N-gram Word Prediction Language Models to Identify the Sequence of Article Blocks in English E-Newspapers

Deepa Nagalavi
Dept. Of Computer Science and Applications,
Bangalore University, Bangalore, India.
deepatnagalavi@bub.ernet.in

M. Hanumanthappa
Dept. Of Computer Science and Applications,
Bangalore University, Bangalore, India.
hanu6572@bub.ernet.in

Abstract*: In the analysis of newspaper page, an identification of individual article is an essential task. Since the articles are the most important information unit in a newspaper. A newspaper contains variety of multiple articles with different heterogeneous page layouts. Consequently the articles in a page are divided into multiple unordered blocks. In this paper the link is established between different blocks of an article with the reading order of a sentence. It is identified with an N-Gram based linguistic processing approach for the retrieval of individual article from newspaper. The model predicts the preceding word knowing the previous content with the probability of a word sequence.

Keywords: N-Gram, Newspaper, Natural Language Processing, Word Prediction, backoff method, Interpolation.

I. INTRODUCTION

Identification of reading order is quite a difficult task for any system when the contents are divided into multiple columns. The blocks of articles in newspapers are not in sequential order moreover the position of each block is heterogeneous. Therefore, this work mainly focuses on assembling the blocks into an individual article with correct reading order and the flow of the article.

Statistical natural language processing methods can be useful in prediction and access the likelihood of various assumptions, for instance the possibility of word sequences, word co-occurrences and so on. The Word prediction technique is used in various systems such as spelling checker, mobile phone, PDA texting, hand writing recognition, word sense disambiguation, and so on. Word prediction technique does the task of guessing the preceding word that is likely to continue with few initial text fragments. Such technique is called as Markov Assumption or Markov Model. A Markov model is a stochastic model which describes the probability of sequence of possible events. This model predict the future state assuming that the future states depend on the state attained in the previous events. Ngram model follows the properties of Markov assumption.

1.1) **N-gram Language Model:**

The Model n-gram is nothing but a contiguous sequence of n items. Thus the model is used to predict upcoming words from a given sequence of text. It is based on the probabilistic language type model where the preceding word are predicted in the form of a (n-1) order of markov model. The simplicity and scalability are the two benefits of n-gram model with larger value n.

The preceding words (Wn) are predicted with a given context (W_1, W_2,…W_{n-1}) it estimates the probability function $P(W_n | W_1 \ldots W_{n-1})$ which is used by using Bayes theorem in this model.

$$P(w) = \prod_{i=1}^{n} P(Wi | W1 \ldots\ldots Wi-1)$$ **Error! No bookmark name given.**

The n-gram model is expressed by $P(W_i | W_1 \ldots W_{i-1}) \approx P(W_i | W_{i-n+1} \ldots\ldots W_{i-1})$ when n=1(unigram,) when n=2 (bigram) and when n=3 it is trigram model. The probabilities extracted from a training corpus helps to design 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).

The fixed occurrence of the word are considered in an unigram Model.. In the news article if a block ends with few starting letters of a word then the remaining part of letters are searched in the article to build a continuation of sequence of words in the article. In this model the most frequent words that begin with few known letters of the word in progress are predicted. When considering the sequence of words or the probability that each word follows with previous words in the blocks of article, there exist bigram, trigram upto ngram models.

1.2) **Dictionaries:**

In the word prediction system the general dictionaries plays a vital role to identify the probability of upcoming words. A large data set of text word in English is used which

is collected from different newspapers and various other sources. It contains a categorized text which are needed to use for training the model. The words and the information about each word contained in the corpus is required to support different word prediction methods such as word frequencies.

II. LITERATURE SURVEY

In an analysis of word prediction language model, the words are selected based on the probability of a word which may appear in a text in corpus. Subsequently, a natural language could possibly considered as a stochastic system. However the prediction language modeling is also named as a probabilistic modeling. For predicting letters, words, phrases, and sentences the model uses the statistical information the probability of a word with its distribution in different corpus. A review is directed to recognize different systems to anticipate the upcoming words of a sentence for English language.

