Evaluating a Model to Disambiguate Natural Language Parses on the Basis of User Language Proficiency

Lisa N. Michaud[1] and Kathleen F. McCoy[2]

[1] lmichaud@wheatoncollege.edu
Dept. of Mathematics and Computer Science, Wheaton College, Norton, MA
[2] mccoy@cis.udel.edu
Dept. of Computer and Information Sciences, University of Delaware, Newark, DE
http://www.eecis.udel.edu/research/icicle

Abstract. This paper discusses the evaluation of an implemented user model in ICICLE, an instruction system for users writing in a second language. We show that in the task of disambiguating natural language parses, a blended model combining overlay techniques with user stereotyping representing typical linguistic acquisition sequences successfully captures user individuality while supplementing incomplete information with stereotypic reasoning.

1 Introduction: the ICICLE System

The name ICICLE represents "Interactive Computer Identification and Correction of Language Errors" and is the name of an intelligent tutoring system currently under development [5, 6, 7]. The system's primary long-term goal is to employ natural language processing and generation to tutor deaf students on grammatical components of their written English. ICICLE accepts as input free-written English texts and responds to the user by highlighting sentences with errors. Our system makes use of a user model to track the user's level of competence in different English syntactic structures; the development of this model is discussed in [6]. This paper addresses how we have evaluated our model within the task of disambiguating natural language parses.

1.1 User Modeling and the Disambiguation Task

ICICLE uses a CFG grammar which is descended from that described in [1], augmented with error-production rules called *mal-rules*, to parse user-written utterances. In the process of seeking a correct analysis of user errors, the ICICLE system needs to choose between multiple parses of each utterance. Some of these parses represent different structural representations of the text, and in the case of ungrammaticality, may place the "blame" for the error on different constituents. To determine which is correct, it is necessary for the system to have at its disposal a model of the student's grammatical proficiency which indicates his

or her mastery of the language rules involved. This knowledge aids in choosing between structurally-differentiated parses by providing information on which grammatical constructs the user can be expected to use correctly or incorrectly[3].

1.2 A Model of Grammar Proficiency

The ICICLE user model, described in depth in [4, 5, 7], attempts to capture what we refer to as "I_i," or the user's current *Interlanguage* state. The concept of interlanguage is that a language learner is generating utterances from a hypothesized grammar I which approaches the language being learned over time [9]. At the current step in the progression, I_i, certain constructs have been mastered, others are currently being learned, and some are still beyond the user's reach.

One component of the ICICLE user model is MOGUL (Modeling Observed Grammar in the User's Language), which captures what is known about the user's interlanguage grammar I_i through an overlay representation in which individual constructs of morphology and syntax (which we refer to as Knowledge Units, or KUs) are scored according to the system's observations of the user's success in executing those KUs in the writing he or she has previously produced. This model compares the number of times the KU has appeared correctly in the user's productions against the total number of times the KU has been attempted and summarizes this information into one of three tags: *Unacquired*, meaning the user has definitely not mastered the KU, *Acquired*, meaning the user consistently uses it correctly, and *ZPD*, meaning the KU is currently being mastered by the user and is therefore exhibiting great variation in successful execution.

Incomplete knowledge in MOGUL results when the system has not yet gathered data on a specific KU. These "gaps" in the profile of the user are filled using the information provided by the second component, SLALOM (Steps of Language Acquisition in a Layered Organization Model). The current implementation of SLALOM involves a representation of three learner stereotype layers (Low, Middle, High) [4]. Each stereotype is associated with a certain level of mastery of each linguistic KU. The current MOGUL tags for a student are compared against the three stereotypes, and the system selects the stereotype profile with the greatest level of similarity to observed performance for this user. That stereotype then provides probable tags for the KUs which have not yet been observed in the user's performance.

1.3 Using the Model to "Score" a Parse

This two-component model enables ICICLE to sift through the multiple syntactic analyses provided by its parser by indicating a maximally likely candidate to represent the user's attempted syntactic structures. The algorithm to accomplish this task was implemented with the following steps:

[3] This is not to say that the user will not make mistakes in already-mastered material. What we wish to select is the most likely parse given the current mastery of the language.

1. Obtain all possible parses for the input sentence.
2. Score each parse tree according to how likely it is given the user's current interlanguage state I_i (as captured in the user model). This scoring process is described below.
3. Select a parse tree with maximal score, i.e. one containing the most likely nodes.

