ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICRITCSA

M S Ramaiah Institute of Technology, Bangalore

Vol. 5, Special Issue 2, October 2016

A Model to Predict Words in the Sentence to Identify an Aligned Sequence of Article Blocks in e-Newspaper

Deepa Nagalavi[1], M. Hanumanthappa[2]

Research Scholar, Dept of Computer Science and Applications, Bangalore University, Bangalore, India[1]

Professor, Dept of Computer Science and Applications, Bangalore University, Bangalore, India[2]

Abstract: e-Newspapers are made up of complex multi article page layout. Accordingly, each individual article is divided into multiple blocks which are not in reading order sequence. This paper proposes an approach to reconstruct the articles which includes the task of article aggregation with the English text reading order of blocks. Therefore an interpolation model is used to combine a part of speech based and a word based n-gram language models to predict the word in a sentence. This sequence probability model identifies the correct sequence of the blocks of article in English e-newspaper. Consequently, the operation is conducted by computing the probability of sequence from the given corpus.

Keywords: HMM, N-Gram, Newspaper, NLP, POS Tagging, Word Prediction.

1. INTRODUCTION

Identification of the sequence of blocks with the reading order is a challenging task. Since, the contents are divided into multiple unordered columns (blocks). A newspaper contains different articles in a page with different page layout and the position of each blocks in article is heterogeneous.

Therefore this research paper proposes a methodology to identify an individual article with correct reading order and the sequence of the blocks of an article.

The Natural Language Processing methods can be used in prediction sequence of blocks and access the likelihood of various premises. The word prediction technique is used for guessing the following word which is likely to continue a given initial text fragment. Word prediction model helps in assigning a probability to the next word in a partial sentence. This prediction is connected to the computing probability problem of a sequence of words. The statistical modeling of the language is a baseline for predicting words. In statistical modeling the selection of words is founded by the probability of a word which may appear in a corpus.

A Markov model is a finite state machine used in probability theory to model randomly changing systems where it predicts the future state depends only on the current state. It means that the next state depends only on the current state and is independent of previous history. The two models n-gram model and Hidden Markov Model are used in this work. The two models are combined together to efficiently identify the sequence of blocks in news article.

2. LITERATURE SURVEY

A survey is conducted to identify different techniques to predict the upcoming words of a sentence in different languages especially for English language. Word prediction is an essential and complex task in the natural language processing to predict the correct sequence of word to complete a sentence in a very meaningful way. Currently there are different word prediction tools like Auto complete by Microsoft, Autofill by Google Chrome, TypingAid, LetMeType etc are available [14]. These tools will predict the words while typing the text on various devices. However in this work words are predicted while reading the content from newspapers blocks with the appropriate reading order of text and sequence of blocks.

In (2007) [11] Sachin Agarwal & Shilpa Arora proposed a context based word prediction system is proposed for SMS messaging in which context is used to predict the most appropriate word for proper English words. In this work, the Hidden Markov Model for POS tag sequence and n-gram for word sequence prediction are used. These models are the machine learning algorithms to predict the current word given its code and previous words part of speech. The algorithm proposed in [11] predicts the current word after training a markov model using Enron email corpus. Whereas in (2014) [1] predicting the part of speech tag of an unknown word in a sentence for Sinhala language. The prediction is done by using Hidden Markov Model for POS tagging of text.

In (2015) [10] Xiaoyi Wu, et.al had proposed an exponential interpolation model to integrate a part of speech based language model with a word based n-gram

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICRITCSA

M S Ramaiah Institute of Technology, Bangalore

Vol. 5, Special Issue 2, October 2016

language model to complete a sentence with proper word prediction. A model based on partial differential equations is used which can mathematically model elements in the natural language processing and to improve the sentic computing methodology. Whereas in (2015) [9] Md. Masudul haque, et.al advocate an endeavor where the word completion and word prediction are two important phenomena in typing that benefit users who type using keyboard or other input devices. It will also assist user to spell any word correctly and to type anything with fewer errors. They focus on the problem when given words are not in the training corpus then the probability of the sentence will be zero for the cause of multiplication. To solve the problem a back-off method is used where for trigram model the word sequences will follow trigram probabilities at first, if it could not match then word sequence will follow bigram model and predict at least a word. If it still not matched, then Unigram probabilities are followed. Back off n-gram modeling is a non-linear method.

