Correcting Grammatical Verb Errors

Alla Rozovskaya  
Columbia University  
New York, NY 10115  
ar3366@columbia.edu

Dan Roth  
University of Illinois  
Urbana, IL 61801  
danr@illinois.edu

Vivek Srikumar  
Stanford University  
Stanford, CA 94305  
v vivek@cs.stanford.edu

Abstract

Verb errors are some of the most common mistakes made by non-native writers of English but some of the least studied. The reason is that dealing with verb errors requires a new paradigm; essentially all research done on correcting grammatical errors assumes a closed set of triggers—e.g., correcting the use of prepositions or articles—but identifying mistakes in verbs necessitates identifying potentially ambiguous triggers first, and then determining the type of mistake made and correcting it. Moreover, once the verb is identified, modeling verb errors is challenging because verbs fulfill many grammatical functions, resulting in a variety of mistakes. Consequently, the little earlier work done on verb errors assumed that the error type is known in advance.

We propose a linguistically-motivated approach to verb error correction that makes use of the notion of verb finiteness to identify triggers and types of mistakes, before using a statistical machine learning approach to correct these mistakes. We show that the linguistically-informed model significantly improves the accuracy of the verb correction approach.

1 Introduction

We address the problem of correcting grammatical verb mistakes made by English as a Second Language (ESL) learners. Recent work in ESL error correction has focused on errors in article and preposition usage (Han et al., 2006; Felice and Pulman, 2008; Gamon et al., 2008; Tetreault et al., 2010; Gamon, 2010; Rozovskaya and Roth, 2010a; Dahlmeier and Ng, 2011).

While verb errors occur as often as article and preposition mistakes, with a few exceptions (Lee and Seneff, 2008; Gamon et al., 2009; Tajiri et al., 2012), there has been little work on verbs. There are two reasons for why it is difficult to deal with verb mistakes. First, in contrast to articles and prepositions, verbs are more difficult to identify in text, as they can often be confused with other parts of speech, and processing tools are known to make more errors on noisy ESL data (Nagata et al., 2011). Second, verbs are more complex linguistically: they fulfill several grammatical functions, and these different roles imply different types of errors.

These difficulties have led all previous work on verb mistakes to assume prior knowledge of the mistake type; however, identifying the specific category of a verb error is nontrivial, since the surface form of the verb may be ambiguous, especially when that verb is used incorrectly. Consider the following examples of verb mistakes:

1. “We discusses*/discuss this every time.”
2. “I will be lucky if I {will find}*find something that fits.”
3. “They wanted to visit many places without spending*spend a lot of money.”
4. “They arrived early to organize*/organize everything”.

These examples illustrate three grammatical verb properties: Agreement, Tense, and non-finite Form choice that encompass the most common grammatical verb problems for ESL learners. The first two examples show mistakes on verbs that function as main verbs in a clause: sentence (1) shows an example of subject-verb Agreement error; (2) is an example of a Tense mistake where the ambiguity is between {will find} (Future tense)
and find (Present tense). Examples (3) and (4) display Form mistakes: confusing the infinitive and gerund forms in (3) and including an inflection on an infinitive verb in (4).

This paper addresses the specific challenges of verb error correction that have not been addressed previously – identifying candidates for mistakes and determining which class of errors is present, before proceeding to correct the error. The experimental results show that our linguistically-motivated approach benefits verb error correction. In particular, in order to determine the error type, we build on the notion of verb finiteness to distinguish between finite and non-finite verbs (Quirk et al., 1985), that correspond to Agreement and Tense mistakes (examples (1) and (2) above) and Form mistakes (examples (3) and (4) above), respectively (see Sec. 3). The approach presented in this work was evaluated empirically and competitively in the context of the CoNLL shared task on error correction (Ng et al., 2013) where it was implemented as part of the highest-scoring University of Illinois system (Rozovskaya et al., 2013) and demonstrated superior performance on the verb error correction sub-task.

This paper makes the following contributions:

- We present a holistic, linguistically-motivated framework for correcting grammatical verb mistakes; our approach “starts from scratch” without any knowledge of which mistakes should be corrected or of the mistake type; in doing that we show that the specific challenges of verb error correction are better addressed by first identifying the finiteness of the verb in the error identification stage.
- Within the proposed model, we describe and evaluate several methods of selecting verb candidates, an algorithm for determining the verb type, and a type-driven verb error correction system.
- We annotate a subset of the FCE data set with gold verb candidates and gold verb type.\(^1\)

2 Related Work

Earlier work in ESL error correction follows the methodology of the context-sensitive spelling correction task (Golding and Roth, 1996; Golding and Roth, 1999; Banko and Brill, 2001; Carlson et al., 2001; Carlson and Fette, 2007). Most of the effort in ESL error correction so far has been on article and preposition usage errors, as these are some of the most common mistakes among non-native English speakers (Dalgish, 1985; Leacock et al., 2010). These phenomena are generally modeled as multiclass classification problems: a single classifier is trained for a given error type where the set of classes includes all articles or the top \(n\) most frequent English prepositions (Izumi et al., 2003; Han et al., 2006; Felice and Pulman, 2008; Gamon et al., 2008; Tetreault et al., 2010; Rozovskaya and Roth, 2010a; Rozovskaya and Roth, 2011; Dahlmeier and Ng, 2011).