In [8] Steffen Bickel, et.al had developed an evaluation metric while using probabilistic n-gram language model to predict the subsequent words when user enters an initial text fragments in a device. Here author uses an EM algorithm to identify the mixing weights which maximizes the generation probability of a adjacent set of sentences which are not used for training corpus. It is identified with the most likely completion, argmax$w_{t+1},..,w_{t+T}$ P(w_{t+1}, . . ,w_{t+T} |w_1, . . ,w_t). Therefore the efficiency of this algorithm grows exponentially even though the size of the search space increases in the number of predicted words. Whereas in [6] Sachin Agarwal & Shilpa Arora has proposed a Context Based Word Prediction system for SMS messaging in which the prediction of appropriate word is conducted based on the context a given tokens. It allows the short forms for proper English words. However Double Meta phone Encoding based on their phonetic similarity is used to map the informal word to its proper English words.

In [4] Md. Masudul haque, et.al had present word completion and word prediction methods which are two important phenomena in typing that benefit users who type using keyboard or other similar devices. It also helps user to spell any word correctly and to type anything with fewer errors. They focus on the problem when the given words are not present in the training corpus in such situation the probability of the sentence will become 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 unigram model and predict at least a word. Back off n-gram modeling is a non-linear method. Whereas in a paper [7] Shashi Pal Singh, et.al had proposed a bilingual word prediction tool that predicts words and phrases for both English and Hindi language. This paper describes about the word prediction tool that predicts the words in both Hindi and English, according to what the user is willing to get predictions. The tool is able to predict using multiple algorithms and also we tried to build our own bilingual database which contains phrases of variable length along with their frequency of occurrence in corpus. Whenever a word or phrase is selected from the prediction list, its frequency is increased by one. Consequently, the most frequently used word will have highest frequency/probability leading to increase in efficiency of predictions. Those words and phrases that are not present in the database but have been typed by the user are fed into the database for being used in future predictions.

III. WORD PREDICTION LANGUAGE MODEL

N-gram language model is used as a probabilistic language model where the approximate matching of preceding item is very high. The Probability of word prediction is based on counting the word in most cases where it calculates the current word depends on the previous words which is called as Markov assumption. In an ngram model unigram looks single item from a given sequence, bigram is called first order markov model, trigram is second order markov model whereas quadrigram is third order markov model. Similarly an n-gram language model is n-1 markov model which looks n-1 words into the past for the prediction of current word. Maximum Likelihood Estimation is calculated in word prediction model for the parameters of an N-gram model by getting the counts from a normalize corpus, and normalize the counts so that they lie between 0 and 1. The Smoothing algorithms provide a more sophisticated way to estimate the probability of N-grams where N-grams rely on lower-order N-gram counts through back off or interpolation methods.

A back-off method is a non-linear method where it model the word sequences which follow trigram probabilities at first, if it could not match then the sequence of word will follow bigram probabilities. If the count is again zero as of not matching then a model backoff to unigram model and predict at least a word. A method "back off" to a lower-order N-gram if count has zero evidence for a higher-order N-gram. The problem with back-off is that the probability estimates can change suddenly on adding more data when the back-off algorithm selects a different order of n-gram model on which to base the estimate.

$$P(Wn|Wn-1Wn-2)=\begin{cases} P(Wn|Wn-1Wn-2), & if\ C(Wn-2Wn-1Wn)>0 \\ P(Wn|Wn-1), & if\ C(Wn-2Wn-1Wn)=0 \\ & if\ C(Wn-1Wn)>0 \\ P(Wn), & Otherwise \end{cases}$$

In an interpolation model, it generally mix the probability estimates from all the N-gram estimators and combines the trigram, bigram, and unigram counts. The simple linear interpolation model is used to consolidate the distinctive ordered N-grams by linearly interpolating all the models. So, this work estimates the trigram probability P(Wn|Wn-2Wn-1) by combining together the unigram, bigram, and trigram probabilities.

$$\hat{P}(Wn|Wn-2Wn-1) = \lambda 1(W_{n-2}^{n-1})P(Wn|Wn-2Wn-1) + \lambda 2(W_{n-2}^{n-1})P(Wn|Wn-1) + \lambda 3(W_{n-2}^{n-1})P(Wn)$$

Identification and extraction of individual article from e-newspapers is the difficult task. The solution of this problem is identified in different phases. In the previous work of mine we have identified the initial point of the articles means that the first block of the article. Also we have grouped the blocks which belong to the same article based on semantic similarity and relatedness between blocks. In this paper the reading order of the content of article is identified between blocks using n-gram linguistic model. This algorithm calculates the probability of the next word based on the previous n terms and merges all the blocks of individual article in reading order and in proper sequence of blocks.