The number work conducted on part of speech and n-gram models previously used has an extensive impact on the current system. Consequently in many cases each one of these language models explores and captures separately specific phenomena of natural language. Part of Speech is an application which uses different kinds of information to retrieve suitable tag for each word. The different information are dictionaries, lexicons, rules and many more. It consists of two types of taggers namely rule based taggers and stochastic taggers. Therefore this work build a complex language models which capable of integrating all language components such as syntactic, semantic and morphological structures. Thus this work proposes an interpolation language model which combines word based and POS based language model.

3. PREDICTION METHODS

The work mainly focuses on the issues of combining multiple blocks of an individual article with sequence of reading order. To overcome such issues, the word prediction language model with the partial differential equation is used. In addition,an interpolation model based on natural exponential interpolation of a word based n-gram model and part of speech language model is also used. The nature of work identifies the reading order of the text in the blocks of English newspapers. It means that the sequence of the blocks are identified while predicting the sequence of the text for a sentence. The work aims at the combination of two methods one is sequence of part of speech tagging and the other one is the sequence of words in sentences to determine the word. Therefore the associated probability of both part of speech based Hidden Markov Model and word based n-gram model is checked to predict the word in sequence. As a main model the n-gram model is more effective in word prediction in a sentence. However the combination with POS based language models provides additional information to predict the sequence of words, sentences and the blocks of articles in English Newspapers.

3.1) **Word based Language Model:**

An n-gram model is a contiguous concatenation of n items where it identifies the next word from a given sequence of text. Consequently, this is one such type of probabilistic language model for predicting the next item in the form of a (n-1) order Markov model. The advantages of n-gram models are simplicity and scalability with larger value n. The n-gram model can store more contexts with a well understood space, time tradeoff enabling small experiments to scale up efficiently.

In an n-gram model the succeeding words (Wn) are predicted with a given context (W_1, W_2,…W_{n-1}) to calculate the probability function $P(W_n \mid W_1 \ldots W_{n-1})$ by using Bayes theorem

$$P(w) = \prod_{i=1}^{n} P(Wi \mid W1 \ldots\ldots Wi-1)$$

Formally n-gram model is denoted by $P(W_i \mid W_1 \ldots W_{i-1}) \approx P(W_i \mid W_{i-(n+1)} \ldots\ldots W_{i-1})$ when n=1 called as unigram, when n=2 called as bigram and when n=3 called as trigram model. The probability of words are identified from a training corpus to construct the model. The implemented n-gram prediction algorithm assumes that one can predict the next word in a phrase based on the previous n-1 words (Markov approximation). Thus, the probability of the occurrence of a word depends only on the previous words which is called markov assumption [5]. Bigram model is also known as first order markov model and Trigram is second order markov model. Whereas, quadrigram is third order markov model. Similarly the model is known to be an n-1 markov model which looks into last n-1 words in the past for the prediction of current word.

An unigram model considers only the fixed occurrence of the word in the news article. If a block ends with few starting letters of a word then the rest part of letters are searched in the article to build a continuation with the sequence of blocks in the article. In this model the most regular words that begin with few known letters of the word in progress are predicted. The bigram, trigram, …, n-gram models are considered whenever there is a sequence of words or the probability that each word follows with previous words in the blocks of article.

