Automated Rating of ESL Essays

Deryle Lonsdale
BYU Linguistics and English Language
lonz@byu.edu

Diane Strong-Krause
BYU Linguistics and English Language
diane_strong-krause@byu.edu

Abstract

To date, traditional NLP parsers have not been widely successful in TESOL-oriented applications, particularly in scoring written compositions. Re-engineering such applications to provide the necessary robustness for handling ungrammatical English has proven a formidable obstacle. We discuss the use of a non-traditional parser for rating compositions that attenuates some of these difficulties. Its dependency-based shallow parsing approach provides significant robustness in the face of language learners' ungrammatical compositions. This paper discusses how a corpus of L2 essays for English was rated using the parser, and how the automatic evaulations compared to those obtained by manual methods. The types of modifications that were made to the system are discussed. Limitations to the current system are described, future plans for developing the system are sketched, and further applications beyond English essay rating are mentioned.

1 Introduction

Rating constructed response items, particularly essays, is a time-consuming effort. This is true in rating essays written by second-language speakers. To make this process more manageable, researchers have investigated how to involve computers in the grading process. Several factors suggest why automating scoring might be desirable: (i) *practicality*: essay grading is costly and time-consuming; (ii) *consistency*: essay grading is somewhat subjective in nature, and consistency may sometimes suffer; and (iii) *feedback*: Providing feedback to a student is important, and automated scoring can provide ways of generating specific suggestions tailored to the needs of the author.

However, computerized rating of essays written by second-language speakers poses unique dilemmas, particularly for responses written by examinees at low levels of language proficiency. Where we expect generally well-formed sentences from native English speaker responses, we find that the majority of the responses by lower proficiency second-language English speakers will be made up of ill-formed sentences.

Previous work in automated essay grading and related technologies has been surveyed and discussed in several different forums (Burstein and Chodorow, 1999; Thompson, 1999; Hearst, 2000; Williams, 2001; Rudner and Gagne, 2001), and a thorough survey of the field has recently been published (Shermis and Burstein, 2003). Typically these approaches have borrowed techniques and tools from several natural language processing (NLP) fields. For example, the knowledge-based engines have been used for analyzing essays: parsers (Carbonell and Hayes, 1984; Schneider and McCoy, 1998), grammar and spelling checkers (Park et al., 1997), discourse processing analyzers (Miltsakaki and Kukich, 2000), and other hand-crafted linguistic knowledge sources.

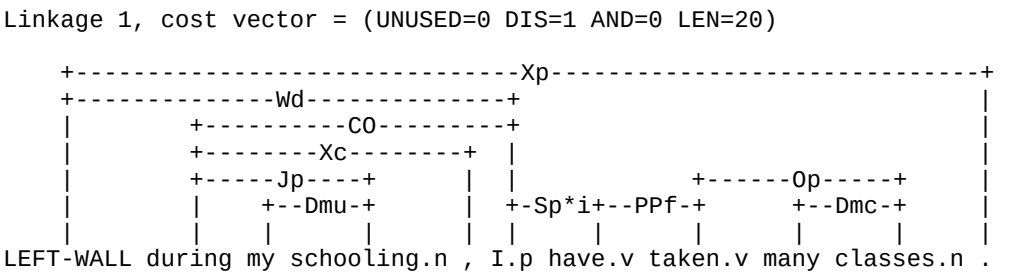

Figure 1: Sample link-parsed sentence with associated cost vector.

On the other hand, much work has leveraged statistical methods in detecting properties of student essays via stylometrics (Aaronson, 2001)[1], latent semantic indexing (Wiemer-Hastings et al., 1998), and feature analysis.

Finally, mirroring noteworthy progress in other NLP fields involving data-driven methods, recent work has involved essay grading via exemplar-based machine learning techniques (Chodorow and Leacock, 2000).

The most visible systems implement one (or more) of these approaches. The Project Essay Grade (PEG) system, for example, uses lexically-based metrics in scoring (Page, 2003). The Intelligent Essay Assessor (IEA) uses latent semantic analysis in calculating its metrics (Landauer et al., 2003). The E-Rater system by Educational Testing Services uses syntactic, discourse, and topical (i.e. vocabulary-based) data analysis (Burstein, 2003).