Algorithm:

Step 1: b=1; //read first block of article.
Step 2: article.append(blk[b]); //reading the content of the article.
Step 3: if(end of block) then read line; //last line of block
Step 4: tokens[]=line.split(" "); //split sentence into tokens
Step 5: // read last few words to find n-gram prediction.
ng[i-1]=tokens[n-1], ng[i]=tokens[n];
Step 6: N=3; //n-gram: start from trigram model.
Step 7: if(count(n_gram(N,ng,nblk))>0) then
Step 8: if (nblk==0) then set flag=1; // error in backoff
Step 9: b=nblk; // address of the block which likely to
//be come in sequence.
Step 10: article.append(blk[b]);
Repeat from step 3 until end of article.
Step 11: if(count(n_gram(N,ng,nblk))==0) then
N=N-1;
Repeat from step 7 to 10.
Step12: if(flag==1) then
Interpolation(n_gram(N,ng,nblk)
Repeat from step 9 to 10.

In this algorithm the blocks of each individual article are merged while checking the reading order of text. An n-gram model is used for the prediction of sequence of words in a sentence. In most of the articles the sentences are broken into two blocks i.e. the last line of the block is end up with half sentence and it is continued in the next block. In such situation after the end of the block the system reads few words of last line and predicts the future word with the n-gram probabilistic model. Later the word is matched with the first word of next block. Sometimes the word of a block end is itself comes in half for example "mur" in one block and the continued letter "der" in next block but the complete word is "murder" which is also predicted by ngram model.

A back-off method is applied in this algorithm where it first predicts word with trigram model and match with first word of next block. If matching is found then the two blocks are merged otherwise the system back-off to bigram model if it could not match then word sequence will follow unigram model and predict at least a word. Similarly all the blocks of individual article are merged to one text block with the reading order sequence. If there is errors with back-off method such as suppose a system identifies more than one block as the sequence of an article then there is an error. In such situation system check for linear interpolation method where it combines the probability of all three n-gram models count and predict the word with Maximum Likelihood Estimation.

IV. IMPLEMENTATION

An approach starts with a sentence fragment and come up with a word using a stochastic language model. The stochastic language models are unigram, bigram, trigram, backoff and linear interpolation. To predict the outcomes of an experiment, a built-in Uppsala English corpus is used.


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 companies' 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.

The Indian EXPRESS epaper editions Thu, 02 June 2016 epaper.indianexpress.com/c/10714969

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

Figure1 is the example taken to test algorithm. After reading the content of first block the system should connect to the 3rd block then to 4th and 5th block. At the end of first block system read few words of last line from last while segmenting the sentence. In this example it takes "the

compa" and predicts the sequence. As "compa" is not complete word the system backoff from trigram to bigram model and check the continued letters for "compa" is then the predicted word is matched with the first word "nies" of other blocks of article which is "companies". If matching is found then the rest of contents of the blocks are merged with the previous block. From the example after retrieving the content of first block it retrieves the content of third block in sequence. Similarly at the end of third block system reads last 3 words "technologies from its" then a word "worldwide" is predicted as sequence of sentence. The approximate matching of next item is identified in corpus. If any of the words are not in the corpus then the probability of the sentence will be zero. Henceforth the size of the corpus should be more so that all the word sequences can be identified efficiently.

An ngram algorithm start with trigram model where it search for the past two words in a text corpus and extracts the preceding word with the probability of maximum likelihood estimated word. If not found then it will search for bigram model and finally with unigram model. This process is known as backoff method. An algorithm also searches for word with interpolation method if there is an error with backoff method. In interpolation method it combines all three model and predict the word which likely comes in sequence.