The disadvantage of word n-gram language model is its large number of parameters.[4] Another disadvantage of the word n-gram language model is its high dependence on the training corpus. Henceforth Hidden Markov Model can also be used with this model where words are also predicted based on part of speech tag.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICRITCSA

M S Ramaiah Institute of Technology, Bangalore

Vol. 5, Special Issue 2, October 2016

3.2) **Part of Speech based Language Model:**

Part of speech based Language Model is the system which predicts the next POS tag to be produced in the current sentence and it reduces the the amount of the word which is read. In other words a syntactic predictor has rights to the below sequence of words and POS tags to predict the current word. The POS tags ti-2 and ti-1 of the words Wi-2 and Wi-1 respectively are the previous words of a sentence taken to predict current word CWi.

… Wi-2/ti-2, Wi-1/ti-1, CWi

In natural language processing part of speech tagging is the process of making up a word in a text related to a particular part of speech. The POS tag is based on its definition and context of a word. Hidden Markov Model (HMM) is intimately associated with n-gram models and widely used as a computational model for language processing. HMM identifies the sequence of part of speech tagging for the words of sentences. For instance a noun followed by a verb, there after the probability of next item may be a preposition, article or noun. The use of certain major POSs such as noun, verb, adjective, etc along with inflections like gender (masculine, feminine, neuter), number (singular, plural, neuter) and person (1st, 2nd, 3rd, 1st/3rd) can generate accurate POS-based word predictors with a comparatively low-speed list of expected words [2]. Thus, an initial POS tagset was first derived by selecting the most functional POS tags corresponding to English. Eventually, the model can identify the probability of next POS tag which can be occur at the end of each block and switch the reading order to the next block while connecting the proper words to the sentence. Therefore Markov models extract linguistic knowledge automatically from large corpora and does part of speech tagging.

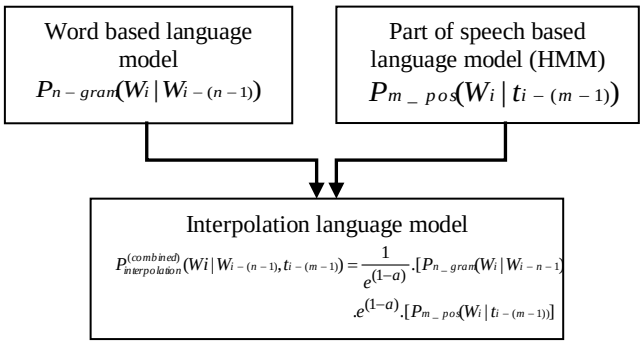
The prediction of part of speech tag can be extended to include as many previous words as desired. However in this paper we considering maximum of two previous words in the prediction model (3−POS model). The difference lies in the calculation and the tag can efficiently be predicted with previous two words tagset.

3.3) **Dictionaries:**

The general dictionaries plays a fundamental role in the natural language processing task. It contains huge number of words and sentences taken from different sources such as newspapers, magazines, academic journals etc. The classified text are needed to the text used for training the model. A large data set of text word in English is used which is collected from different newspapers and other sources. A corpus comprising of words, its POS tags and word frequencies which provides information about each word required to support all the word prediction methods. Hence the next possible words are proposed based on the probability established on the corpus provided.

3.4) **Exponential Interpolation Model**

Considering the word based n-gram model plays an important role in predictive modeling systems and can be improved in combination with a part of speech based language model[3]. The word prediction system is taken as an important element within the context of natural language processing [4]. In this regard this paper presents an interpolation language model which is determined by the exponential combination of a part of speech based language model and word based n-gram language model.



An interpolation language model integrates the benefits of each language model and assuming the language modeling as a natural system problem [3]. Here the relationship between independent models and the mathematical models are merged. The main challenge involved in modeling using differential equations, which are used to formulate the equations describing the problem from a set of limited information of the system. Since the overall goal of the interpolation model is to incorporate the benefit of two language models.

$$P^{(geometric)}_{interpolation}(Wi \mid W_{i-(n-1)}, t_{i-(m-1)}) = [P_{n\_gram}(W_i \mid W_{i-n-1})]^{a}. .[P_{m\_pos}(W_i \mid t_{i-(m-1)})]^{1-a} \quad -(1)$$