Determining a parse tree's compatibility to I_i is done as a two-step process. First, the tree is traversed so that a score for each node is determined in the following manner:

1. Determine the parsing rule used to construct the constituent represented by this node and the KU to which this rule belongs.
2. Determine the tag on this KU. This will be Unacquired, ZPD, or Acquired. If there is insufficient data in MOGUL to supply this tag, the tag is inferred using the SLALOM information on typical performance for the user's stereotype level.
3. Translate this marking into a score for this rule, giving high scores to those rules which should be in I_i given the tag on the KU, and low scores to those rules which are not expected to be in I_i.

The process of obtaining the score in Step #3 reflects an answer to the question: *Do we believe that this rule is in I_i?* If the answer is yes, the node receives a positive score of 1. If the answer is no, the node receives a negative score of -1. "Unacquired" KUs imply that rules representing correct execution of the structure are not in I_i, but rules representing malformations of the structure are. Conversely, "Acquired" KUs are represented by correct (regular) rules in I_i, not mal-rules. KUs in the ZPD represent structures realized by competing rules, both correct and incorrect, which result in the variation in ZPD-level performance; for that reason, both mal-rules and correct rules are believed to co-exist in I_i for those structures.

Once all of the node scores for a tree are determined, these scores are combined to obtain an average score to represent the likelihood of the entire tree overall.

2 Evaluating the Model

This parse scoring mechanism and the user model on which it is based have been implemented within the ICICLE system. In order to demonstrate the efficacy of the implementation, we set out to show the following:

- Parse selection based on a stereotype successfully selects parses which are the closest match to the "expected performance" depicted in the stereotype image in SLALOM.
- When a user builds up a history of performance that deviates significantly from the assigned stereotype—for instance, when the student's proficiency changes because he or she is learning—the stereotype assignment is updated to better reflect the user.

– When a user is correctly placed in a stereotype and yet has individual deviations in his or her MOGUL tags from that stereotype, representing a history of "atypical" performance, the parse selector correctly recognizes the appropriateness of parse interpretations which are consistent with that user's individuality.

For this evaluation, we used a corpus of sentences contained in 106 samples of writing by deaf individuals at various levels of English proficiency.

2.1 Parse Selection Depending Upon the Stereotype

The parse selection process as it operates when all decisions are based on a selected stereotype level is consistent with the mode of operation with a new user, and also reflects the system's ability to select parses which are consistent with a complete performance profile. To illustrate this process, we selected a stereotype level of "Middle" for a hypothetical user, and we parsed the following Middle-rated sentence from our corpus:

(1) I really like wrestling.

The parser found six possible trees to span this input. Several of these parses received low scores; they all involved a syntactic interpretation containing a dropped copula verb[4]. This interpretation would be consistent with reading the sentence as "I *am* really like wrestling," whose parse is similar to the standard parse for "I am really like my mother." The parse involved dropping the copula verb *be*, an error common for some learners in this population but inconsistent with Middle-level performance. The mal-rule which handles dropped copulae (`-MV22>`) participates in a Knowledge Unit which, according to the SLALOM model, is in the ZPD for a Low-level learner, but is Acquired at the Middle level. Therefore, the parses involving the dropped copula were themselves "dropped" for involving a mal-rule reflecting an error we would not expect from this learner. The parses receiving high scores from the parse selection mechanism were far more consistent with the Middle-level performance profile represented by that stereotype layer in SLALOM.

To test the stereotype aspect of our model further, we used this sentence with a user of each of the other stereotypes. The tree scores we obtained are shown in Table 1. The expectations generated by the Low stereotype did not penalize the dropped copulae and, in fact, rewarded one parse (#1) which involved both the dropped copula and a dropped determiner. In this parse, *like* was treated as a noun without its required determiner[5]. The expectations of the High stereotype paralleled those of the Middle stereotype, except in that the gerundive use of "wrestling" as a noun was considered more likely at this level, raising almost all of the scores. This also resulted in more than one parse receiving a maximal

[4] Although *wrestling* is a verb form, the gerund is used as an NP in this interpretation and therefore does not function as a verb in this sentence.

[5] In this instance, "[a] like wrestling" was parsed as would be "a horse jumping."

Table 1. Parse Tree Scores for "I really like wrestling," All Stereotypes.

Tree	Low	Middle	High	Notes
0	0.75	0.75	**1.0**	GOOD
1	**1.0**	0.5	0.5	Dropped copula and determiner
2	0.75	0.5	0.75	Dropped copula
3	0.78	0.56	0.778	Dropped copula
4	0.78	0.56	0.778	Dropped copula
5	0.82	**0.82**	**1.0**	GOOD

score under the High stereotype; in Section 3, we discuss methods for empowering the scoring procedure to make greater differentiation between parses in order to minimize this occurrence.

