Automatic Grammar Correction for Second-Language Learners

John Lee, Stephanie Seneff

Spoken Language Systems
MIT Computer Science and Artificial Intelligence Laboratory
Cambridge, MA 02139 USA
{jsylee, seneff}@csail.mit.edu

10.21437/Interspeech.2006-542

Abstract

A computer conversational system can potentially help a foreign-language student improve his/her fluency through practice dialogues. One of its potential roles could be to correct ungrammatical sentences. This paper[1] describes our research on a sentence-level, generation-based approach to grammar correction: first, a word lattice of candidate corrections is generated from an ill-formed input. A traditional n-gram language model is used to produce a small set of N-best candidates, which are then reranked by parsing using a stochastic context-free grammar. We evaluate this approach in a flight domain with simulated ill-formed sentences. We discuss its potential applications in a few related tasks.

Index Terms: computer-assisted language learning, dialogue systems, natural language generation.

1. Introduction

In the past few years, our group has been developing a conversational language learning system [1], which engages students in a dialogue in order to help them learn a foreign language. An important component of such a system is to provide corrections of the students' mistakes, both phonetic [2] and grammatical, the latter of which is the focus of this paper. For example, the student might say, "**I will like to see flight arrive Dallas next day.*" The system would be expected to correct this to, "*I would like to see flights arriving in Dallas the next day.*"

An important point to consider is that the system need not feel obliged to alert the student of every error it detects. In usage, it is anticipated that the student would first engage in an interactive dialogue with the system, during which it would conceivably apply an error-correction algorithm in order to increase the probability of obtaining a correct meaning analysis. Any differences between the "corrected" hypothesis and the original input would be recorded in a log file, along with associated parse scores. In a follow-up interaction, the system would provide explicit feedback about the previous dialogue. It could afford to be selective at this point, informing the student only of errors where it has high confidence. This conservative approach would greatly reduce the likelihood that the system misinforms the student.

[1]This work is sponsored by the Department of Defense under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the United States Government. This work is also supported in part by the Natural Sciences Engineering Research Council of Canada. We thank Chao Wang and the four judges, Ed Filisko, Alex Gruenstein, Mitch Peabody and Daniel Schultz.

Previous approaches to grammar correction focus on parsing ill-formed sentences. To cope with grammatical errors, new mechanisms are incorporated into parsers which, otherwise, are intended for analyzing well-formed sentences. One example is *constraint relaxation* in a unification framework, such as in [3] and [4]. The constraints, such as subject-verb agreement, are progressively relaxed, until the sentence can be parsed. A correction can then be easily generated by examining the violated constraints.

Another example is the use of error-production rules in context-free grammars, such as in ICICLE [5], a broad-coverage system designed for teaching English to American Sign Language signers. In ARBORETUM [6], such rules are used in conjunction with an aligned generation strategy, so that "the corrected form should match the input in all ways except those affected by the correction." In this way, the grammatical errors may be clearly shown to the student.

A disadvantage with the above *parsing-based* approaches is that, as more and more types of errors need to be handled, the grammars become increasingly complicated, exponentially growing the number of ambiguous parses. We instead propose a two-step, *generation-based* framework. Given a possibly ill-formed input, the first step paraphrases the input into an over-generated word lattice, licensing possible corrections; and the second step utilizes language models and parsing to select the best rephrasing. This approach sidesteps the need to model ungrammaticality in the natural language understanding (NLU) component.

The rest of the paper is organized as follows. §2 identifies the types of errors we intend to handle and describes our two-step, generation-based framework for grammar correction. §3 presents some experiments on a corpus of flight queries. §4 reviews previous related research. §5 sketches our future directions.

2. Correction Scope and Procedure

According to the Japanese Learners' English corpus [7], which consists of transcripts of native Japanese speakers conversing in English, the three most frequent error classes are articles, noun number and prepositions, followed by a variety of errors associated with verbs. Motivated by this analysis, we consider errors involving these four parts-of-speech:

- All **articles** and ten **prepositions**, listed in Table 1.
- **Noun** number.
- **Verb** aspect, mode, and tense.

