Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics
San Antonio, TX, USA - October 2009

An Overview on the Existing Language Models for Prediction Systems as Writing Assistant Tools

Masood Ghayoomi*, Saeedeh Momtazi†
Department of Computational Linguistics
Saarland University
Saarbruecken, Germany
*masoodg@coli.uni-saarland.de
†saeedeh.momtazi@lsv.uni-saarland.de

***Abstract*—The prediction task in national language processing means to guess the missing letter, word, phrase, or sentence that likely follow in a given segment of a text. Since 1980s many systems with different methods were developed for different languages. In this paper an overview of the existing prediction methods that have been used for more than two decades are described and a general classification of the approaches is presented. The three main categories of the classification are statistical modeling, knowledge-based modeling, and heuristic modeling (adaptive).**

***Index Terms*—Word prediction, Assistant technology, Language modeling.**

I. Introduction

After the Second World War, the number of people with disabilities was increased dramatically. In order to help them to communicate with the outside world, assistant technology such as word prediction was used. Researchers tried to develop systems that are alternative to the users disabilities and could augment his abilities too. The prediction systems have been in use since the early 1980s.

Prediction refers to those systems that guess which letters, words, or phrases are likely to follow in a given segment of a text. Such systems are very useful for user, specially the ones with disabilities. The systems typically operate by displaying a list of the most likely letters, words, or phrases for the current position of the sentence being typed by the user. As the user continues to enter letters of the required word, the system displays a list of the most probable words that could appear in that position. Then, the system updates the list according to the sequence of the so-far entered letters. Next, a list of the most common words or phrases that could come after the selected word would appear. The process continues until the text is completed.

The goal of all writing assistance systems is increasing the KeyStroke Saving (KSS) which is the percentage of keystrokes that the user saves by using the word prediction system. A higher value for KSS implies a better performance; as a result, decreasing the user's effort to type a text. In other words, the amount of text to be typed needs to be as short as possible for the user with the least effort. KSS is one of the important standard performance metrics to evaluate prediction systems [1], [2].

Word prediction, or in general predicting a segment to be completed, is one of the important tasks in most natural language processing applications such as speech recognition, word-sense disambiguation, context-sensitive spelling correction, statistical machine translation, handwriting recognition, and optical character recognizer. In addition, it could be used for people who are non-native speakers of a language and want to learn the language. Moreover, it could be used in recent technologies such as cell phones and Personal Digital Assistants (PDA).

There are numbers of prediction systems that were developed and are developing with different methods for different languages. In this paper, an overview of all the existing approaches used in the prediction systems will be described; also a general classification of the methods that are introduced by researches will be presented to show the progress of developing prediction systems and the place where we stand now.

The next three sections describe the existing prediction methods that are used to model the natural language since the early 1980s. In these sections three major approaches are described: statistical modeling, knowledge-based modeling, and heuristic modeling (adaptive). The paper is summarized in the last section.

II. Statistical Modeling

Traditionally, predicting words has solely been based on statistical modeling of the language [3]. In statistical modeling, the choice of words is based on the probability that a string may appear in a text. Consequently, a natural language could be considered as a stochastic system. Such a modeling is also named probabilistic modeling. The statistical information and its distribution could be used for predicting letters, words, phrases, and sentences.

A. Letter Prediction

Prediction approaches have entered into recent technologies such as cell-phones and PDAs. Letter prediction could be used as an aiding tool to enter a text on Short Message Service (SMS), to chat on Instant Message, and to write an email. Because of being portable, such systems could not have a single key for a letter. So, a text should be entered with a limited number of keys (e.g., in cell-phones, a text is written with only 9 keys on the phone). This means that a key should

978-1-4244-2794-9/09/$25.00 ©2009 IEEE

carry three or four letters. The reduced keyboard makes it hard for the user to enter a text; so, the letter prediction method would be an efficient way. The reason to have such a system is that the user will need to press only one key for each character on the mobile phone [4]. But how the letters would be disambiguated in a single key? Three methods are presented:

1) The Lexical-based Predictive Text Entry Method: One of the methods is to press one key for one character. This requires a program that matches the key sequence to the corresponding words in a lexical dictionary. In this method, the most frequent words that match with the key sequence will be presented. In fact, the method merely aims at disambiguating the sequence of letters rather than predicting them. If the key sequence corresponds to two or more words, then the user can browse through the resulting word list and choose the intended word. T9 developed by Tegic is such a system [5]. Also systems such as EziText by Zi Corporation [6] and iTAP by Motorola [7] have used this method. The keystrokes per character (KSPC) for such systems are greater than one.