Several criticisms have been aimed at automatic scoring systems on both theoretical and implementational grounds. For example, many systems exhibit an inherent Achilles' heel since it is possible to trick them into evaluating a nonsensical text purely by reverse-engineering the scoring mechanism and designing a text that responds to the criteria. Another problem is the cost of development, which can be substantial. In addition, most systems are designed around certain specific topics in order to focus terminology and vocabulary in limited subdomains; introducing new subject areas requires building a new model, often a nontrivial process. Thus, many systems are often not adaptable enough to meet the particular needs of an individual, class, teacher, or institution.

[1]This work also uses the Link Grammar parser.

The purpose of our research is to explore the use of a particular natural language processing (NLP) approach for automated scoring of essays written by ESL students at lower levels of language proficiency Our goal for the system reflects commonsense (though ambitious) criteria: to have the system's scores agree with those assigned by human raters at least as often as human raters agree among themselves.

2 The parser

As mentioned above, one approach to grading is to use a parsing system, along with its associated knowledge sources, to analyze the correctness of a text. The NLP field has produced a wide range of parsers and grammars to support them. Most of the widely used and highly accurate parsers are closely (or even inalienably) tied to a particular syntactic theory: XTAG with Tree-Adjoining Grammar[2], LFG-WB with Lexical-Functional Grammar[3], ALE with HPSG or Categorial Grammar[4], and so on. Principled coverage of grammatical phenomena can thus be tied closely to the linguistic theory in question. Some parsers are designed to skip over ungrammatical and disfluent portions of input and have been successfully applied to speech and dialogue processing (Rosé, 1997), with perhaps possible future application to rating ESL essays. There are disadvanges to traditional parsers, though, which offset their usefulness for automated grading.

Consider, for example, that the encoding of any parser's phrase-structure component is costly, complex, and dependent on significant lexical resources. This precludes involvement of the uninitiated. Even

[2]See http://www.cis.upenn.edu/~xtag/

[3]See http://www2.parc.com/istl/groups/nltt/medley/

[4]See http://www.cs.toronto.edu/~gpenn/ale.html

more serious is the lack of robustness that most parsers entail. Most linguistic formalisms focus precisely on what is grammatical, and not on what is ungrammatical. This often becomes an architectural assumption in the way parsers are implemented. The result is that such systems are rather inflexible, particularly in the face of ungrammatical input—ungrammaticaly is almost always avoided in both the theory and in its implementation. Yet crucially for the essay grading ungrammatical input is frequent and expected.

One method used to sidestep the robustness issue is to explicitly encode rules reflecting ungrammaticality, called "mal-rules" (McCoy et al., 1996). For example, the following is a possible mal-rule for an LFG parser:

```
S  --> NP (agr ?a) VP (agr ~?a)
```

This rule says that a sentence can consist of an NP and a VP whose respective agreement features do NOT agree. While such a technique allows for detection of ungrammatical sentences, it introduces two problems. First, the computational complexity of a parsing system increases as such rules are added to the phrase-structure component. Second, maintaining a knowledge base of such information is a complicated and never-ending proposition, as student errors vary in a seemingly infinite number of ways.

In our work we chose to use a different kind of parser, the link grammer parser (Sleator and Temperley, 1993). This parser has been developed for robust, efficient processing of dependency-style syntax (Grinberg et al., 1995). Freely available for research purposes, it is more robust than traditional parsers and has been widely used in such NLP applications as information retrieval, speech recognition, and machine translation[5]. Written in the C programming language, it is comparatively fast and efficient.

The link grammar parser does not seek to construct constituents in the traditional linguistic sense—instead, it calculates simple, explicit relations between pairs of words. A link is a targeted relationship between two words and has two parts: a left side and a right side. For example, links associate such word pairs as: subject + verb, verb + object, preposition + object, adjective + adverbial modifier, and auxiliary + main verb. Each link has a label that expresses the nature of the relationship mediating the two words. Potential links are specified by a set of technical rules that constrain and permit word-pair associations. In addition, it is possible to score individual linkages and to penalize unwanted links.

Figure 1 shows an example link parse of a sentence from a student essay. Ten links of various types span the various relationships observable in the sentence.

