Word Prediction Using a Systemic Tree Adjoining Grammar

Gijoo Yang, Kathleen McCoy, Patrick Demasco

Applied Science and Engineering Laboratories
University of Delaware/ A.I. duPont Institute
Wilmington, Delaware USA

Introduction

Several augmentative communication systems attempt to predict the most likely input from the user in order to make that input available with minimal time/effort. To date these systems have relied on frequency data (e.g., word frequency statistics, word category statistics) However, none of them have employed the approaches developed in linguistics or computational linguistics to the fullest power. In this work the application of several of these approaches to the area of word prediction is investigated.

Background

AAC systems with predictive capabilities can potentially increase communication rate for non-vocal users. Word prediction systems typically present a list of likely words to the user. This list is based on previous user input and general linguistic knowledge. Early word prediction systems relied on word frequency statistics (Gibler, 1983). Typically, these systems used initial letter selections to constrain the possible choices from a frequency ordered dictionary. While this approach did offer rate enhancement over traditional spelling systems, the use of additional linguistic information could potentially improve the accuracy of predictions.

Syntax was used to enhance the predictive system PAL (Arnott, 1984) (Swiffin, 1987). It maintained a table of probabilities for all possible pairs of word categories such as adjective-noun, verb-noun, etc., and used these probabilities, in conjunction with word frequency data, to compute the probability of the next word. For example, after an adjective is introduced in a sentence, the probability of a specific noun as its successor is computed by multiplying the pre-computed probability of an adjective-noun pair and the noun's static frequency.

While PAL's addition of syntax was somewhat successful, the researchers reported only small improvement in prediction accuracy (4.3% to 6.4%). This modest gain was attributed to a number of factors including the size of the dictionary. The system was somewhat inaccurate because many words belong to more than one syntactical class; yet the system always assured the most frequent class assignment. Inaccuracy also derived from the small scope of application of syntactic knowledge. The authors suggested increasing the transition matrix to three dimensions (i.e., prediction based on 2 previous words) to improve the scope limitation.

A more refined model of syntax (e.g., one that parses the entire sentence using a grammar of English) would certainly increase prediction accuracy. It would allow the entire sentence preceding the current word to be taken into account to yield more efficient statistics. For instance, a noun at the beginning of a sentence is likely to be followed by a verb; a noun following a verb is very unlikely to be followed by another verb. This could be captured using the more global context suggested here.

Ultimately, however, the use of syntax is limited by the fact that it only provides information about sentence structure. It would be advantageous to consider semantic and pragmatic information. In addition, to incorporate advanced syntax semantics and pragmatics into a predictive system, it will be necessary to go beyond a purely stochastic paradigm (e.g, frequency tables) and utilize rule-based formalisms from the field of computational linguistics.

Proposed Work

One goal of this work is to improve predictability of the next word in a sentence by considering a number of linguistic information sources. To improve prediction using syntax, we turn to a standard computation oriented grammatical formalism such as described in (Allen, 1987) and (Winograd, 1983). While incorporating any such theory is likely to improve the situation, it still leaves much to be desired. For instance, consider a situation where *If you do that again, I will smack...* is uttered and the system must predict the next word. Syntax alone can predict that a noun phrase will follow. However, if semantics is also taken into account, the system can predict that the noun phrase will most likely be a person.

Systemic grammar (Halliday, 1985) which has been developed by Halliday and other researchers is a formalism which provides information about many aspects of language. It views an utterance as some meaning whose expression will vary depending on the situation. Thus, it takes into account the semantics of the intended utterance as well as such things as the relationship between the speaker and hearer.

Prediction Using S-TAG

While the above aspects of systemic linguistics make it very attractive for use by a prediction system, in its present formulation systemic linguistics does have its draw-backs. In particular, while the aspects noted in systemic linguistics do affect the syntax chosen for the intended utterance, systemic linguistics itself has no formal explicit treatment of syntax. In this work we propose to augment the systemic grammar framework with a formal syntactic component. In particular, our investigation has led us to Tree Adjoining Grammar (Joshi, 1983) as a formalism which is particularly well suited for this task.