2) WordWise: WordWise is developed by Eatoni Ergonomics [8]. It uses an auxiliary key. In this approach a key is selected as an auxiliary key that is pressed simultaneously with the corresponding key to the character. The auxiliary key is red and the letters that need to be pressed simultaneously by the key are also red. Key 1 is the auxiliary key, and letters c, e, h, l, n, s, t, y are red. The disadvantage of such a system is to press two keys for one character [4].

3) LetterWise: LetterWise is another method that is also developed by Eatoni Ergonomics. It considers the letter diagram probabilities. The program selects the most probable letters knowing the previous one. Since disambiguating of letters are based on the already entered characters and not on the lexical dictionary in itself, as a result the method needs a small amount of memory and it is much easier to enter new words [9]. The KSPC for LetterWise is 1.15.

B. Word Prediction

Word prediction systems guess the words that the user intends to use. These words are suggested in a list to the user that might be used in that position. The method tries to ease writing a text to the user. Statistical language modeling is widely used in these systems. Statistical word prediction is made based on the Markov assumption in which only the last *n-1* words of the history affects the next word [10]. Thus, the model could be named *n*-gram Markov model. Word frequency and word sequence frequency, generally, are the methods that are commonly used in prediction systems, especially for the ones that are developed commercially:

1) Word Frequency: The early predictive systems, specially the ones which were used as a writing aid tool in the 1980s, used merely the frequency information of each word independently to complete a word in the current position of a sentence being typed by the user without considering the previous context (the history). In other words, the systems used unigram word model with a fixed lexicon. Such systems always come up with the same prediction suggestions for a particular sequence of letters. Since solely independent statistical information has been used in this model, most of the time the suggestions might be inappropriate [1], [3]. PAL [11] is a system which uses this approach with 50% saving in keystrokes. Profet [12], [13] is another system that uses the unigram model with 26.1% KSS developed for four languages: English, Norwegian, Dutch, and French.

2) Word Sequence Frequency: Using a unigram word model for early systems, it was cleared that some of the suggestions are not appropriate in that position of a sentence; suggestions will be better if context is taken into account. As a result, researchers tried to develop systems that used history as a clue for appearance of the next words. If only the previous word was used to predict the next word in the current position of the sentence being typed, then it was named bigram word model or first order Markov model. If the last two words were used to predict the next word, then it was named trigram word model or second order Markov model [10], [14]. WordQ [15] and Gus [16] are examples of this method for English. The KSS for WordQ is 53.1% [17]. Ghayoomi and Assi [18] used a combination of un-, bi-, and trigram word model for Persian. Their system achieved 57.57% in KSS.

FinishLine [19] and another system developed by Bickel et al [20] have used the *n*-gram word model to predict sentences.

III. Knowledge-based Modeling

The systems that merely used statistical modeling for prediction often present words that are syntactically, semantically, or pragmatically inappropriate [21], [22]; then they impose a heavy cognition load on the user to choose the intended word and decrease the writing rate as a result. Omitting inappropriate words from the prediction list gives more comfort and confidence to the user. The linguistic knowledge that could be used in prediction systems is syntactic, semantic, and pragmatic that would be discussed below:

A. Syntactic Prediction

Syntactic prediction is a method that tries to present words that are appropriate syntactically in that position of the sentence. It means that the syntactic structure of the language is used. In syntactic prediction Part-of-Speech (POS) tags of all words are identified in a corpus and the system uses the syntactic knowledge for prediction [2], [1]. Statistical syntax and rule-based grammar are two general syntactic prediction methods that will be described in more detail. This method includes various types of probabilistic and parsing methods such as Markov model and artificial neural network [23].

1) Statistical Syntax: This approach uses the sequence of syntactic categories and POS tags for predictions. The appearance of a word in this method is based upon the correct usage of syntactic categories. In other words, the Markov assumption about *n*-gram word tags is used.