When parses are not possible, the system's robustness allows it to discard words (or alternatively posit spelling corrections) in order to arrive at a tenable description of the input. Figure 2 shows two ungrammatical sentences that the parser has nonetheless coped with. In the first, it skips over words that don't seem to fit into any grammatical pattern, parsing instead a core sentence "The class is mathematical." The cost vector for this sentence records the fact that there were 4 unused words. In the second example, only one word must be discarded to arrive at a reasonable parse.

The LG parser as distributed was not completely suited to handle the grading of ESL students' essays, so some modifications had to be made. Lexical items had to be added to the system's lexicon to cover terms frequently used by students, such as acronyms: E.L.C. (the English Language Center), R.O.C. (Republic of China), and so on. Other constructions not supported in the standard release were also added, for example variant ordering within dates (e.g. 24 May as well as May 24). The grammar as originally distributed did not allow for optional commas where unexpected. It also did not penalize certain ungrammatical constructions (e.g. missing determiners, as in "I am student of English.") since such constructions were not anticipated.

With the system slightly modified as described above, it was well suited to parsing ESL essays. Two more example parses of student essay sentences are illustrated in Figure 3. In the next section we discuss how it was used to score such essays.

3 The corpus

Our study involved using the LG parser to rate essays based on the results of a link parse for each sentence. We used ESL essays written by Intensive En-

[5]For a bibliography see http://link.cs.cmu.edu/link/papers/

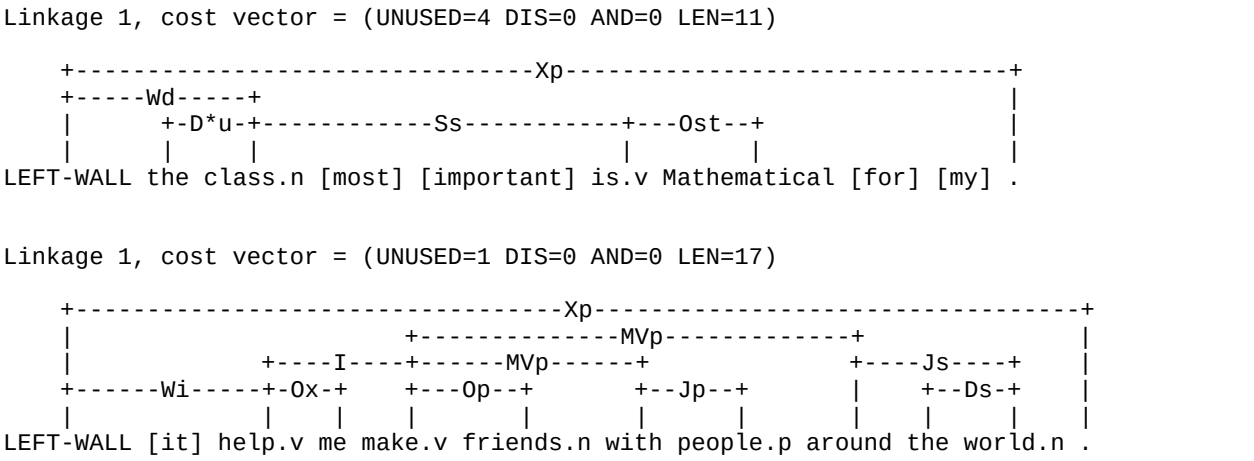

Figure 2: Link parser results for highly ungrammatical sentences. Note the discarded words indicated in square brackets.

glish students who spanned a range of ability from Novice-mid to Intermediate-high. The essays were on a variety of assigned topics, and each had to be written within a 30-minute time limit.

Our corpus consists of 301 human-rated essays in total consisting of some 50,000 words and 3400 sentences. These were sub-divided further by semester into 5 sub-corpora. The essays exhibited the following characteristics: each had (on average) 165 words and 11.2 sentences, and each sentence had on average 14.75 words. Note the wide variety of errors in this typical sentence from one essay:

> Iwork really hard and occacionally
> I don't have time for have fun
> whith mt friens but i don't mind
> becausse i knew ,when i grow up
> i will have a profesion and have
> a good job and i will be very
> happy.