In the first step, an input sentence, hypothesized to contain errors, is first reduced to a "canonical form" devoid of articles,

Part-of-speech	Words
Articles	a, an, the
Modals, Verb aux.	can, could, will, would, must, might, should be, have, do
Prepositions	about, at, by, for, from, in, of, on, with, to
Nouns	flight, city, airline, friday, departure, ...
Verbs	like, want, go, leave, take, book, ...

Table 1: *The parts-of-speech and the words that are involved in the experiments described in §2. The five most frequently occurring (in the test set) nouns and verbs are listed in their base forms. Other lists are exhaustive.*

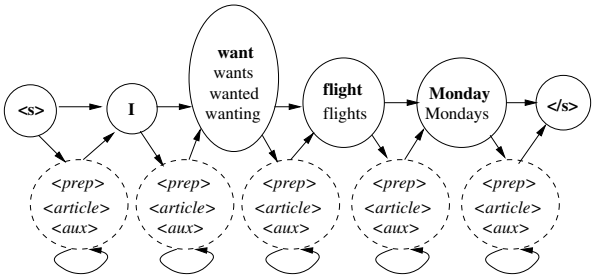

Figure 1: *Lattice of alternatives formed from the reduced input sentence, "I want flight Monday". The lattice encodes many possible corrections, including different noun and verb inflections, and insertions of prepositions, articles, and auxiliaries. One appropriate correction is "I want a flight on Monday".*

prepositions, and auxiliaries ("can," "would," "be," etc.). Furthermore, all nouns are reduced to their singular forms, and all verbs are reduced to their root forms. All of their alternative inflections are then inserted into the lattice in parallel. Insertions of articles, prepositions and auxiliaries are allowed at every position. This simple algorithm thus expands the sentence into a lattice of alternatives, as illustrated in Figure 1, for the reduced input sentence "*I want flight Monday*".

In the second step, a language model is used to score the various paths in the lattice. Both n-grams and a stochastic context-free grammar language model are utilized in a complementary fashion.

A grammar, which was originally designed for our flight domain spoken dialogue system [8], incorporates both syntactic and semantic information into the grammar rules. We use the TINA [9] framework to perform the parsing step. A set of probabilistic context-free rules describes the sentence structure, and a constraint-unification mechanism handles feature agreement and movement phenomena. The probability model is applied to nodes in the parse tree, where each node's category is conditioned on its parent and left sibling. The statistics are trained on a large corpus of in-domain utterances.

3. Training and Evaluation

3.1. Experimental Set-Up

Because our methods first reduce a sentence to an impoverished, uninflected form, we can both train the system and evaluate its performance by applying it to a corpus collected from a general population. We generated reduced sentences using the scheme described in the first step in §2. We then measured the system's ability to recover the original function words (articles, modals, prepositions) and to produce correct inflectional forms for both nouns and verbs. This set-up provides an idealized situation: the error classes in the data are essentially restricted to the classes we model. Thus we are able to measure the effects of the reranking algorithm in a controlled fashion.

We also performed individual experiments on each class. For example, we removed all the articles from the sentences, while retaining all other words. A similar procedure was repeated for the prepositions, verbs and nouns.

3.2. Data Source

The training set consists of 10,369 transcripts of utterances, produced by callers in spoken dialogues with the MERCURY flight domain [8]. The test set consists of 1317 sentences from the same domain, all at least four words long. These utterances, whose average length is 7.6 words, serve as our "gold-standard" in the automatic evaluation, although we recognize that some of the users may be non-native speakers.

3.3. Overgeneration Algorithm

Starting with a backbone lattice consisting of the reduced input sentence, the following operations on the lattice are allowed:

Free Insertions Articles, prepositions, auxiliaries and modals are allowed to be inserted anywhere.

It is possible, based on a shallow analysis of the sentence, to limit the possible points of insertion. However, at least in this restricted domain, these constraints have a negligible effect on the final performance. For example, in the articles-only experiment, over 99% of the time, the correct solution is among the 10-best proposed by the trigram language model.