The interpolation model based on geometric interpolation has some negative characteristics that occur when, any independent language model has a zero probability causing the interpolation model to present a zero probability output. The new interpolation model (2) can be found by solving geometric interpolation model (1) [3].

$$P^{(combined)}_{interpolation}(Wi \mid W_{i-(n-1)}, t_{i-(m-1)}) = \frac{1}{e^{(1-a)}}.[P_{n\_gram}(W_i \mid W_{i-n-1})]^{a}. .e^{(1-a)}.[P_{m\_pos}(W_i \mid t_{i-(m-1)})] \quad -(2)$$

It is worth noticing that the proposed interpolation model has a form that is similar to conventional maximum entropy model. [3] It addresses the difficulty of building interpolation models and opens the way for the use of different mathematical tools to analyze the language modeling process[4].

4. IMPLEMENTATION

Word prediction language model is used to connect reading order of text in the blocks. The text each

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICRITCSA

M S Ramaiah Institute of Technology, Bangalore

Vol. 5, Special Issue 2, October 2016

individual blocks can be read but it is difficult to get the correct reading order when it has to shift to the next block which should come in sequence. Thus the system predicts POS tag as well as the corresponding word which likely should come as an upcoming word in a sentence. Therefore word prediction model combines the part of speech based model with word based n-gram model. The following algorithm identifies the correct sequence in multiple blocks of an individual article.

Algorithm:

step 1) Read blocks of news article.

step 2) Read the first block and at the end of a block read the last sentence.

step 3) Predict the upcoming word of a sentence with n-gram model. $P_{n-gram}(W_i \mid W_{i-(n-1)})$

step 4) Preprocess the block while assigning the part of speech tag for each words and retrieve the last sentence of the block.

step 5) Predict upcoming Part of Speech tag for the sentence with Hidden Markov Model.

$$P_{m\_pos}(W_i \mid t_{i-(m-1)})$$

step 6) Develop an interpolation model for step 3 and 5.

$$P^{(combined)}_{interpolation}(Wi \mid W_{i-(n-1)}, t_{i-(m-1)}) = \frac{1}{e^{(1-a)}}.[P_{n\_gram}(W_i \mid W_{i-n-1})]^{a}. .e^{(1-a)}.[P_{m\_pos}(W_i \mid t_{i-(m-1)})]$$

step 7) Identify the list of words which most likely to follow the sentence.

step 8) Match the words with the first word of the remaining blocks of an article.

step 9) If match found then read the content of the next block with the continuation to the previous block and repeat from step 3 to 9, until end of article.

step 10) Retrieve the blocks of an individual article in sequence.

Initially this work is driven by the hypothesis that n-gram models could be more effective in word prediction task. However the combination of POS based language models to the work proposes a novel exponential approach based theoretically on partial differential equations. As in many natural processes, the differential equations characterize a particular system which has been observed and, can be possible to extract relevant information about them.

5. EXPERIMENTAL RESULT

The system searches for the most probable word to identify the reading order of the news article blocks. However the probability given by the interpolation model is the list of predicted words making it to have a correctly predicted word count. In order to evaluate the word prediction system with different interpolation models, the experiments have been conducted with the text blocks of news article. In order to evaluate the proposed word prediction model with n-gram model, the experiments have been conducted with the different English newspapers.


Xiaomi buys 1,500 patents from Microsoft

PRESS TRUST OF INDIA
SHANGHAI/NEW DELHI, JUNE 1

CHINESE HANDSET maker Xiaomi has bought 1,500 patents across technologies like video, cloud and multimedia from US software giant Microsoft.

As part of the deal, it will ship smartphones and tablets loaded with Microsoft software, including Office and Skype, to buyers in India and China.

No financial details of the partnership were disclosed.

"Xiaomi and Microsoft announced today a deepening of our partnership... The compa-

"The companies' new collaboration also includes a cross-license and patent transfer agreement," Xiaomi said in an emailed response

nies' new collaboration also includes a cross-license and patent transfer agreement," Xiaomi said in an emailed response.