Each essay was given a holistic rating by two human judges. Different raters participated each semester, though there was likely a small degree of overlap among raters across subsequent semesters.

Scores ranged from 1 to 5 with half-points possible (i.e. essays could receive 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, and 5). Occasionally a judge gave a rating of 0 indicating that no comprehensible language was present. Inter-rater agreement, where each human assigned a score within one point of the other, was 98% over the corpora.

The following categories describe scoring levels:

1. Demonstrates limited ability to write English words and sentences. Sentences and paragraphs may be incomplete and difficult to follow.

2. Writes a simple paragraph with a fair control of basic, not complex, sentences structures. Errors occur in almost all sentences except for the most basic, formula-type (memorized) structures. Little detail is present.

3. Writes a fairly long paragraph with relatively simple sentence structures. Personal experiences and some emotions can be expressed, but much detail is missing. Frequent errors in grammar and word use make reading somewhat burdensome.

4. Writes long groups of paragraphs with some complex sentence patterns. Some transitions are used effectively. Vocabulary is broadening, but some wrong word use. Grammar errors may detract from meaning. Some ideas are supported with detail. Some notion of an introduction and conclusion is included.

5. Writes complex thoughts using complex sentence patterns with effective transitions between ideas and sentences. Errors in grammar exist but do not obscure meaning. A variety of advanced vocabulary words are used

but some wrong use occurs, including problems with prepositions and articles. Ideas are clearly supported with details. Effective introduction and conclusion are included.

Although judges gave a holistic rating based on a number of features of the essay, the category descriptions hint that syntax (and to some degree vocabulary use) is the focus for at least categories 1 through 3, and even much of category 4.

4 Results and analysis

The corpus was partitioned into two classes: the development set and the test set. The former was used to develop and tune the system, and consisted of the most recent set of essays (60 in number) dating from Winter 2002 semester. The other 4 (earlier) corpora were used for testing the system.

Each essay was sent through the LG parser a sentence at a time, and each sentence was given a 5-point score based on the parse's cost vector. An overall score for each essay was computed as the average across sentence scores, after discarding the lowest and highest ones. The overall score was then compared with those of the human raters. For the development set, the system agreed 67% of the time with human raters, where agreement follows the standardly accepted definition of falling within 1.0 of the closest human's score. Note that since humans only gave ratings involving integers and half-steps between them, all computer-generated scores were rounded to the nearest integer or half-step as appropriate.

The system was then run on the previously unseen test corpus, actually consisting of essays from four separate semesters. The test corpus scores were as follows:

Fall01:	82 students	69.5% agreement
Summer01:	58 students	62.1% agreement
Winter01:	36 students	66.7% agreement
Winter92:	75 students	62.2% agreement

Hence over a corpus of some 300 essays and five sub-corpora, system agreement with human raters was achieved about 66% of the time. We now turn to a brief analysis of interesting results that emerged from the system's performance.

Generally, the system tended to over-score essays with very low human scores (i.e. those in the 1-2 point range). It also tended to under-score essays with high human scores and complex run-on sentences. This reflects the observation that run-on sentences, which were very plentiful, were penalized by the system but largely forgiven by human raters. Also, the system's scoring matched human values better for midrange-scored essays, and worse for extreme examples (i.e. with average score < 2 or > 4.5). Finally, system panics (when the system ran out of allotted time without successfully parsing a sentence) occurred most frequently when several conjunctions appeared in a single sentence.

It is informative to look at an essay that reflects one of the most extreme mismatches between human and computer ratings. In this case, the two human raters gave the essay scores of 1 and 2 respectively, whereas the computer scored the essay at 4.40.

```
My free time is very fun.  Because
I meet my friends.  We goes to
play.  For exmple, I went to
movies, recreation ground, trip
and shopping with them.  I can't
write English.
```

Another observation from the present work is that performing a purely syntactic parse does not always assure appropriate ratings. The current system's scoring mechanism occasionally results in artificially high scores. Consider, for example, the sentence in Figure 4. Even though there are no egregious syntactic errors, violations of selectional restriction, collocation, determiner selection, and verbal aspect render the sentence highly unnatural, though this is undetected by the current parser. Addition of hand-coded postprocessor rules may help avoid such situations, and is possible with the parser.

