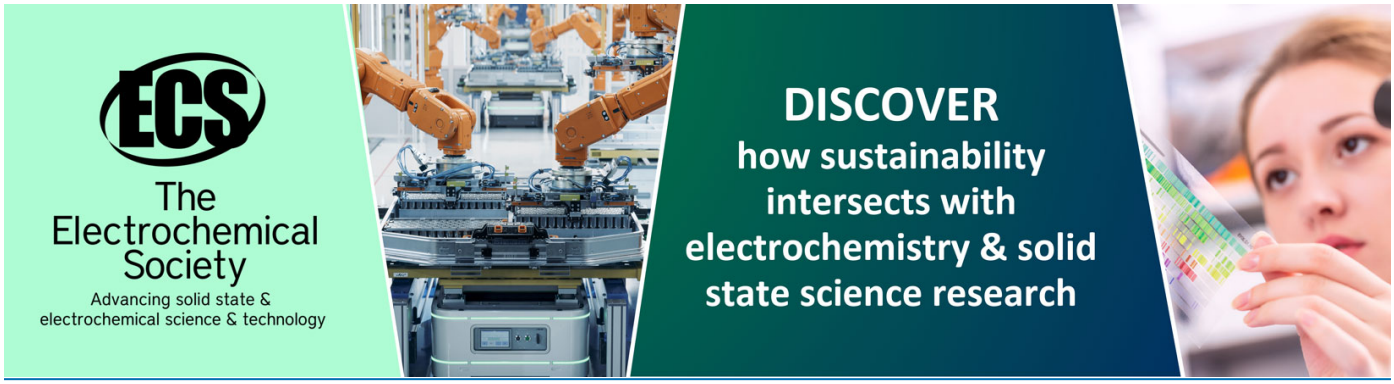
A Stochastic Approach to Identify POS in Iraqi National Song using N-Iterative HMM using Agile Approach



A Stochastic Approach to Identify POS in Iraqi National Song using N-Iterative HMM using Agile Approach

Abbood Kirebut Jassim and Boshra F Zopon Al_Bayaty

[1] Computer Science Department, College of Science, University of Baghdad, Iraq
[2] Computer Science Department, College of Science, Mustansiriyah University, Iraq

E-mail: abboodkj_comp@csw.uobaghdad.edu.iq

Abstract. This research article conveys the use of an agile approach for building an N-iterative HMM model for POST (Part of Speech Tagging) analysis. The agile model is a phenomenon or approach which has vast application. The implementation of such an iterative model is discussed in this paper. Most effectively, the information is conveyed with the help of the exact word we use during the communication. The sentence may not be a complete sentence or grammatically correct sentence, but the purpose of communication is served without any hurdle. Ages witnessing the evolution of the language earlier medium was sign language that might not contain any language communication rules that can be completed. Now complete language with so many tools and API is available, but the way preferred for the processing is the same (logically). The designed iterations lead to improved quality validation by using the word by word score calculation. The experiment is conducted to complete Part of Speech Tagging (POST) for Iraqi National Song (data set of translated Iraqi national song), and the stochastic approach used for completing the tagging is Iterative Hidden Markov Model. The results of the experiments convey the positive impact of the iterative approach over accuracy.
Keywords. Agile Approach, POST, HMM, N-Iterative Model, Iraqi national song.

1. Introduction

A software engineering approach to testing the system checks the difference between requirements and bridges the gap between the expected system and the actual system. This is well interpreted by the experiment where the Agile- like approach is used to check the system and improve it through a number of iterations. The hidden Markov model is the system used for experimentation purposes here. The Hidden Markov model is used to address the Part of Speech Tagging. This POST analysis is improvised using iteration by following the logic used in the agile methodology of the Software Engineering Approach. Part of speech is the most crucial property associated with the word. The POS tagging is the process of identification of Part of Speech (POS) based on their occurrence, meaning, frequency, and many other influencing factors. Applications of Part of Speech (POS) are translation, text to speech conversion, plagiarism check by article spinning, search engine empowerment. Research, copyright infringements, etc. requires a thorough analysis of POS is necessary. To check the POS against all

contents available cannot be done manually, considering time and cost constraints; that is why machine learning and artificial intelligence are necessary. The N-Iterative model works on a threshold-driven decisive approach to decide values of N. This threshold is nothing but an acceptable range of accuracy for Part Of Speech (POS) tagging.