Noun/Verb Inflections The nouns and verbs, appearing in their uninflected forms in the reduced input, can be substituted by any of their inflected forms.

There were 92 unique nouns (excluding proper nouns) in the training set, which are seen 733 times in the test set; 79 unique verbs occur 967 times.

3.4. Reranking Strategies

For verbs and nouns, the MAJORITY baseline simply uses the inflected form that occurs most frequently in the training set. In all error classes, the following two reranking strategies are contrasted:

TRIGRAM A word trigram language model is trained from the sentences in the training set.

PARSE The flight domain context-free grammar is trained with the sentences in the training set. The highest scoring parse obtained from parsing the 10-best list produced by the trigram language model is selected. If no parsed hypothesis is obtained, it defaults to the highest scoring trigram hypothesis. Nearly 87% of the 10-best hypotheses[2], contain the "gold-standard".

[2]Naturally, this percentage can be improved with a larger N-best list. However, any significant improvement would come with a trade-off in computing time.

Reranker	Noun Number	Verb Inflection
MAJORITY	84.1	75.4
TRIGRAM	90.2	89.8
PARSE	94.8	92.2

Table 2: *Accuracy results for noun and verb inflectional endings.*

Class	Reranker	Precision	Recall	F-Score
Articles	TRIGRAM	85.7	72.8	0.79
	PARSE	85.7	76.4	0.81
Prep.	TRIGRAM	83.4	70.0	0.76
	PARSE	88.2	78.4	0.83
Aux.	TRIGRAM	90.2	86.1	0.88
	PARSE	91.4	88.8	0.90

Table 3: *Precision/recall for reinstating function word classes.*

3.5. Results

3.5.1. Automatic Evaluation

Results for experiments on the individual part-of-speech error classes are shown in Tables 2 and 3. Table 2 reports inflection accuracies for verbs and nouns, while Table 3 reports precision/recall for prepositions, auxiliaries, and articles, for which free insertions are allowed. Note that the grammar correction task is made easier in these individual experiments. For example, knowing that the noun is plural rules out the use of "*a*" as its article. When all error classes are combined, as expected, the performance level degrades, as shown in Table 4.

In addition, we separately computed the results for the PARSABLE subset, which includes only those utterances for which a parse was produced from the 10-best candidate list (73.5% of the 1317 utterances). The substantially improved performance of this subset suggests that parsability can be used as a pre-condition for offering a correction; i.e., as a confidence score for the quality of its proposed rewording.

Reranking with PARSE achieved higher F-scores than TRIGRAM across all experiments. Two properties of the parser were responsible for this improvement. First, it was able to reject candidate sentences that do not parse, such as "**When does the next flight to Houston?*" or "**When does the next one leaving?*". Trigrams are unable to detect these long-distance dependencies. Second, the parser was able to apply domain knowledge to reject syntactically valid, but semantically implausible sentences, such as "**I would like to fly from Boston in Bangkok.*"

3.5.2. Human Evaluation

As in machine translation, there are often multiple valid corrections for one sentence. To better gauge the performance of the PARSE model, we conducted a human evaluation on the PARSABLE subset. Four native English speakers, not involved in this research, were given the ill-formed input, and were asked to compare the corresponding transcript ("gold-standard") and the PARSE output, without knowing their identities. They were asked to decide whether the two are of the same quality, or that one of the two is better. An example is shown in Table 5.

