A Connectionist Model for Part of Speech Tagging

Brent A. Olde, James Hoeffner, Patrick Chipman, Arthur C. Graesser, and the Tutoring Research Group
The University of Memphis
Department of Psychology
Campus Box 52640
Memphis, TN 38152-6400
Baolde@memphis.edu

Abstract

AutoTutor is a fully automated tutoring system that attempts to comprehend learner contributions and formulate appropriate dialogue moves. This paper reports the mechanisms and performance of one of AutoTutor's language modules, the word tagging module. AutoTutor's word tagging module determines the part of speech tag for every word in the learner's contributions. It uses a two part procedure: it first consults a lexicon to identify the set of possible tags for each word, then it uses a neural network to select a single tag for each word. Performance assessments were made on a corpus of oral tutorial dialogue, as opposed to well-formed printed text. The lexicon provided the correct tag, as one member of a set, for 97% of the words and 91.6% of the neural network's first-choice tags matched assignments by humans.

Introduction

An intelligent tutoring system (ITS) is fully automated if the learner can use the system without extensive training on the human-computer interface and if there is no need for an auxiliary agent (i.e., human or software) to interpret the learner's contributions. There are a number of fully automated systems on topics such as mathematics, medicine, and technical equipment (Anderson et all. 1995; Hume et all. 1996; Lesgold et all. 1992; Van Lehn, 1990).

Three serious barriers have limited the extent to which learners can use an ITS by holding a conversation with the system. These barriers include: (1) the inherent difficulty of getting the computer to "comprehend" the language of users, including utterances that are not well-formed syntactically and semantically, (2) the difficulty of getting computers to effectively use a large body of potentially relevant world knowledge, and (3) the lack of research on human tutorial dialog. However, recent advances in computational linguistics, cognitive science, artificial intelligence, and discourse processing have reduced these barriers substantially. The tutoring system that we have developed, called AutoTutor, incorporates some of these recent advances in addition to more established computational procedures.

AutoTutor was developed to tutor high school students and college students on the fundamentals of computer literacy, such as hardware, the operating system, and the Internet (Graesser, Franklin, Wiemer-Hastings, & the Tutoring Research Group, 1998; Wiemer-Hastings, Graesser, Harter, & the Tutoring Research Group, 1998). AutoTutor follows a curriculum script that asks questions and presents problems that involve lengthy answers. That is, the answers and solutions require an interactive dialogue between computer and tutor that involves many turns. At each step of the exchange, AutoTutor attempts to "comprehend" the speech acts within the turn of the learner (typed into the keyboard) and to formulate one or more dialogue moves that are sensitive to the quality of the learner's contributions. AutoTutor's dialogue moves include short feedback (positive, negative, neutral), assertions, prompts for learner information, hints, corrections, and other categories of speech acts. These dialogue moves are delivered by a talking head with appropriate facial expressions and synthesized speech. However, this paper focuses on the comprehension mechanisms of AutoTutor rather than the production of dialog moves.

AutoTutor attempts to comprehend what the human learner types into the keyboard by using several language modules. Each language module analyzes some aspect of the content of the message that the learner types into the keyboard during a particular conversational turn. First, the sequence of words and punctuation marks in a turn are segmented into a sequence of word units; punctuation marks are regarded as a special class of words. Second, a part of speech (POS) tag is assigned to each word unit. This is accomplished by accessing words in a lexicon and computing the best tag (i.e., syntactic class) through a neural network. Third, the sequence of words are segmented into speech act units and classified into speech act categories by another neural network. Examples of speech act categories are wh-questions, yes/no-question, short responses, directives, and contributions. Fourth, the meaning of each speech act is

interpreted by latent semantic analysis (Landauer & Dumais, 1997) and other semantic analyses. However, the deeper levels of comprehension are not under focus in this paper. Our immediate concern is in the language module that assigns POS tags to word units.

The Role of Part of Speech Tagging in AutoTutor

Each word unit in the learner's turn is assigned to a POS tag, such as noun, main verb, adjective, determiner, punctuation mark, and so on. The input is the sequence of words and punctuation marks handed up by the segmentation module. The output of the POS tagger is the first-choice tag for each word unit; these sequence of tags are passed on to the speech act classification system. The POS tags play a crucial role in speech act classification. For example, the network learns that when the first word is a main verb, the speech act is often a directive.