2. Leterature surey

There are different experiments performed by the scientists in this domain POST. Applications, utility, corpus selection is different, but the crux remains the same. POST can be completed by using supervised and unsupervised modeling. Some of the essential models for POST from supervised and unsupervised modeling are as mentioned below.

A. Decision tree model

In the decision tree model, the tags are stored at a leaf node, and their occurrence is observed with respect to word each time by introducing the recursion in a tree [1], as mentioned in Figure 1. Part of Speech, whether love is a verb or noun, is decided by looking at the occurrence of the word love with its usage, as mentioned in the following decision tree.

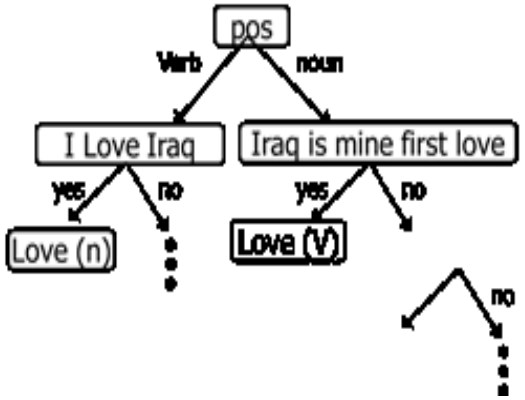


Figure 1. Decision tree model.

B. Cluster base model

Parameters specifying the cluster's information are noted and assume that these clusters have a higher chance of occurrence in a group together [2]. The logic used in cluster modeling depends on the assumption, the group of word group may probably reveal the correct POS values. For example, one should love the country that person lives in. I love my job, where I am working. Here the use of the word indicates POS of love and live along with the cluster of words. If the sentence like I love Iraq country, where I am living now, is to be used as a target object to know POS of love and live. It gives the verb as a POS based on the cluster.

C. Maximum entropy model

In the maximum entropy model, a word's entropy is calculated based on the history that earlier values [3].

D. Hidden markov model

The hidden Markov model only calculates the probabilities based on the current state [4]. These models cannot be compared on the same grounds; only the HMM is considered the HMM performance is enhanced by calculating the frequency and the iterations of their probabilities. The usage of HMM will reveal the facts that can be logically derived, but there is no refinement or repetition, so the accuracy obtained might be incidental, or we can say that there is no clear

evidence supporting the accuracy. On the other hand, in the case of the N-Iterative model, the minutes of every word is examined thoroughly, and every component's contribution is calculated. This logic strengthens the accuracy and lets us adhere to the in-depth analysis of POS based on the theorem's basic principles. Grammatical form labeling may not be the answer for a specific NLP issue. It is something that is done as a pre-essential to rearrange a variety of issues [5]. Give us a chance to consider a couple of uses of POS labeling in different NLP assignments.
For Example:
1. I live in Iraq; I love Iraq.
2. My first love is Iraq.
Input – Sentence 1 and Sentence2
Problem-
1. How to pronounce love? Phonetics - Text to Speech Conversion.
2. Whether love is a noun or a verb? To solve problems like this, Part Of Speech Analysis is required.
Types of Part of Speech Tagging
1. Language-Based Part of Speech
2. Frequency Based Part of Speech

3. Language-based part of speech
In this type of Part of Speech tagging, the complete set of Regulations are formed, and Expressions are designed. To be a complete sentence or correct sentence, the following rule related to Regular Expression is necessary [6],[7]. For example, a sentence is generally made up of subject, verb, and object. Regular Expression for any sentence in the generic form is
Sentence = Subject (Mandatory if there is no Object with a verb) +Verb (Mandatory) + Object (Mandatory if there is no subject with a verb)
Considering the same sentence mentioned above, Subject is My, Object is Iraq, and the verb is First Love is a description of the object that can be considered an Adjective.