Microsoft sold Xiaomi nearly 1,500 patents that read on a variety of technologies from its

worldwide portfolio, including wireless communications, video, cloud and multimedia, it added.

"This agreement reflects the deep respect Xiaomi and Microsoft have for intellectual property rights. We are committed to a healthy patent system and standing against the abusive practices we've seen at times in our industry," the Chinese firm said.

Xiaomi will start pre-loading Microsoft Office and Skype on its phones starting September 2016, initially in China, with other markets including India to follow.

This will include Mi 5, Mi Max, Mi 4S, Redmi Note 3 and Redmi 3 devices from Xiaomi.

"Microsoft Corp and Xiaomi Inc have expanded their global partnership to provide innovative user experiences on mobile devices... As a result, tens of millions of consumers and business customers in China, India and around the world will have new ways to work, collaborate and communicate," Microsoft said in a statement. While Microsoft has done licensing deals with many Android device makers over the last few years, it has had tough luck with Chinese manufacturers.

Figure1: **Example**-The blocks of news article are identified with reading order.

The results are presented in Table1, considering the word-based n-gram model as a measure. The comparative improvements were evaluated and presented in Table. The result shown in table1 is the efficiency of correctly identified individual article among number of blocks which is calculated by f-measure. F-Measure is defined as the harmonic mean of precision and recall. In information retrieval system precision can be described as positive predicate value such as fraction of retrieved instances. Whereas recall is a sensitive and fraction of relevant instances. Table1 shows the experimental result of the 25 tested newspapers.

Table 1: Experimental result

News paper	Article	Blocks	N-Gram Model (%)	Interpolation Model (%)
1	6	24	52	57
2	5	20	85	86
3	6	26	94	94
4	7	21	65	68
5	8	28	82.5	87.5
6	5	22	78.5	83.5
7	9	32	91	96
8	5	20	95	96
9	8	29	90	92
10	4	20	58.5	63.5
11	6	24	98	98.2
12	7	27	63	68
13	5	25	49	54
14	6	24	87.5	92.5
15	7	20	87	92
16	8	26	87.5	92.5
17	9	27	95	95.1
18	6	24	83	88
19	7	20	91	93.2
20	4	26	88	90

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICRITCSA

M S Ramaiah Institute of Technology, Bangalore

Vol. 5, Special Issue 2, October 2016

21	5	26	93.5	98.5
22	7	24	97	99
23	8	20	89	92.5
24	6	18	87.5	92.5
25	6	20	88	87.5

The proposed method identifies the reading order of sentence from one block to other block and merger the content and extract individual article. From the table1 it is understood that the aggregation of POS based and word based models gives more accurate result than word based n-gram model.

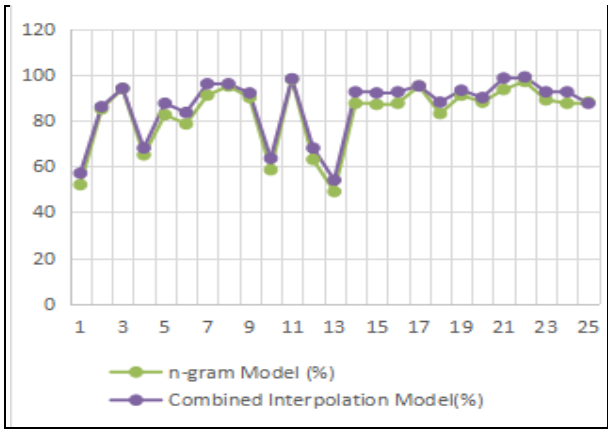


Figure 2: Graphical Representation of the result

The result given in table is further analyzed with manually identified individual article.

$$\text{Accuracy} = \frac{\text{No. of correctly identified article by the system}}{\text{Total No. of manually identified article}}$$

The accuracy is 98.8% and the 1.2% of error is due to the size of corpus and other typographic feature of newspaper such as addition of related sub article in a main article or addition of jumpline section to continue to following pages.

