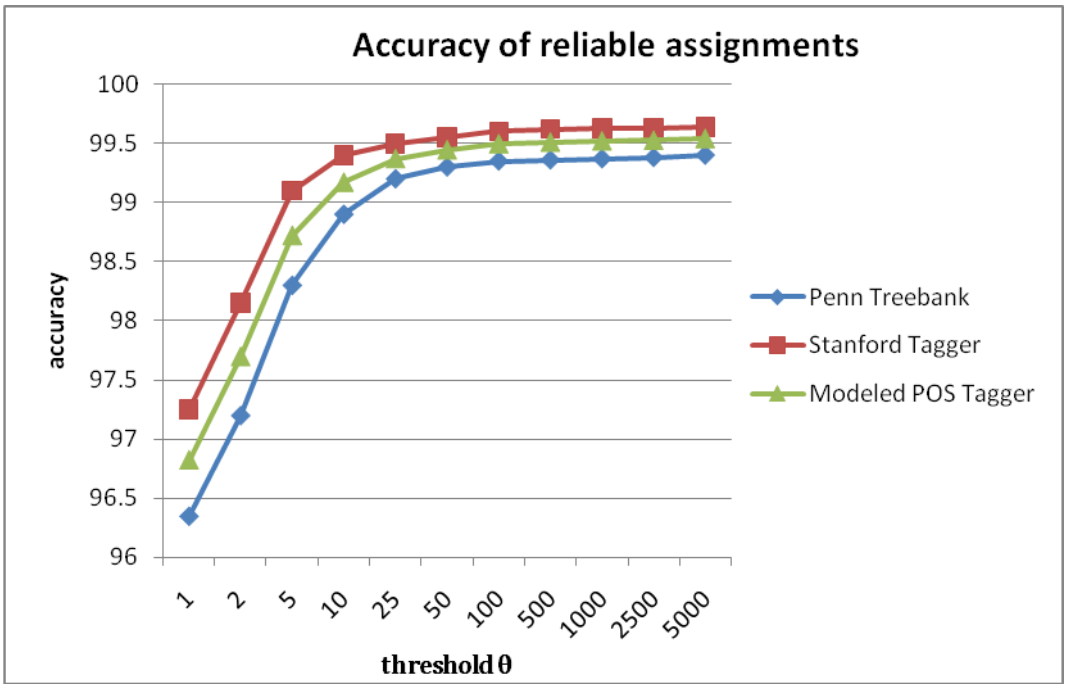
After creating a model from training date, the test sentence and model are passed to the part of speech tagger which consists of sentence tokenizer, HMM and Viterbi engine [9,12]. The sentence is broken down into tokens according to the Penn Treebank guidelines. These tokens serve as the observation nodes for the HMM of length N, where N is the number of tokens. Once the HMM is created, the most probable sequence of tags for the token sequence is generated using the Viterbi algorithm [9].



3. Conclusion

The overall accuracy of the modeled part of speech trigger is 96.35% which is high enough comparing to other speech taggers available in the market at the moment [1,2,4,7]. Although the overall accuracy is high comparing to other taggers but there has been a drop detected when gathering data from Penn Treebank.

The graph below shows the accuracy for reliable assignments for Penn Treebank, Stanford Tagger and a modeled tagger.



Bibliography:

1. Brill E. A simple rule-based part of speech tagger //Proceedings of the workshop on Speech and Natural Language. – Association for Computational Linguistics, 1992. – С. 112-116.
2. Daelemans W. et al. MBT: A memory-based part of speech tagger generator //Proceedings of the Fourth Workshop on Very Large Corpora. – 1996. – С. 14-27.
3. Cutting D. et al. A practical part-of-speech tagger //Proceedings of the third conference on Applied natural language processing. – Association for Computational Linguistics, 1992. – С. 133-140.
4. Brill E. Some advances in transformation-based part of speech tagging //arXiv preprint cmp-lg/9406010. – 1994.
5. Smirnov A. Artificial Intelligence: Concepts and Applicable Uses. – 2013.
6. Smirnov A. Creating utility-based agent using POMDP and MDP //Ledentsov Readings. – 2013. – С. 697.
7. Carlberger J., Kann V. Implementing an efficient part-of-speech tagger //Software-Practice and Experience. – 1999. – Т. 29. – №. 9. – С. 815-32.
8. Ratnaparkhi A. et al. A maximum entropy model for part-of-speech tagging //Proceedings of the conference on empirical methods in natural language processing. – 1996. – Т. 1. – С. 133-142.
9. Forney Jr G. D. The viterbi algorithm //Proceedings of the IEEE. – 1973. – Т. 61. – №. 3. – С. 268-278.
10. Smirnov A. Describing how the data obtained from DEA services is being used nowadays// Научный журнал НИУ ИТМО. Серия «Экономика и экологический менеджмент». – 2014
11. Smirnov A., Vorobiev S., Abraham A. The Potential Effectiveness of the Detection of Pulsed Signals in the Non-Uniform Sampling. ISDA 2013 Proceedings. IEEE, 2013
12. Eddy S. R. Hidden markov models //Current opinion in structural biology. – 1996. – Т. 6. – №. 3. – С. 361-365.