Existing POS tagging systems can function at high levels of accuracy when applied to a corpus of printed texts. Allen (1995) states, for example, that most POS taggers can achieve an accuracy rating of 90%. This raises the question of why we attempted to create our own POS tagger. The language of many learners who use AutoTutor is more akin to oral conversation than to printed text. Much of the language is ungrammatical, vague, semantically ill-formed, incoherent, and replete with repairs and metacommunication markers (e.g., uh-huh, uh). Naturalistic tutorial dialogues, such as those collected by Graesser and Person (1994) are more "noisy" than concisely worded, preprocessed text. Redington, Chater, and Fich (1998) analyzed an oral corpus and reported an accuracy of only 72% when considering distributional analyses of syntactic categories. Our attempt to analyze an oral corpus of tutorial dialogues is a good match to the keyboard input of learners of AutoTutor. It will also help facilitate the eventual transition from keystroke contributions to speech recognition as the primary input for AutoTutor.

Another reason to create our own POS tagger is that it is more flexible in inducing patterns of word tag sequences in the corpus. The neural network can induce unexpected patterns in the words that surround any given word, at least compared to alternative symbolic POS taggers. The first stage of AutoTutor's POS tagger is to have the computer lexicon generate a list of candidate POS tags. Once this is done, the second stage uses a neural network to incorporate surrounding contextual cues to determine the single most likely POS tag. As has been noted, we tested AutoTutor's POS tagger on a corpus of oral tutorial dialogues, namely the corpus collected by Graesser and Person (1994). To our knowledge, there are no POS taggers for oral communication that have reached a high accuracy (i.e., over 90%).

Our POS Tagging Method

The part of speech tagging system consists of two separate components. The first component is a lexicon developed by Francis and Kucera (1982) and by the MRC Psycholinguistic Database. Each word is matched to its entries in the computer lexicon. The computer pulls out the set of possible POS tags and associated frequencies listed in the lexical entry for the word. The second component of the POS tagging system is a neural network. It uses local context (the words preceding and following the target word and its position in the sentence) and base rate frequency information to select the most likely POS tag from the set of candidate tags handed up by the lexicon.

The lexicon assigns each word or punctuation unit to one or more POS categories. We used a set of 17 parts of speech, including major syntactic categories (such noun, adjective, and main verb) and less frequent categories that are particularly diagnostic for speech act classification (e.g., wh-words). Commas, periods, question marks, etc. were assigned to the category "punctuation mark". The major syntactic categories included verb, preposition, noun, adverb, pronoun, interjection, adjective, conjunction, auxiliary/modal, and determiner/article. The minor syntactic categories included reply (words such as "yes" and "OK" that are often used as responses to questions), mathematical operators (e.g., plus and minus signs), digits, question words (e.g., what, who, how, etc.), be/have/do (since these verbs can be both main verbs and auxiliaries), and "other" (for words that do not have an entry in our lexicon).

In our second set of simulations, an additional set of contextual cues indicating the position of the target word was added. We used a set of 4 position tags, the first word in a sentence, the second word in a sentence, the third word in a sentence, and a last word in a sentence.

The computation of frequency had a number of steps. The Kucera and Francis (KF) was first examined because it contains frequency information for each separate word tag (for a particular word) and we use this information to calculate activation in the neural network. If the word is found, the set of word classes and associated frequencies are considered candidates. The frequency data were normalized as percentage scores by taking the total number of frequencies for that word and dividing the frequency for each class by this total frequency. Next, the words in the special categories had only one assigned word tag so the activation for that tag was assigned 100%. When a word was not found in the KF corpus, the word was assessed by the MRC corpus. This was used as a backup since the individual tag frequencies of a given

word in the MRC all had the same frequency, and hence would all have the same calculated percentage. In summary, the word, its classes, its position in the sentence, and the activation calculated from the class frequencies were passed on to the neural network.

The first neural network is a feed-forward network with 51 input units, 16 hidden units, and 17 output units (corresponding to the 17 POS categories). The second neural network is a feed-forward network with 55 input units, 16 hidden units, and 17 output units. The network is trained with the backpropagation algorithm.