CONCLUSION

In this work the blocks of an individual article is identified efficiently with the reading order from English e-newspaper. To establish a connection from one block to the other, word prediction model is used. Word prediction technique refers to the systems which are used to guess the letters, words or phrases that are liable to follow a given piece of text. Thus word based n-gram language model play an important role in predictive modeling systems. Moreover the work proposes an exponential interpolation language model, that combines a word based model with a part of speech based language model to predict the sequence of words in an article blocks more efficiently. Thus the combination language model gives an accurate result than a word based n-gram model alone. It can also be improved by finding other linguistically relevant factors.

REFERENCES

[1] A.J.P.M.P. Jayaweera and N.G.J. Dias, "Hidden Markov Model Based Part Of Speech Tagger For Sinhala Language ", International Journal on Natural Language Computing (IJNLC) Vol. 3, No.3, June 2014, DOI:10.512/ijnlc.2014.3302.

[2] Carlo Aliprandi, et.al. "A word predictor for inflected languages:System design and user-centric interface", In Proceedings of the Second IASTED International Conference on Human Computer Interaction, IASTED-HCI '07, pages 148–153, Anaheim, CA, USA, 2007. ACTA Press.

[3] Cavalieri, Daniel C., Sira E. Palazuelos- Cagigas, Teodiano F. Bastos-Filho, and Mario Sarcinelli-Filho. "Combination of Language Models for Word Prediction: An Exponential Approach", IEEE/ACM Transactions on Audio Speech and Language Processing, 2016.

[4] Daniel Cruz Cavalieri, et.al. "On Combining Language Models to Improve a Text-based Human-machine Interface", International Journal of Advanced Robotic Systems,14 October 2015 DOI: 10.5772/61753.

[5] Daniel Jurafsky & James H. Martin, "N-Grams", Speech and Language Processing,chapter 4, 2014,Draft of September 1, 2014.

[6] Gerald R. Gendron, " Natural Language Processing: A Model to Predict a Sequence of Words", MODSIM World 2015, 2015 Paper No. 13 Page 1 of 10.

[7] Karl Wiegand, Rupal Patel, "Non-SyntacticWord Prediction for AAC", NAACL-HLT 2012 Workshop on Speech and Language Processing for Assistive Technologies (SLPAT), pages 28–36, Montreal, Canada, June 7–8, 2012. c 2012 Association for Computational Linguistics.

[8] Masood Ghayoomi, Saeedeh Momtazi, "An Overview on the Existing Language Models for Prediction Systems as Writing Assistant Tools", Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics San Antonio, TX, USA - October 2009.

[9] Md. Masudul haque, et.al, "automated word prediction in bangla language using stochastic language models", international journal in foundations of computer science & technology (ijfcst) vol.5, no.6, november 2015, doi:10.5121/ijfcst.2015.5607.

[10] Riya Makkaret.al., "Word Prediction Systems: A Survey", Advances in Computer Science and Information Technology (ACSIT) Print ISSN: 2393-9907; Online ISSN: 2393-9915; Volume 2, Number 2; January-March, 2015 pp. 177-180.

[11] Sachin Agarwal & Shilpa Arora, "Context Based Word Prediction for Texting Language", Conference RIAO2007, Pittsburgh PA, U.S.A. May 30-June 1, 2007.

[12] Shashi Pal Singh, et.al, "Word and Phrase Prediction Tool for English and Hindi language", International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) – 2016, 978-1-4673-9939-5/16 ©2016 IEEE.

[13] Smt.M.Humera Khanam, et al. "Mix Hidden Markov Model Based Part-of-Speech Tagging for Urdu in Limited Resource Scenario", International Journal of Advanced Research in Computer Science and Software Engineering,Volume 3, Issue 8, August 2013 ISSN: 2277 128X .

[14] Qaiser Abbas, "A Stochastic Prediction Interface for Urdu", International Journal of Intelligent Systems and Applications, I.J. Intelligent Systems and Applications, 2015, 01, 94-100 Published Online December 2014 in MECS DOI: 10.5815/ijisa.2015.01.09 .