Before the implementation of the work a study is conducted to test the models. To understand the merits of the work in predicting words in newspaper blocks the stochastic language models are compared with each other. For linear interpolation model, [4] analytically the linear weights $\lambda 1 = 0.5$, $\lambda 2 = 0.33$ and $\lambda 3 = 0.17$ are taken such that the sum of λi is 1. The probability equation for interpolation is converted into as follows:

$$P(Wn|Wn-1Wn-2) = 0.5P(Wn|Wn-1Wn-2) + 0.33P(Wn|Wn-1) + 0.17P(Wn)$$

The probabilities of few sentences are taken from news article to test the models, and the result is displayed in table1. From the table1 the trigram, backoff and interpolation language model have performed almost in the same trend-line. However the average accuracies of trigram, backoff and interpolation model are close. Henceforth all the three models are used in this work for different conditions. Consequently most of the words are predicted by interpolation method in the proposed work.

Table1: Results of models used for sentences

No of Words in Sentence	Trigram	Backoff	Interpolation
5	52.16%	52.16%	52.16%
6	62.68%	63.18%	62.68%
7	56.32%	66.32%	67.95%
8	60.05%	61.05%	62.68%
9	56.95%	62.45%	68.65%
10	69.58%	71.58%	72.58%
11	59.32%	65.32%	62.68%
12	67.45%	68.95%	69.95%
13	62.68%	63.18%	63.88%
14	50.95%	58.95%	69.95%
15	67.95%	67.95%	68.95%
16	60.05%	60.55%	60.96%
17	65.32%	65.32%	65.68%
19	59.95%	68.45%	68.82%
20	41.63%	41.63%	41.63%

Table 1 defines the results of models used for sentences of different range of words. The plots in figure 3 represents the performance of the models given in table1. The plot clearly shows that interpolation model predicts the words in sequence than other model.

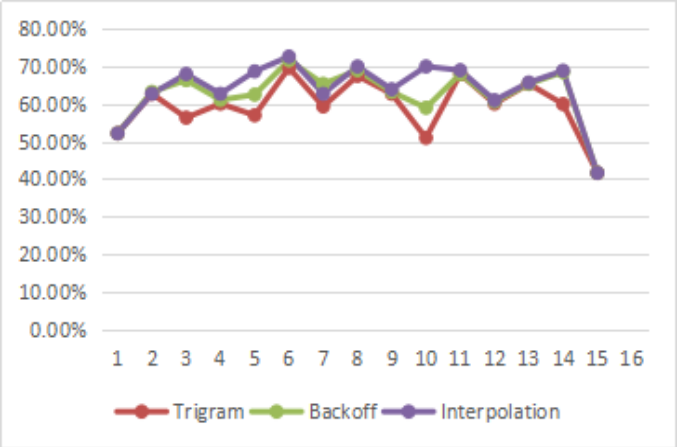

Figure2: Graphical representation of an experimental result

V. EXPERIMENTAL RESULT

The proposed method is test with 25 English newspapers of different publishers. In total it consists of around 900 pages among which approximately 6000 articles tested for experimental result. The proposed method identifies the reading order of sentence form one block to other block and merger the content and extract individual article. The result is 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 96.5% and the 3.5% of error is due to the size of corpus and the grammar. Since ngram model predict the words based on the probability of words found in corpus. Hence the efficiency can be increased by adding the part of speech tagging method to this work.

CONCLUSION

In this work an individual article is identified efficiently with the reading order of the blocks. To establish a connection from one block to the other, an ngram word prediction model is used. Ngram based word prediction works well for English newspaper. But it is more challenging task to get 100% performance as it depends only on the training corpus of large data. However to improve the word prediction task a linear interpolation model is used which combines trigram, bigram and unigram models. In this work single word prediction is conducted but a set of words or phrases can also be predicted to complete a sentence and merge the blocks of article with the matching sequence.

REFERENCES

[1] 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

[2] 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, Montre´al, Canada, June 7–8, 2012. c 2012 Association for Computational Linguistics.

[3] 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

[4] 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

[5] 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

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

[7] 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/$31.00 ©2016 IEEE.

[8] Steffen Bickel, et.al., "Predicting Sentences using N-Gram Language Models", Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), pages 193–200, Vancouver, October 2005.

[9] Xiaoyi Wu, et.al., "An Improved Hierarchical Word Sequence Language Model Using Word Association", Springer International Publishing Switzerland 2015 A.-H. Dediu et al. (Eds.): SLSP 2015, LNAI 9449, pp. 275–287, DOI: 10.1007/978-3-319-25789-1 26.