A. Frequency based part of speech
The POS tagging is achieved with the help of an analysis of the frequency of the given the word in the context. POS's possible options are identified, and the system is trained to classify POS into these different options. Once the training phase is completed, the testing phase is initiated where the given the word's occurrence in a context is noted. The maximum frequency option is identified and declared as POS of that word [8],[9].

B. Hidden markov model
The following dimension of unpredictability that can be brought into a stochastic tagger joins the past two methodologies, utilizing both label grouping probabilities and word recurrence estimations. This is known as the hidden Markov Model (HMM).[10]
Before continuing with the hidden Markov model, let us first see what a Markov model is. That will better help comprehend the importance of the term Hidden in HMMs
The HMM concept can be explained clearly using a UML diagram called a state transition diagram, which helps interpret the relation amongst the states involved in the process. Hidden Markov model has two essential phases' emission and transition. Emission: - Emission is the observation of state, where values are noted as they can be observed. In Figure 2, the emission is indicated where the words can be observed. W1…Wn
Transition: - Transition phase is not directly explored to the user, but looking at the final state or emission state can be depicted. It is represented by t in the diagram, which deals with the tag.

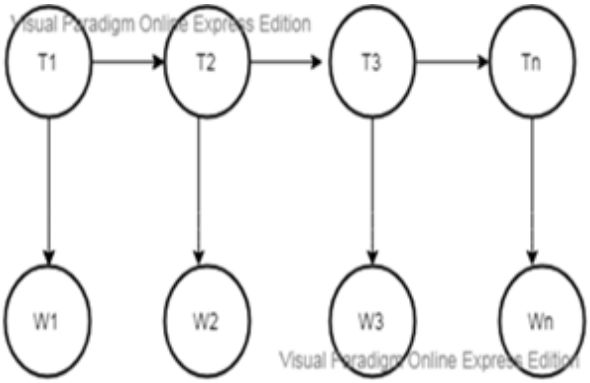


Figure 2. Markovian model for part of speech.

C. Markov mode

Say that there are just three sorts of climate conditions, to be specific

a) Happy
b) Successful
c) Satisfied

Any person living in a country like Iraq may have states like Happy, Successful, and Satisfied. Based on the information about the current states, the information about the future state can be predicted, Figure 3.

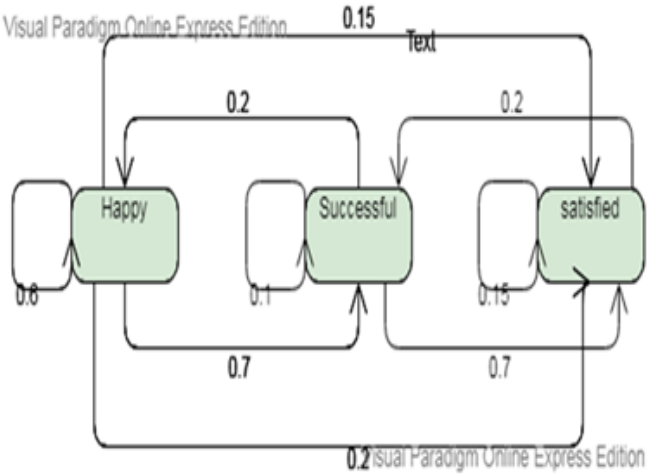


Figure 3. Markov model for the states happy, satisfied, and successful.

As mentioned in the state diagram and the table 1, three states, Happy, Successful, and Satisfied, are available. The prediction or comment about the future state can be made based on the information about the current state; this phenomenon is nothing but the Markovian model.

As for the state Happy,

If a person is happy, then the probability of being Successful is 0.2.

If a person is happy, then the probability of being Happy is 0.6

If a person is happy, then the probability of being satisfied is 0.2

Table 1. Probabilities for the states happy, satisfied and successful.