5 Future work

There are several ways in which the base system described in this paper can be improved. For instance, sentence and essay scores are currently based on straightforward values from the cost vector, whereas more sophisticated measures can be implemented. Future work will involve using statistical smoothing to improve performance in the extreme (high-scoring and low-scoring) situations.

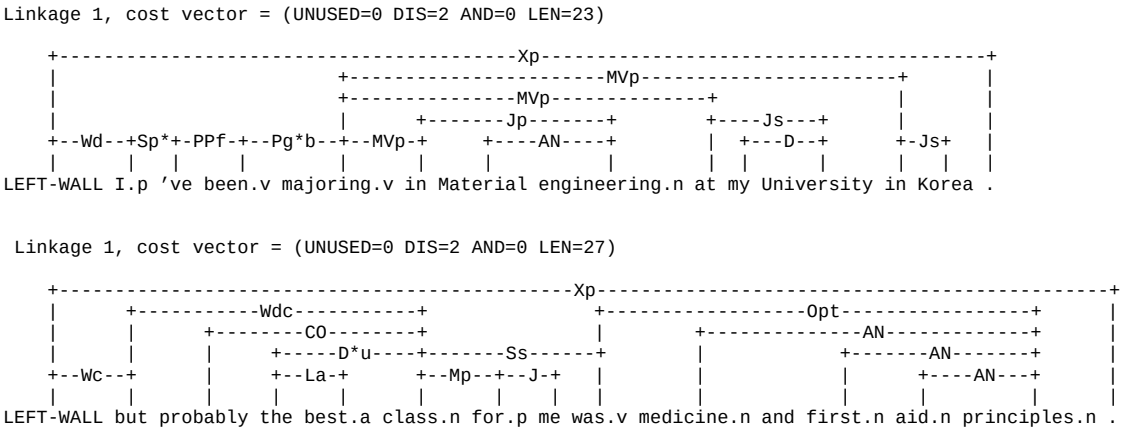

Figure 3: Sample link-parsed sentences from the student essay corpus.

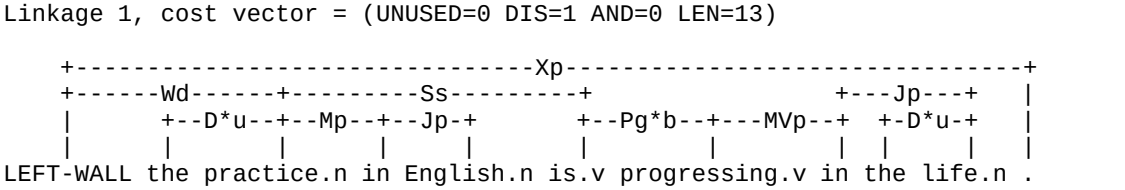

Figure 4: Sentence illustrating collocational and selectional problems.

The system could also achieve more human-like scoring by integrating data-driven, exemplar-based approaches. Training the system to relate salient features and vector costs of the essays with the corresponding human scores can be done using any of a variety of available techniques, such as memory-based learning or analogical modeling.

Finally, further linguistic processing can improve the system. Some syntactic improvements can be made, including specifying licit structures that are not recognized as well as unacceptable structures with their associated coses. As mentioned above, pushing the analysis beyond a syntactic LG parse will be necessary. This might involve leveraging resources such as WordNet (Fellbaum, 1998) for providing lexical semantic information which could prove useful in scoring compounding strategies, collocations, and verb argument structure formation. Current research in discourse processing using such devices as centering, anaphor/coreference, coherence, and topic continuity can also be integrated into the system as has been done in other scoring programs.

In addition, the LG parser is also being developed for other languages (e.g. French and Spanish) with the eventual goal of providing a scoring engine for learners of these languages as well.

6 Conclusions

We have shown how the output from a non-traditional syntactic parser can be used to grade ESL essays. With a robust enough parser, reasonable results can be obtained, even for highly ungrammatical text. We anticipate that this foundation can be improved upon by using other commonly adopted NLP techniques. Its applicability should extend to other languages besides English given a comparable LG parser implementation.

References

Scott Aaronson. 2001. Stylometric clustering: a comparison of data-driven and syntactic features. Manuscript. http://www.cs.berkeley.edu/~aaronson/sc.doc.