In sum, we see the reflections of the different stereotype expectations working as we had hoped. Given a Low level stereotype, "I like wrestling" is assumed to contain errors because a Low level user is unlikely to have acquired this use of gerunds. At the higher levels, the gerund is more likely (and the dropped copula is less likely), resulting in an error-free interpretation.

2.2 Updating the Stereotype Assignment

What if the stereotype the system has recorded for a user is wrong? Recall that we expect our user's language proficiency to be dynamic as learning progresses, and that eventually the stereotype recorded for a user at step i in his or her language acquisition will no longer be appropriate when the user is at some later step $> i$. In our next task, we sought to illustrate how the system may recognize the inappropriateness of a stereotype for a given learner and update that stereotype assignment over time[6].

We chose to create a new user for this task, again with the stereotype level Middle. We wished to design a situation in which our learner was *previously* a Middle-level English user, but has now progressed to more advanced proficiency. In this situation, we wanted to update the stereotype selection to High. We selected a batch of 20 sentences from samples in our corpus. Fifteen of these came from samples which had been scored by our judges as representing a High proficiency level, and 5 came from samples which received a Low or Middle rating, but which contained some High-level syntactic structures. Our objective was to assemble sentences which clearly exhibit structures expected primarily of a High-level learner, the new stage to which our user had progressed[7].

[6] There is no bias toward upward revision of the stereotype. Although we chose to illustrate a learner's upward transition in this example, the ability of the user modeling component to adjust the stereotype is the same whether it is being revised higher or lower.

[7] These sentences were extracted from the corpus through a search for specific types of error and specific levels of competence, and were screened to ensure that ICICLE

In our first step, we fed these 20 sentences into the ICICLE analyzer (with the SLALOM-based stereotype set to Middle) together in a batch as if they were a user-written essay. This way, the parser would analyze them each in turn, but would not send data back to the user model to update it until the analysis of all 20 had finished. We sought to see how successful the suggestions of the Middle-level stereotype would be for these High-level sentences.

We compared ICICLE's parse selections against the "optimal" choices of a human judge[8]. The system gave the maximum score to an optimal parse 12 of 20 times (60%), despite having the "wrong" stereotype (Middle rather than High) on which to base these decisions[9]. We then inspected the resultant MOGUL markings to see if the system had learned something about the user in just this one sample of 20 sentences. In fact, eight Knowledge Units now bore MOGUL tags, all marking Acquired structures. Furthermore, those eight were now a closer match to a High stereotype than to a Middle one; the stereotype level of the user was therefore no longer set to Middle, but had been changed to High.

To further test the system's ability to adapt to a learning user over time, we next started with another new user (again set to "Middle") and iteratively entered these sentences one at a time, requesting an analysis (and subsequent user model update) after each sentence. The question we sought to answer here was whether the cumulative evidence provided by the earlier sentences would enable the system to make more appropriate decisions about the sentences at the end than it had in the batch run, and if the total number of correct choices would be higher, given the access to the evidence provided in the first several sentences. Despite the fact that 20 sentences provide relatively little evidence, the number of maximum-scoring optimal parses rose to 15/20 (75%). These results are compared with those of the batch run on the left in Figure 1.

We also examined the results obtained for the last five sentences in particular. In the original batch run, only in one of these five had the optimal parse been maximally scored by the system. In the results of this iterative run, however, we observed the following:

- Before reaching the final five sentences, the stereotype had already been revised to High, so that the decisions the system made on these were already affected significantly by the input from the earlier sentences.
- In all of the sentences for which the optimal parse tree's score was originally lower than the maximum, the score increased.

---

was capable of parsing them with an appropriate interpretation among those parses obtained.

[8] In some cases, more than one parse is considered optimal because of inconsequential syntactic differences which allow multiple parses to be acceptable. In all cases, however, the judge's choice was non-controversial; we would expect any native speaker judge to have selected the same choice.

[9] We report the "maximum score" rate here because, as mentioned above, there are cases in which the system is forced to make a random selection between parses receiving equal scores. In these situations, the system has recognized the validity of the optimal parse.

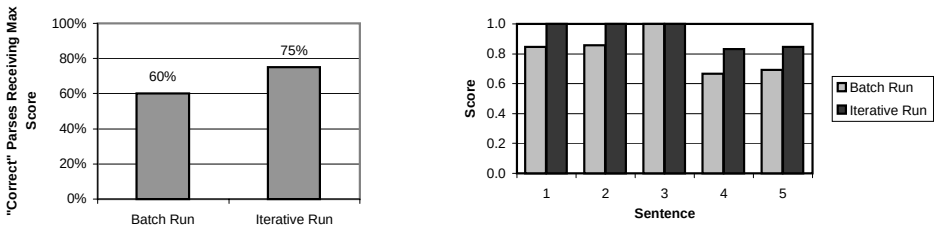

Fig. 1. Comparing the Batch and Iterative Runs.