	Successful			Happy			Satisfied	
1	P(Su/Su)	0.1	1	P(Su/H)	0.2	1	P(Su/Sa)	0.7
2	P(H/Su)	0.7	2	P(H/H)	0.6	2	P(H/Sa)	0.15
3	P(Sa/Su)	0.2	3	P(Sa/H)	0.2	3	P(Sa/Sa)	0.15
	Total	1.0		Total	1.0		Total	1.0

4. Architecture of the system

As mentioned in the architecture above, the administrator provides the input of the word to be analyzed. The relevant database is referred for the analysis, frequency is calculated, and the POST process is initiated. After POST, the accuracy of the word is compared with the threshold for accuracy verification. The process is repeated with the N-Iterative model until the threshold is achieved for the word. This process is repeated for the entire database, Figure 4.

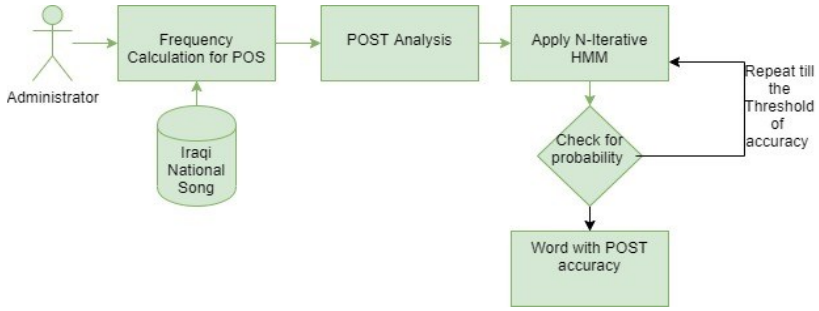


Figure 4. Architecture of the system.

5. Experimental set up

The experiment is performed to identify the possible probabilities in the "Iraqi National Song." Necessary information about the experimental set up as shown in Table 2 and Figure 5.

Table 2. Experimental set-up.

Sr. No.	Item	Specification
1.	Parser	Stanford parser
2.	Technology	JDK 1.8
3.	IDE	Eclipse 2018-09
4.	Database	MySQL 5.0 + Unstructured
5.	A groovy-based domain-specific language (DSL)	Gradle

The Part of Speech based on the frequency calculation is applied. This approach is also known as the N-Gram approach. This means calculating the occurrence of the word in each database. The Part of Speech type is denoted along with the probabilities. The example for word homeland is mentioned as below-Word: homeland Lemma: homelandPOS: NOUN
Sense Key homeland%1:15:00:: Probability: 1.0
This means the word homeland is lemmatized to the homeland. The process lemmatization removes all the extensions to the given the word like ing, ed, s etc. This leads to part of Speech tagging using the Stochastic Approach.

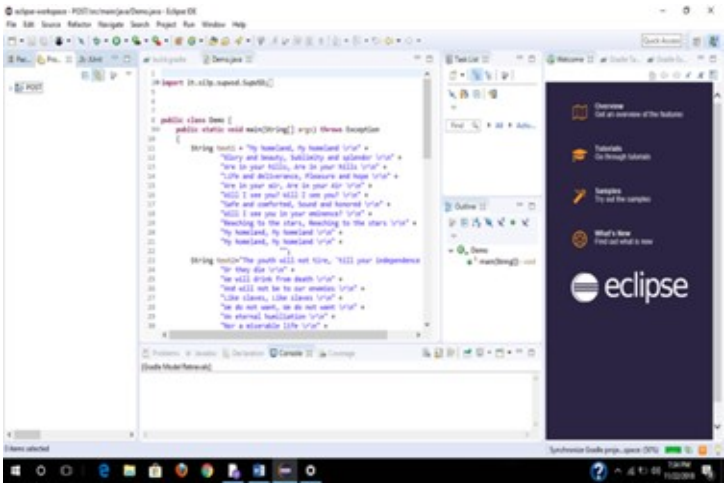


Figure 5. Screenshot of stanzas in Iraqi national song.