In the simplest method, the POS tags are sufficient for prediction. Therefore a probability would be assigned to each candidate word by estimating the probability of having this word with its tag in the current position and regarding the most

probable tags for the previous word(s). Fazly [2] experimentally gained the result of 49.8% in KSS for 5 word suggestions in the prediction list.

In another approach, the predictor tries to estimate the probability of each candidate word according to the previous word and its POS tag, and the POS tag of its preceding word(s). In other words, the system uses word bigram and POS trigram model. SyntaxPal [1], [24], [25], New Profet [26], FASTY [27], The Predictive Program [28] are systems that have used such a method. The Predictive Program model achieved 53.3% in KSS [2]. Ghayoomi and Daroodi [32] have used quadrogram model for both word and POS tags in a word prediction system for Persian. Their system achieved 42.45% KSS.

A linear combination model of POS tags tries to estimate the probability of POS tag for the current position according to the two previous POS tags. Then it attempts to find words that have the highest probability of being in the current position according to the predicted POS tag. Then, it combines this probability with the probability of the word given the previous word. So, there are two predictors in which one predicts the current tag according to the two POS tags and the one that uses bigram probability to find the most likely word. Fazly [2] reported that experimentally the system gained the result of 53.14% in KSS.

2) Rule-based Grammar: In this approach, syntactic word prediction would be made by using the grammatical rules of the language. A parser will parse the current sentence by using the grammar of the language to reach to its categories. The parsing method can be either top-down or bottom-up [29]. Phrase Structure Rule Grammar (PSRG), Context Free Grammar (CFG) [1], and Head-driven Phrase Structure Grammar (HPSG) are the methods that could be used in prediction systems based on grammatical rules. Windmill [1] is a sample of systems in which has used CFG along with PSRG. The system achieved a KSS between 37.3% to 55.1% depending on the type of text and the prediction algorithm.

B. Semantic Prediction

Some of the predicted items in the prediction list could be wrong semantically even though they are syntactically right. So, suggesting the words that are syntactically and semantically correct would increase the accuracy of the predictions [1], [22]. To reach the goal, a great semantic knowledge is tagged to the words and phrases in a corpus. Mostly in semantic prediction appearance of specific word with special content is a clue to increase the probability of appearing other words that have semantic relationships to that word. PROSE [22] is a sample of systems used semantic knowledge of English language.

Two methods are used for semantic prediction. One of these methods is lexical source like WordNet in English which measures the semantic probability of words to get assured that the predicted words are semantically related in that context.

The other method is lexical chain that assigns the highest priority to the words which are related semantically in that context; the unrelated words to that context would be removed from the prediction list.

C. Pragmatics Prediction

Pragmatics affects the predictions too. Adding the method to the prediction procedure tries to filter the words that are probably correct syntactically and semantically, but wrong according to discourse. The pragmatic knowledge is also tagged to the words in a corpus. Suggesting the words that are correct pragmatically would increase the accuracy of predictions as well [22]. CHAT and TalksBack [22] have used the pragmatic knowledge of English.

TOPIC [30] is a system that has used a combination of semantic and pragmatic knowledge of English.

IV. Heuristic Modeling (Adaptation)

To make predictions more appropriate for a specific user, the adaptation methods are used. This approach tries to get adapted the system to every individual user [2], [1], [3]. There are two general methods that make the system adapted to the users. One of the methods is short-term learning and the other one is long-term learning that will be described in this section.

A. Short-term Learning

In this approach, the system adapts to the user on a current text that is going to be typed by an individual user. Recency promotion, topic guidance, trigger and target, and *n*-gram cache are the methods that a system could use to adapt itself to a user in a single text. The methods are commonly used in prediction systems.

1) Recency Promotion: The term "recency" has come from cognitive psychology. This concept means a word that has already occurred in a text will be given a higher probability of use; thus, more likely to be used in that text again. Such a method usually assigns dynamically higher probabilities to the words that recently are used in the text; so, it does not only take into account what words have been typed; but further, how recent they have been used [23].

2) Topic Guidance: This approach is a way of adapting the predictor to the overall subject of the current text. To do so, the general lexicon is complemented with a domain specific lexicon that contains words which are frequently occurring within certain domains, though not very common in general [3].