– The number of maximum-scoring optimal parses rose from 1/5 to 3/5.

The differences in the two sets of scores are shown for all five sentences on the right in Figure 1.

2.3 Overriding the Stereotype

The previous two sections have illustrated how ICICLE will prefer a parse which is consistent with stereotype expectations in the absence of other information, how the system can recognize when a user deviates significantly from that stereotype (and can update the stereotype categorization accordingly), and how it can adjust its parse selection decisions to meet a user's changed acquisition status. In a final evaluation task, we explored the integration of the two facets of the user model; specifically, we tested ICICLE's ability to integrate the stereotype expectations which SLALOM provides with the specific history of parses that it records in the MOGUL model.

For this task, we chose to create a High-level user who is appropriately classified in that stereotype but in whose language mastery there exist certain fossilized structures which are executed with error in a fashion atypical of the High-level stereotype. We sought to illustrate that if the system has built up a performance history for this user illustrating these differences, it will select parses more appropriate to the user's actual interlanguage I_i even if that deviates from stereotype expectations.

We chose from the corpus 26 example sentences to represent our learner with fossilized errors. Specifically, we chose 12 sentences exhibiting Middle- and High-level structures executed correctly (as would be expected with the High-level stereotype), and 18 sentences containing Low-level errors, focusing on errors in noun pluralization, subject/verb agreement, and determiner usage.

Our first step in this test was to determine the "base case" of stereotype-based parse selection on these sentences. We therefore ran the entire set of sentences together as a single sample with only the stereotype setting of "High." We then noted in each case which parse the stereotype selected and compared it against the optimal human-selected parse. Of the 26 sentences, ICICLE gave a maximum score to the optimal parse in a total of 19 cases (73%).

We checked to see if there was a difference between those sentences which were "typical" performance for a High-level learner (and therefore true to the stereotype) and those which were "atypical." Of the 12 typical sentences, the system had given a maximum score to the optimal parse in all 12 (100%). In the atypical sentences, only 7/14 (50%) of the optimal parses received the maximum score. As may be expected, the parse selection process fared much better when the sentences were consistent with the stereotype than when the system had to go against the stereotype expectations to chose an optimal parse.

As shown in the previous section, however, the system is capable of adjusting to a user over time and the parse selection process will reflect that. To simulate an accumulated performance history, we "trained" the MOGUL model by hard-wiring the parse selector to choose the parse which was optimal according to human judgment and then running the entire set of 26 sentences through the analyzer 10 times. This recorded the rules from the optimal parse trees for each sentence 10 times into the MOGUL model.

After this history was constructed, it was inspected to see how different it was from the stereotypical MOGUL tags of a High-level learner. We noted that 44 of the 114 Knowledge Units were now marked with MOGUL tags. As expected, there were differences between these and the High-level stereotype. Most noticeably, the KUs representing plural noun morphology, third person singular verb morphology, and determiner usage had atypical markings, the first two in the Unacquired range, and the last in the ZPD. This would not be expected in a High-level learner. We now had a MOGUL model reflecting a user who was still classified as High but who had those fossilized linguistic difficulties which were inconsistent with High-level stereotype expectations.

Following this inspection, the hard-wired parse selection was removed in order to investigate whether the acquired history would positively affect the selected parses. The sentences were given to ICICLE one final time. This time, the scoring process was not relying upon just the stereotype, but also upon this "history" we had constructed.

The difference between the choices based on the stereotype alone versus those making use of this stored performance history on MOGUL are summarized in Figure 2. The number of optimal parses receiving the maximum score rose to 81%. More specifically, although the "typical" sentences lost one in that category (as may be expected, because many of these sentences exhibit structures which this user has not fully mastered, such as noun pluralization), the percentage of atypical sentences where the optimal parse received the maximum rose from 50% to 79%. In 21/26 sentences overall, the optimal parse now received top marks.