Reranker (# utterances)	Noun/Verb Accuracy	Aux/Prep/Article Prec.	Recall	F-score
TRIGRAM (1317)	89.2	80.9	67.7	0.74
PARSE (1317)	91.6	85.9	74.1	0.80
PARSABLE (968)	95.1	90.2	84.0	0.87

Table 4: *Experimental results for all error classes combined.*

Reduced input:	when delta flight leave atlanta
Correction 1:	when does the delta flight leave atlanta
Correction 2:	when does the delta flight leave from atlanta

Table 5: *A sample entry in the human evaluation. Correction 1 is the transcript, and correction 2 is the* PARSE *output. Both evaluators judged these to be equally good. In the automatic evaluation, the* PARSE *output was penalized for the insertion of "from".*

Test Set 1	OK	WORSE	Test Set 2	OK	WORSE
OK	101	6	OK	91	3
WORSE	15	45	WORSE	25	41

Table 6: *Agreement in the human evaluation. The test set was randomly split into two halves, Test Set 1 and Test Set 2. Two human judges evaluated Test Set 1, with kappa = 0.72. Two other evaluated Test Set 2, with kappa = 0.63. Both kappa values correspond to "substantial agreement" as defined in [10].*

To measure the extent to which the PARSE output is distinguishable from the transcript, we interpreted their evaluations in two categories: (1) category OK, when the PARSE output is as good as, or better than the transcript; and (2) category WORSE, when the transcript is better. As shown in Table 6, the human judges exhibited "substantial agreement" according to the kappa scale defined in [10].

In the automatic evaluation, 641 of the PARSE outputs are identical to their corresponding transcripts. The remaining 317 sentences were then judged in the human evaluation as either better or as good as the transcript. A summary of the results is provided in Table 7. Over all, 88.7% of the corrections of the sentences in the PARSE set were at least as good as the transcript.

In many cases where the PARSE was judged to be worse, the parse tree contains certain parse tree patterns that are not well modelled in the probability model. Consider the ill-formed sentence "**I want a flight 324.*", whose parse tree is partially shown in Figure 2. Well-formed sentences such as "*I want flight 324*" or "*I want a flight*" have similar parse trees. The only "unusual" combination of nodes is thus the co-occurrence of the `indef` and `flight_number` nodes. The grammar could likely be reconfigured to capture the distinction between a generic flight and a specific flight. Such combinations can perhaps also be expressed as features in a further reranking step similar to the one used in [11].

Human Judgment	Count	Percentage
Identical output	641	66.2%
Same quality or PARSE better	218	22.5%
Transcript better	109	11.3%
Total	968	100%

Table 7: *Of the 968 corrections proposed in the* PARSABLE *subset, 641 were identical to the transcript. The rest were considered in the human evaluation. Overall, 88.7% of the time, they were judged to be of the same quality or better than the transcript.*

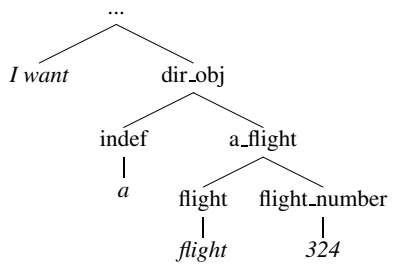

Figure 2: *Parse tree for the ill-formed sentence "*I want a flight 324". The co-occurrence of* `indef` *and* `flight_number` *nodes should be recognized to be a clue for ungrammaticality.*

4. Related Research

4.1. Natural Language Generation

Our generation step in §2 may be viewed as a type of natural language generation (NLG) that is intermediate between conventional NLG from meaning representations, and NLG from keywords.

For the former, the input is either a hierarchical semantic frame, such as in [12] and [13], or a set of attribute-value pairs, such as in [14]. There are two main problems with this type of NLG and grammar correction. First, a meaning representation is difficult to obtain from an ill-formed sentence. Second, in most systems, the same surface string is always hypothesized for the same meaning representation. For example, if the language model determines that `query(time=Monday)` is best generated as "*I want a flight on Monday*", then it will give the same output even if the input is "*I* **would like** *a flight on Monday*". In grammar correction, however, the corrections are to remain as close to the input as possible, a pedagogical principle also followed in [6].

In [15], the input is a sequence of three Japanese keywords, to which particles and connectives are added to form a complete Japanese sentence. This task may be viewed as a grammar correction approach that is similar to ours in the sense of stripping away and then regenerating the function words and inflectional endings.