6. Result validation

After applying the lemmatization and Part of Speech Tagging to the database, the accuracy is improved to 97%. If the funnel is observed carefully for a given word "Moves." The accuracy is 82.22%, which is transformed from 53.91%to 82.26% based on the frequency of the word "Moves.". The algorithm's accuracy is reasonably well accepted for conducting the experiment related to Part of Speech Tagging, Figure 6.

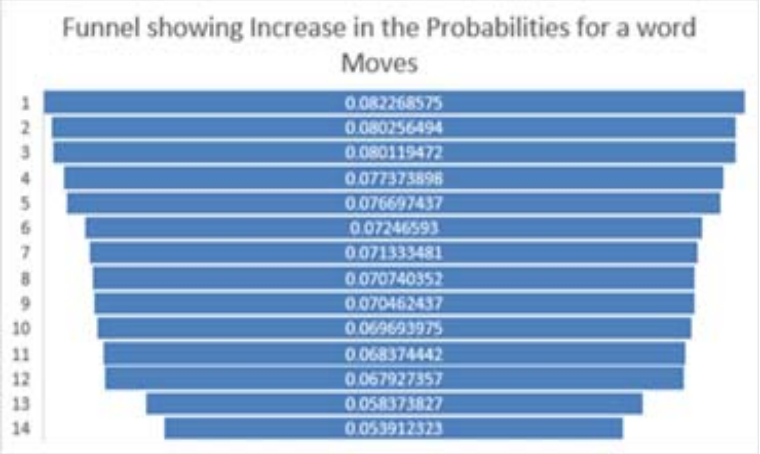


Figure 6. Accuracy of a word in terms of probability.

7. Conclusion

A novel approach is used to implement the frequency-based Part of Speech Tagging method, where the occurrence of words and tags along with sequence is measured in terms of probability. To accomplish this, the Markov model is used. Accuracy delivered by the system is considerably good. There are various applications of the experiment conducted, especially in the text to speech conversion. The future scope of this experiment is to convert speech to text and vice versa with an acceptable range of accuracy. To perform the result validation as there are no likewise standards available so the performance is compared with result processes by normal HMM are used. The authors would like to thank the University of Baghdad and Mustansiriyah University (www.uomustansiriyah.edu.iq) Baghdad– Iraq for its support in the present work.

8. References

[1] Abhijit Paul, Bipul Syam Purkayastha, Sunita Sarkar and Hidden Markov 2015 *Model based Part of Speech Tagging for Nepali language* (2015 International Symposium on Advanced Computing and Communication (ISACC))

[2] Boshra F Zopon AL_Bayaty, Shashank Joshi 2015 *Empirical Comparative Study to Supervised Approaches for WSD Problem: Survey* (International conference IEEE Canada, IHTC, Ottawa, 31 May- 4th June, 2015. (IEEE Explore paper)

[3] Kamal Sarkar, Vivekananda Gayen and V A Trigram 2013 *HMM-Based POS Tagger for Indian Languages* (Proceedings of the International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA)) pp 205–212

[4] Genoveva Galarza Heredero, Subhadip Bandyopadhyay and Arijit Laha 2011 ACM 978-1-4503- 0750-5/11/03, Copyright 2011.

[5] Mehmet Ali Yatbaz, Deniz Yuret and Coling 2010 Poster Volume, pages 1391–1398, Beijing

[6] Lichi Yuan 2010 *Sch.of Inf. Technol* (IEEE ,Jiangxi Univ. of Finance & Econ, Nanchang, China) vol 978 no 1 pp 4244–6892-8

[7] Michael Connor, Yael Gertner, Cynthia Fisher and Dan Roth 2010 *Association for Computational Linguistic*

[8] Taesun Moon, Katrin Erk and Jason Baldridge 2010 (Conference on Empirical Methods in Natural Language Processing, Proceedings of the 2010)

[9] Shane Bergsma, Emily Pitler and Dekang 2010 *LinAssociation for Computational Linguistics*

[10] Jonathan K. Kummerfeld, Jessika Roesner,Tim Dawborn, James Haggerty, James R. Curran and Stephen Clark 2010 *Association for Computational Linguistics*