In these results, the point is not the accuracy of the analysis (this would be an instance of testing on the training set). What we have shown is that the analysis changes in the face of the performace history. This change allows the system to begin to recognize the atypical structures that a writer at a particular stereotype level may exhibit; a history of correctly parsed atypical constituents better enables the system to correctly parse sentences containing similar atypical constituents. The user modeling component then has the ability to recognize the

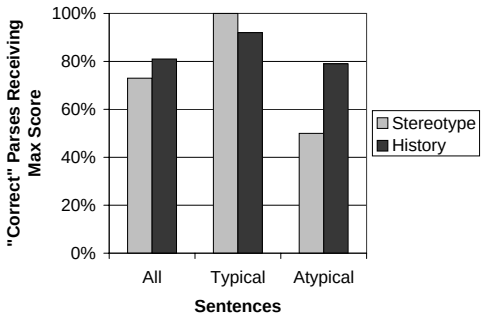

Fig. 2. Summary of ICICLE's Ability to Pick Optimal Parses Given a User History.

validity of parses whose rule usage is consistent with what it knows about the *individual*, not just expectations based on a population, using an integration of the SLALOM stereotype information and the MOGUL individual performance data.

3 Conclusions

Explicitly-modeled errors in a parsing grammar have been used in several other systems. However, the difficulty of handling the ambiguity resulting from the application of such a grammar to parsing in an unrestricted domain is well-known, resulting in the fact that most existing CALL systems restrict their task to well-defined domains (BELLOC, [3]), the parsing of prompted translations (HyperTutor, [8]), or specific subsets of the syntactic spectrum such as pronouns (Mr. Collins, [2]). Perhaps because ICICLE is such an ambitious project with a large and broadly-defined domain, our user modeling effort is far more precise than what can be found in most comparable language instruction systems.

The evaluative runs discussed in this paper illustrate that the ICICLE parse selection mechanism scores and selects trees appropriately given a profile of expected user performance, and that the adaptive nature of the model allows it to shift to adapt to differences in user behavior. They also clearly illuminate paths toward future improvement. Because there were several instances where the system gave the maximum score to many trees, the need for even more intelligent scoring is clear. While parse node scoring on the basis of rule membership in I_i is helpful for the selection of appropriate parse trees, taken alone it does not discriminate strongly enough; in some cases, the number of trees obtaining the highest score is fairly large. In fact, rule membership in I_i is only *part* of what signifies *the most likely tree*; other factors must be taken into account. Future improvements to the system may include taking into account the likelihood of part-of-speech tags on the lexical items in the utterance.

Another future direction for this work is to show whether ICICLE can recognize correct parses for atypical constituents when they first occur (and thus

create a performance history). This would involve providing a greater body of writing from a high-level learner containing some atypical errors, and then testing if, over time, the user model comes to correctly reflect the user's unique language profile and thus to correctly parse subsequent input from this user. This is planned for future analysis.

The evaluation discussed in this paper shows, however, that the design of the ICICLE user model has already found success in melding stereotypical and individual user information, creating a dynamic form which poses a novel approach to the challenge of ambiguity in the natural language task.

4 Acknowledgments

This work has been supported by NSF Grants #GER-9354869 and #IIS-9978021.

References

[1] James Allen. *Natural Language Understanding.* Benjamin/Cummings, California, second edition, 1995.

[2] Susan Bull, Paul Brna, and Helen Pain. Extending the scope of the student model. *User Modeling and User-Adapted Interaction*, 5(1):45–65, 1995.

[3] Thierry Chanier, Michael Pengelly, Michael Twidale, and John Self. Conceptual modelling in error analysis in computer-assisted language learning systems. In M. L. Swartz and M. Yazdani, editors, *Intelligent Tutoring Systems for Second-Language Learning*, volume F80 of *NATO ASI Series*, pages 125–150. Springer-Verlag, Berlin Heidelberg, 1992.

[4] Lisa N. Michaud. *Modeling User Interlanguage in a Second Language Tutoring System for Deaf Users of American Sign Language.* PhD thesis, Dept. of Computer and Information Sciences, University of Delaware, 2002. Tech. Report #2002-08.

[5] Lisa N. Michaud and Kathleen F. McCoy. Error profiling: Toward a model of english acquisition for deaf learners. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 386–393, Toulouse, France, July 5-11 2001. ACL.

[6] Lisa N. Michaud and Kathleen F. McCoy. Empirical derivation of a sequence of user stereotypes. *User Modeling and User-Adaptive Interfaces*, to appear.

[7] Lisa N. Michaud, Kathleen F. McCoy, and Litza A. Stark. Modeling the acquisition of English: an intelligent CALL approach. In *Proceedings of the 8th International Conference on User Modeling*, pages 14–23, Sonthofen, Germany, July 13-17 2001. Springer.

[8] Ethel Schuster and Jennifer Burckett-Picker. Interlanguage errors becoming the Target Language through student modeling. In *Proceedings of the Fifth International Conference on User Modeling*, pages 99–103, Kailua-Kona, Hawaii, January 2-5 1996. UM96, User Modeling, Inc.

[9] Larry Selinker. Interlanguage. *International Review of Applied Linguistics*, 10(3):209–231, August 1972.