55 Input nodes: Before (17) Target word (17) Target word position (4) After (17)

1. Noun (N)
2. Pronoun (U)
3. Verb (V)
4. Auxiliary (X)
5. Special Auxiliary (S)
6. Adjective (J)
7. Adverb (A)
8. Preposition (R)
9. Conjunction (C)
10. Article Determiner (L)
11. Interjection (I)
12. Question words (Q)
13. Reply (Y)
14. Mathematical Operators (M)
15. Digits (D)
16. Punctuation (T)
17. Not found in Database (Z)

1. First word in a sentence
2. Second word in a sentence
3. Third word in a sentence
4. Last word in a sentence

Hidden nodes: 16
Output nodes: 17 (N U V X S J A R C L I Q Y M D T Z)

The corpus in these analyses consisted of 420 randomly sampled learner turns in the naturalistic tutoring transcripts collected by Graesser and Person (1994). The Graesser & Person tutoring sessions consisted of transcribed records of videotaped college undergraduates who were tutored by graduate students on the topic of scientific research methods. Our random sample of learner turns had 3,170 words. These words were rated by humans on first-choice POS tags. The computer's POS tags were compared to the human's in order to assess the accuracy of AutoTutor's tagging procedure.

Results of Look-up Procedure

First-choice tags. When the human's POS tag was compared to the computer's first choice (i.e., highest frequency) POS tag, there was a .867 **hit rate**. The hit rate is the proportion of computer's first-choice tags that were also the human's first-choice tags. Table 1 shows hit rates for each separate word class. Other measures include the **incidence** in the sample (proportion of time a tag occurred in the sample, according to human raters), **false alarm rate** (proportion of time the lexicon assigned a first-choice tag which was not assigned by a human as first choice), and a **d' score** (a pure measure of how discriminating AutoTutor is in assigning tags).

Word Class	Incidence in the sample	Hit Rate	False Alarm	d'
Noun	.098	.806	.017	2.89
Pronoun	.149	.742	.000	2.96
Verb	.093	.902	.008	3.60
Auxiliaries	.021	.924	.000	3.72
Special Auxiliaries	.084	1.000	.000	4.64
Adjective	.062	.687	.022	2.52
Adverb	.062	.813	.014	3.10
Preposition	.054	.959	.002	4.07
Conjunction	.035	.919	.030	3.28
Article Determiner	.053	.994	.032	4.20
Interjection	.039	.696	.000	2.84
Question Words	.012	1.000	.000	4.64
Reply	.033	.865	.004	3.38
Mathematical*	NA	NA	.000	NA
Digits	.005	1.000	.000	4.64
Punctuation	.200	1.000	.000	4.64
Not Found*	NA	NA	.000	NA
Total		**.867**	**.001**	**3.43**

Table 1: POS Tag Comparison Between the Humans' First Choice and the Lexicon's First Choice

Set of tags. The full set of tags generated for each word by the lexicon was compared to the humans' choices. A hit was scored whenever the humans' choice was in the **set** of candidates generated by the lexicon. In this case, the hit rate was extremely high (.970). Thus, the lexicon is nearly always providing the correct tag as one of its choices.

The mean number of tags in a set produced by the computer was 1.70 (s.d. = 1.00). Thus, the high rate of agreement between human and computer was not a result of the computer generating a very large set of candidates. Table 2 presents incidence scores and hit rates for each of the 17 word classes.

* These categories were not present in our corpus.

Word Class	Incidence in the sample	Hit Rate
Noun	.098	.971
Pronoun	.149	.970
Verb	.093	1.000
Auxiliaries	.021	.924
Special Auxiliaries	.084	1.000
Adjective	.062	.795
Adverb	.062	.919
Preposition	.054	1.000
Conjunction	.035	.991
Article	.053	1.000
Determiner		
Interjection	.039	.952
Question Words	.012	1.000
Reply	.033	.981
Mathematical*	NA	NA
Digits	.005	1.000
Punctuation	.200	1.000
Not Found*	NA	NA
Total		**.970**

Table 2: POS Tag Comparison Between the Humans' First Choice and the Lexicon's Choices

Results of Neural Network

The network was trained on 80% of the oral corpus. The training resulted in .926 correct classification. The network was tested on the other 20% of the oral corpus. In the first test of the network there were three sources of information available to the network, 1) the set of tags generated by the lexicon for each target word, 2) the immediate context, that is, the sets of tags generated for the word preceding and the word following the target word, and 3) the relative frequencies for the tags in all three sets.

The network with the highest activation was considered to be the network's first choice. The network's choice agreed with the human rater's choice 88.9 % of the time. A chi-square was performed to assess whether the lexicon's first choice POS tag rating, based on the frequency of the tag for the word (.867), was significantly different than the neural network test (.889). This difference was not statistically significant, $p > .05$.

In an effort to improve performance we attempted to fine-tune the system by varying the number of hidden units, learning rate, and training time. Although these efforts produced an increase in the correct classification rating to 89.1%, this was also not significantly better than the lexicon's performance level.