3) Trigger and Target: In this method, the appearance of a word is highly correlated with other word sequences. It means when the word A, the trigger, occurs in the text, it triggers the word B, the target. Then it causes the target word's probability estimation to change [21].

4) N-gram Cache: It is assumed that if a word is used once, it is more likely to be used again. In other words, the previous use of a word in a context increases the probability of that word to be used again. Using *n*-gram cache is a way to capture the most common words and sequences that are frequently used. These words would be put in the cache to get an increased probability [21].

B. Long-term Learning

In this method, the system gets adapted to the user by considering not only the current text, but previous texts that are produced by the user. As a result, gradually by using the system more, it adapts to the user heuristically [3]. Some of the methods for heuristics adaptations that are language specific are adding new words, automatic capitalization, providing inflected form of words, and compounding.

1) Adding New Words: Heuristic adaptation may involve adding new words to the lexicon of the system whenever the user types unknown words to the system. The added new words could be called in the prediction list for future use [23].

2) Automatic Capitalization: Depending on the language that the system is running for, some letters should be capitalized. For example, the first letter of a word at the beginning of a sentence and also proper words must be capitalized. Automatic capitalization allows the user to save more keystrokes [23].

3) Providing Inflected Form of Words: For some languages which are very inflected such as German, French etc, the prediction system would be more efficient to the user if the system takes the inflected forms of the words into account. The result is having higher percentage of KSS [31].

4) Compounds: Compounding is a method to make new words from other words. Compound words are written as a single unit. Compounds are numerous in languages like German, French etc. Adding such a method to the prediction systems allows the user to write compound words more easily with higher KSS [31].

V. Summary and Conclusion

Briefly, different approaches which are introduced by different researches along with their prediction methods overviewed. Traditionally, prediction systems used only statistics and frequency of use. Since some predictions were not appropriate, syntactic knowledge of the language was added to the systems to have better predictions. Gradually it was cleared some of the suggestions should be filtered based on semantic knowledge of the language; and later it was found out although some of the predictions are appropriate syntactically and semantically, inappropriate in terms of discourse and pragmatic knowledge of the language to achieve higher saving in keystrokes. Beside the methods, it is needed for the system being user-friendly to adapt itself to the user. For achieving this goal, heuristic methods are used in the systems.

Some systems use a combination of methods for predictions which could use all the methods simultaneously in one system to have better predictions with higher percentage in KSS. It seems this method is the best approach to have appropriate predictions by utilizing more linguistic knowledge such as syntactic, semantic, and pragmatic beside the statistical knowledge at once to save more keystrokes. Having such a system needs a great body of available knowledge of the language to the system. This approach might be a step forward to be closer to the 100% KSS.

References

[1] M. E. J. Woods, *Syntactic Pre-Processing in Single-Word Prediction for Disabled People*. PhD. dissertation, University of Bristol, Bristol, 1996.

[2] A. Fazly, *The Use of Syntax in Word Completion Utilities*. Master dissertation, University of Toronto, Canada, 2002.

[3] E. Gustavii and E. Pettersson, *A Swedish Grammar for Word Prediction*. Uppsala University, Stockholm, 2003.

[4] J. Hasselgren, E. Montnemery, P. Nugues, and M. Svensson, "HSM: A predictive text entry method using bigrams", 10th Conference of EACL, In *Proceedings of the Workshop on Language Modeling for Text Entry Methods*, Budapest, Hungary, pp. 59-99, 2003.

[5] C. L. James, and K. M. Reischel, "Text input for mobile devices: Comparing model prediction to actual performance", In *Proceedings of CHI-2001, ACM*, New York, pp. 365-371, 2001.

[6] Zi Corporation, eZiText. Technical report, 2002. http://www.zicorp.com

[7] Lexicus Division, iTap. Technical report, Motorolla, 2002. http://www.motorola.com/lexicus

[8] http://www.eatoni.com

[9] S. Mackenzie, H. Kober, D. Smith, T. Jones, and E. Skepner, "Letterwise: Prefix-based disambiguation for mobile text input In Proceedings of the ACM Symposium on User Interface Software and Technology UIST, New York. ACM, 2001.

[10] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice-Hall, NewJersy, 2000.