4.2. Reranking

Various language models have been explored in reranking N-best lists of candidate sentences. Dependency models were used in [14] and [15]. Bigrams were used in [13], and they can be improved upon using a lexicalized syntax model [16]. The benefits of such a model are observed in other tasks, such as [11], as was also observed in our experiments.

5. Conclusions and Future Plans

We presented a generation-based approach for grammar correction, and evaluated this approach in the flight domain using simulated data based on the four most common error classes. Among those sentences that can be parsed by our NLU component, the quality of 88.7% of the corrections offered is indistinguishable from the original transcript.

In this research, we limited our attention to a number of the most common error categories, which were artificially introduced into the transcripts. We would like to collect data from real second-language learners. This data would allow us not only to expand the error-correction coverage, but also to develop a confidence score, which can measure how well the system is able to distinguish grammatical sentences from errorful ones.

We would also like to pursue research in two directions. First, to improve the generation module of an interlingua-based translation system with the techniques outlined here; and second, to investigate the feasibility of scaling these techniques up to a broader domain.

6. References

[1] Seneff, S., Wang, C., Peabody, M., and Zue, V., "Second Language Acquisition through Human Computer Dialogue," *Proc. ISCSLP*, Hong Kong, 2004.

[2] Dong, B., Zhao, Q., Zhang, J. and Yan, Y., "Automatic Assessment of Pronunciation Quality," *Proc. ISCSLP*, pp. 137–140, Hong Kong, 2004.

[3] Fouvry, F., "Constraint Relaxation with Weighted Feature Structures," *Proc. 8th International Workshop on Parsing Technologies*, Nancy, France, 2003.

[4] Vogel, C., and Cooper, R., "Robust Chart Parsing with Mildly Inconsistent Feature Structures," *Nonclassical Feature Systems*, vol. 10, 1995.

[5] Michaud, L., McCoy, K. F. and Pennington, C. A., "An Intelligent Tutoring System for Deaf Learners of Written English," *Proc. 4th International ACM SIGCAPH Conference on Assistive Technologies (ASSETS)*, Arlington, VA, 2000.

[6] Bender, E. M., Flickinger, D., Oepen, S., Walsh, A. and Baldwin, T., "ARBORETUM: Using a Precision Grammar for Grammar Checking in CALL," *Proc. InSTIL/ICALL Symposium on Computer Assisted Learning*, Venice, Italy, 2004.

[7] Izumi, E., Uchimoto, K., Saiga, T., Supnithi, T. and Isahara, H., "Automatic Error Detection in the Japanese Learners' English Spoken Data," *Proc. ACL*, Sapporo, Japan, 2003.

[8] Seneff, S., "Response Planning and Generation in the MERCURY Flight Reservation System,", *Computer Speech and Language*, vol. 16, 2002.

[9] Seneff, S., "TINA: A Natural Language System for Spoken Language Applications," *Computational Linguistics*, 18(1):61–86, 1992.

[10] Landis, J. R. and Koch, G. G., "The Measurement of Observer Agreement for Categorical Data", *Biometrics*, 33(1), p.159–174, 1977.

[11] Collins, M., Roark, B. and Saraclar, M., "Discriminative Syntactic Language Modeling for Speech Recognition," *Proc. ACL*, Ann Arbor, MI, 2005.

[12] Knight, K. and Hatzivassiloglou, V., "Two-Level, Many-Paths Generation," *Proc. ACL*, 1995.

[13] Langkilde, I. and Knight, K., "Generation that Exploits Corpus-based Statistical Knowledge," *Proc. ACL*, 1998.

[14] Ratnaparkhi, A., "Trainable Methods for Surface Natural Language Generation," *Proc. NAACL*, Seattle, WA, 2000.

[15] Uchimoto, K., Sekine, S. and Isahara, H., "Text Generation from Keywords," *Proc. COLING*, 2002.

[16] Daumé, H. III, Knight, K., Langkilde-Geary, I., Marcu, D. and Yamada, K., "The Importance of Lexicalized Syntax Models for Natural Language Generation Tasks," *Proc. INLG*, 2002.