In systemics, an utterance is made by going through a system of choices, each of which can be a choice at different strata; situation, semantics, or syntax. Selecting a particular choice on higher stratum affects the choices on lower strata, i.e., selecting a choice narrows down the possibilities for how the meaning can be conveyed. For instance, the above utterance might result from a mother attempting to warn her child to stop doing some activity. Given the relationship (e.g., mother-to-child) a particular semantic choice (e.g., *if further activity, physical warning to the hearer*) is made, which is a mild physical warning. This semantic choice is made because the relationship (choice at situation stratum) is mother-to-child.

In order to use this formalism for prediction we must take the utterance made so far and infer which choices in the grammar must have been taken in order to get that partial utterance. The predicted next word must also be consistent with these inferred settings.

Compared to systems equipped with syntax alone, the prediction power of this system is substantially increased. For instance, using the same example above, the blank can be predicted to be a noun phrase because that is the only syntactic category consistent with the syntactic choices made so far. As explained above, semantics helps predict that the noun phrase be a person for the same reason.

With systemic grammar it is possible to go even further because situational strata of systemic grammar takes the patterns of utterances into account. This may allow it to infer the appropriate situation (e.g., mother-to-child control). If this is the case, the system can predict that the person should be *you*.

Conclusion

In this work the use of TAG as a formal syntactic component in a systemic grammar framework has been investigated. The work has then been applied to word prediction in the augmentative communication. With systemic grammar it is possible to infer a more global setting by taking into account a piece of linguistic information which appears in a partial utterance. The inferred setting then can be easily used to aid in the prediction of the next word/phrase. Systemic grammar is useful in this task because only certain choices can be made depending on a particular setting.

We intend to implement a prototype system using NIGEL (Mathiessen, 1984), an implementation of systemic grammar which was developed at the Information Sciences Institute. The system will take an input from the user and provide a list of words/phrases which are predicted by taking several aspects of the input into account.

Acknowledgments

This work is partially supported by Grant #H133E80015 from the National Institute on Disability and Rehabilitation Research. Additional support has been provided by the Nemours Foundation.

References

James Allen. *Natural Language Understanding*. Benjamin/Cummings, CA, 1987

J.L. Arnott, J.A. Pickering, A.L. *An adaptive and predictive communication aid for the disabled that exploits the redundancy in natural language*. In RESNA 7th annual conference, pages 349-350, RESNA, Ottawa, Canada, 1984

R. Foulds, M. Soede, H. Balkom, and L. Boves. *Lexical prediction technique applied to reduce motor requirements for augmentative communication*. In RESNA 10th annual conference, pages 115-117, RESNA, San Jose, CA, 1987

C.D. Gibler, D.S. Childress *Adaptive dictionary for computer-based communication aids*. In RESNA 6th annual conference, pages 165-167, RESNA, San Diego, CA, 1983

M. A. K. Halliday. *An Introduction to Functional Grammar*. Edward Arnold, London England, 1985

A. K. Joshi. *Factoring recursion and dependencies*. In Proc. 21st Annual Meeting of the ACL, pages 7-15, ACL, Cambridge MA, June 1983

Christian Mathiessen. *Systemic Grammar in Computation: The NIGEL case*. ISI/USC ISI/RS-83-121 February, 1984

A. L. Swiffin, J.L. Arnott, and A. F. Newell. *The use of syntax in a predictive communication aid for the physically handicapped*. In RESNA 10th annual conference, pages 124-126, San Jose, CA, 1987

T. Winograd. *Language as a Cognitive Process*, Vol. 1: Syntax. Addison-Wesley Publishing Company, Reading MA, 1983

Gijoo. Yang
Applied Science and Engineering Laboratories
Alfred I.duPont Institute
P.O. Box 269
Wilmington, DE 19899
Phone: (302) 651-6830
Email: yang@wizard.asel.udel.edu