[11] L. Booth, W. Beattie, and A. Newell, "I know what you mean", *Special Children*, pp. 26-27, 1990.

[12] S. Hunnicutt, "Lexical prediction for a text-to-speech system in communication and handicap: Aspects of psychological compensation and technical aids" E. Hjelmquist and L.-G. Nillsson, eds. *Elsevier Science Publisher*, 1986.

[13] S. Hunnicutt, "Using syntactic and semantic information in a word prediction aid", *Eurospeech*, vol. 1, pp. 191-193, Paris, France, 1989.

[14] C. D. Manning and H. Schdze, *Foundation of Statistical Natural Language Processing*. The MIT Press, 1999.

[15] F. Shein, T. Nantais, R. Nishiyama, C. Tam, and P. Marshal, "Word cueing for persons with writing difficulties: WordQ", In *Proceeding of 16th Annual Conference on Technology and Persons with Disabilities*, Los Angeles, CA, 2001.

[16] http://www.gusinc.com/wordprediction.html

[17] T. Nantais, F. Shein, and M. Johnsson, "Efficacy of the word prediction algorithm in WordQTM", In *Proceedings of REZNA*, Reno, Nevada, 2001.

[18] M. Ghayoomi and S.M. Assi. "Word prediction in a running text: A statistical language modeling for the Persian language", In *Proceedings of the Australasian Language Technology Workshop*, University of Sydney, Australia, pp. 57-63, 2005.

[19] http://www.finishline.featureditems.com

[20] S. Bickel, P. Haider, and T. Scheffer, "Predicting Sentences using N-Gram Language Models", In *Proceedings of Conference on Empirical Method in Natural Language Processing*, 2005.

[21] R. Rosenfeld, *Adaptive Statistical Language Modeling: A Maximum Entropy Approach*. PhD. dissertation. Canegie Mellon University, Pittsburgh, 1994.

[22] K. McCoy and P. Demasco, "Some application of natural language processing to the field of augmentative and alternative communication", In *Proceeding of the IJCAI 95 Workshop on Developing AI Applications for People with Disabilities*, 1995.

[23] S. Hunnicutt and J. Carlberger, "Improving word prediction using markov models and heuristic methods", *Augmentative and Alternative Communication*, vol. 17, pp. 255-264, 2001.

[24] C. Morris, A. Newell, L. Booth, and J. Arnott, "Syntax pal a system to improve the syntax of those with language dysfunction", In *Proceedings of REZNA*, pp. 105-106, 1991.

[25] C. Morris, L. Booth, I. Ricketts, N. Alm, and A. Newell, "Evaluation of a syntax-driven word prediction for children with language impairments", In *Proceedings of the Annual Conference of REZNA*, pp. 423-425, 1993.

[26] J. Carlberger, *Word Prediction: Design and Implementation of a Probabilistic Word Prediction Program*. Master dissertation, Royal Institute of Technology, Stockholm, 1997.

[27] M. Wester, *User Evaluation of a Word Prediction System*. Master dissertation, Uppsala University, 2003.

[28] J. Treviranus and L. Norris, “Predictive programs: Writing tools for severely physically disabled students”, In *Proceedings of the REZNA*, Washington D.C., pp. 353-354, 1987.

[29] Garay-Vitoria and Gonzlez-Abascal, “Word prediction for inflected languages: Applications to Basque language”, *Copestake, A., Langer, S. and Palazuelos-Cagigas S., editors, Natural Language Processing for Communication aids, In Proceedings of a workshop sponsored by ACL, Madrid, Spain*, pp. 23-28, 1997.

[30] N. Alm, J. Arnott, and A. Newell, “Database desighn for stories and accessing personal conversational material”, In *Proceedings of the 21th Annual Conference of REZNA*, pp. 147-148, New Orleans, Louisiana, USA, 1989.

[31] P. Boissiere, “An overview of existing writing assistance systems”, In *Proceeding of French-Spanish Workshop on Assistive Technology*, Paris, France. 2003

[32] M. Ghayoomi, and E. Daroodi, “A POS-based word prediction system for the Persian language”, In *Proceeding of the 6th International Conference in Advancec in Natural Language Processing (GoTAL2008)*, Gothenburg, Sweden. 2008.