An analysis of the network results revealed that the network was performing poorly on word classes where the position of the word in the sentence could be an important cue. For example, the human raters coded "what" as a question word when it appeared in the first position in the sentence. However, "what" can belong to several other POS categories when it is found in other sentence positions. Therefore, an additional set of cues were added to the network's input. Four new input units coded for the position of the target word in the sentence (indicating if the target word was the first, second, third, and/or last word in the sentence).

With these additional cues available to the network, the performance scores went up for nearly every POS category. The overall hit rate increased to .916, which was significantly greater than the .867 hit rate achieved by using the lexicon alone ($p < .01$ in a chi-square test). Table 3 presents the breakdown of the network results as a function of the 17 categories.

Word Class	% Correct
Noun	83.9
Pronoun	93.0
Verb	90.2
Auxiliaries	100.0
Special Auxiliaries	100.0
Adjective	75.9
Adverb	76.9
Preposition	93.8
Conjunction	88.9
Article	100.0
Determiner	
Interjection	74.2
Question Words	100.0
Reply	93.8
Mathematical*	NA
Digits	100.0
Punctuation	99.2
Not Found*	NA
Total	**91.6**

Table 3: Breakdown of the Network Results as a Function of the 17 POS Categories.

Conclusion

The word tagging module was able to reach performance of 91.6% correct on our test sample. The use of a two stage procedure gives us great flexibility in improving word tagging performance and in adapting the tagging module to new tutoring environments. Use of the lexicon allows us to easily add new words as we deal with the specialized vocabularies found in tutoring different academic subjects. The neural network has great flexibility as well, since we can add new cues or additional context to enhance performance whenever analyses indicate that these additional constraints may be informative.

Acknowledgments

This research was funded by the national Science Foundation (SBR 9720314). The members of the Tutoring Research Group are: Ashraf Anwar, Patrick Chipman, Scotty Craig, Rachel DiPaolo, Stan Franklin, Max Garzon, Art Graesser, Barry Gholson, Doug Hacker, Peggy Halde, Derek Harter, Jim Hoeffner, Xiangen Hu, Jeff Janover, Bianca Klettke, Roger Kreuz, Kristen Link, Johanna Marineau, William Marks, Brent Olde, Natalie Person, Victoria Pomeroy, TeeJay Shute, Shannon Whitten, Peter Wiemer-Hastings, Katja Wiemer-Hastings, Holly Yetman, and Zhaohua Zhang.

References

Allen, J., (1995). Natural Language Understanding. Redwood City, CA: Benjamin/Cummings Publishing Company, Inc.

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. The Journal of the Learning Sciences, 4, 167-207.

Francis, W.N., & Kucera, N. (1982). Frequency Analysis of English Usage. Houghton-Mifflin.

Graesser, A., Franklin, S., Wiemer-Hastings, P., and the Tutoring Research Group (1998). Simulating smooth tutorial dialogue with pedagogical value. Proceedings of the 11th International Florida Artificial Intelligence Research Symposium, Menlo Park, CA: AAAI Press.

Graesser, A., & Person, N.K. (1994). Question asking during tutoring. American Educational Research Journal, 31, 104-137.

Hume, G., Michael, J., Rovick, A., & Evans, M. (1996). Hinting as a tactic in one-on-one tutoring. The Journal of the Learning Sciences, 5, 23-47.

Landauer, T.K., & Dumais, S.T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological review, 104, 221-240.

Lesgold, A., Lajoie, S., Bunzo, M., & Eggan, G. (1992). SHERLOCK: A coached practice environment for an electronics troubleshooting job. In J. H. Larkin & R. W. Chabay (Eds.), Computer-assisted instruction and intelligent tutoring systems (pp. 201-238). Hillsdale, NJ: Erlbaum

MRC Psycholinguistic Database. Machine Usable Version 2.00. Informatics Division, Science and Engineering Research Council.

Redington, M., Chater, N., &Fich, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. Cognitive Science, 22, 425-469.

VanLehn, K. (1990). Mind bugs: The origins of procedural misconceptions. Cambridge, MA: MIT Press.

Wiemer-Hastings, P., Graesser, A., Harter, D., and the Tutoring Research Group, (1998). The foundations and architecture of Autotutor. Proceedings of the 4th International Conference on Intelligent Tutoring Systems, San Antonio, Texas (pp. 334-343). Berlin: Springer-Verlag.