Jill Burstein and Martin Chodorow. 1999. Automated essay scoring for nonnative English speakers. In *Computer Mediated Language Assessment and Evaluation in Natural Language Processing*, pages 68–75. Association for Computational Linguistics.

Jill Burstein. 2003. The E-*rater* scoring engine: Automated essay scoring with natural language processing. In Mark D. Shermis and Jill C. Burstein, editors, *Automated Essay Scoring: A Cross-Disciplinary Perspective*. Lawrence Erlbaum, Mahwah, NJ.

Jaime G. Carbonell and Phillip J. Hayes. 1984. Coping with extragrammaticality. In *Proceedings of COLING '84*, pages 437–443. Association for Computational Linguistics.

Martin Chodorow and Claudia Leacock. 2000. An unsupervised method for detecting grammatical errors. In *Proceedings of ANLP-NAACL 2000*, pages 140–147. Morgan Kaufmann Publishers.

Christiane Fellbaum. 1998. *WordNet: An electronic lexical database*. MIT Press, Cambridge, MA.

Dennis Grinberg, John Lafferty, and Daniel Sleator. 1995. A robust parsing algorithm for Link Grammars. Technical Report CMU-CS-95-125, School of Computer Science, August.

Marti A. Hearst. 2000. The debate on automated essay grading. *IEEE Intelligent Systems*, September/October 2000:22–37.

Thomas Landauer, Darrell Laham, and Peter Foltz. 2003. Automated scoring and annotation of essays with the Intelligent Essay Assessor. In Mark D. Shermis and Jill C. Burstein, editors, *Automated Essay Scoring: A Cross-Disciplinary Perspective*. Lawrence Erlbaum, Mahwah, NJ.

Kathleen F. McCoy, Christopher A. Pennington, and Linda Z. Suri. 1996. English error correction: A syntactic user model based on principled "mal-rule" scoring. In *Proceedings of the Fifth International Conference on User Modeling*, pages 59–66. User Modeling, Inc.

Eleni Miltsakaki and Karen Kukich. 2000. The role of centering theory's rough-shift in the teaching and evaluation of writing skills. In *Proceedings of ACL-2000*. Association for Computational Linguistics.

Ellis Batten Page. 2003. Project Essay Grade: PEG. In Mark D. Shermis and Jill C. Burstein, editors, *Automated Essay Scoring: A Cross-Disciplinary Perspective*. Lawrence Erlbaum, Mahwah, NJ.

Jong C. Park, Martha Palmer, and Gay Washburn. 1997. An English grammar checker as a writing aid for students of English as a Second Language. In *Proceedings of the Conference of Applied Natural Language Processing (ANLP)*.

Carolyn Penstein Rosé. 1997. *Robust Interactive Dialogue Interpretation*. Ph.D. thesis, School of Computer Science, Carnegie Mellon University.

Lawrence Rudner and Phill Gagne. 2001. An overview of three approaches to scoring written essays by computer. *Practical Assessment, Research & Evaluation*, 7(26).

David Schneider and Kathleen McCoy. 1998. Recognizing syntactic errors in the writing of second language learners. In *Proceedings of COLING-ACL 1998*, pages 1198–1204. Morgan Kaufmann Publishers.

Mark D. Shermis and Jill C. Burstein, editors. 2003. *Automated Essay Scoring: A Cross-Disciplinary Perspective*. Lawrence Erlbaum, Mahwah, NJ.

Daniel Sleator and Davy Temperley. 1993. Parsing English with a Link Grammar. In *Third International Workshop on Parsing Technologies*.

Clive Thompson. 1999. New word order: The attack of the incredible grading machine. *Linguafranca: The Review of Academic Life*, 9(5).

Peter Wiemer-Hastings, Arthur C. Graesser, and Derek Harter. 1998. The foundations and architecture of AutoTutor. *Lecture Notes in Computer Science*, 1452:334–343.

Robert Williams. 2001. Automated essay grading: An evaluation of four conceptual models. In *Expanding Horizons in Teaching and Learning: Proceedings of the 10th Annual Teaching Learning Forum*. Curtin University of Technology.