Mistakes on verbs have attracted significantly less attention in the error correction literature. Moreover, the little earlier work done on verb errors only considered subsets of these errors and assumed the error sub-type is known in advance. Gamon et al. (2009) mentioned a model for learning gerund/infinitive confusions and auxiliary verb presence/choice. Lee and Seneff (2008) proposed an approach based on pattern matching on trees combined with word n-gram counts for correcting agreement misuse and some types of verb form errors. However, they excluded tense mistakes, which is the most common error category for ESL learners (40% of all verb errors, Sec. 3). Tajiri et al. (2012) considered only tense mistakes. In the above studies, it was assumed that the type of mistake that needs to be corrected is known, and irrelevant verb errors were excluded (e.g., Tajiri et al. (2012) addressed only tense mistakes and excluded from the evaluation other kinds of verb errors). In other words, it was assumed that part of the task was solved. But, unlike in article and preposition error correction where the type of mistake is known based on the surface form of the word, in verb error correction, it is not obvious.

The key distinction of our work is that we propose a holistic approach that starts from “scratch” and, given an instance, first detects a mistake and identifies its type, and then proceeds to correct it. We also evaluate several methods for selecting verb candidates and show the significance of this step for improving verb error correction performance, while earlier studies do not discuss this aspect of the problem. In the CoNLL shared task (Ng et al., 2013) that included verb errors in agreement and form, the participating teams did not provide details on how specific challenges were handled, but the University of Illinois system obtained the highest score on the verb sub-task, even though

\(^1\)The annotation is available at http://cogcomp.cs.illinois.edu/page/publication/view/743
Table 1: Grammatical verb errors in FCE.

<table>
<thead>
<tr>
<th>Tag</th>
<th>Error type</th>
<th>Rel. freq. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TV</td>
<td>Tense</td>
<td>40.0</td>
</tr>
<tr>
<td>FV</td>
<td>Form</td>
<td>22.3</td>
</tr>
<tr>
<td>AGV</td>
<td>Verb-subject agreement</td>
<td>11.5</td>
</tr>
<tr>
<td>MV</td>
<td>Missing verb</td>
<td>11.7</td>
</tr>
<tr>
<td>UV</td>
<td>Unnecessary verb</td>
<td>7.3</td>
</tr>
<tr>
<td>IV</td>
<td>Inflection</td>
<td>5.4</td>
</tr>
<tr>
<td>DV</td>
<td>Derivation</td>
<td>1.8</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>6640</td>
</tr>
</tbody>
</table>

all teams used similar resources (Ng et al., 2013).

3 Verb Errors in ESL Writing

Verb-related errors are very prominent among non-native English speakers: grammatical misuse of verbs constitutes one of the most common errors in several learner corpora, including those previously used (Izumi et al., 2003; Lee and Seneff, 2008) and the one employed in this work. We study verb errors using the FCE corpus (Yannakoudakis et al., 2011). The corpus possesses several desirable characteristics: it is large (500,000 words), has been annotated by native English speakers, and contains data by learners of multiple first-language backgrounds. The FCE corpus contains 5056 determiner errors, 5347 preposition errors, and 6640 grammatical verb mistakes (Table 1).

3.1 Verb Finiteness

There are many grammatical categories for which English verbs can be marked. The linguistic notion of verb finiteness or verb type (Radford, 1988; Quirk et al., 1985) distinguishes between verbs that function on their own in a clause as main verbs (finite) and those that do not (non-finite). Grammatical properties associated with each group are mutually exclusive: tense and agreement markers, for example, do not apply to non-finite verbs; non-finite verbs are not marked for many grammatical functions but may appear in several forms.

The most common verb problems for ESL learners – Tense, Agreement, non-finite Form – involve verbs both in finite and non-finite roles. Table 2 illustrates contexts that license finite and non-finite verbs.

Our intuition is that, because properties associated with each verb type are mutually exclusive, verb finiteness should benefit verb error correction models: an observed verb error may be due to several grammatical phenomena, and knowing which phenomena are active depends on the function of the verb in the current context. Note that Agreement, Tense, and Form errors account for about 74% of all grammatical verb errors in Table 1 but the finiteness distinction applies to all English verbs – every verb is either finite or non-finite in a specific syntactic context – and is also relevant for the remaining mistakes not addressed here.2

4 Annotation for Verb Finiteness

In order to evaluate the quality of the algorithm for verb finiteness and of the candidate selection methods, we annotated all verbs – correct and erroneous – in a random set of 124 documents from our corpus with the information about verb finiteness. We refer to these 124 documents as gold subset. We also annotated erroneous verbs in the remaining 1120 documents of the corpus. The annotation was performed by two students with background in Linguistics. The inter-annotator agreement is shown in Table 3 and is high.

Annotating Verb Errors

For each verb error that was tagged as Tense (TV), Agreement (AGV), and Form (FV), the annotators marked verb finiteness. Additionally, the annotators also specified the type of error (Tense, Agreement, or Form) (Table 4), since the FCE tags do not always correspond to the three error types we study here. For example, the FV tag may mark errors on finite verbs. Overall, about 7% of verb errors have to do with phenomena different from the three verb properties considered in this work and thus are excluded from the present study.

Annotating Correct Verbs

Correct verbs were identified in text using an automated procedure that relies on part-of-speech information (Sec. 5.1). Valid candidates were specified for verb finiteness. The candidates that were identified incorrectly due to mistakes by the part-of-speech tagger were marked as invalid.

5 The Computational Model

The verb error correction problem is formulated as a classification task in the spirit of the learn-

Table 3: Inter-annotator agreement based on 250 verb errors and 250 correct verbs, randomly selected.

<table>
<thead>
<tr>
<th>Category</th>
<th>Agreement</th>
<th>Kappa</th>
<th>Random</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct verbs</td>
<td>0.97</td>
<td>0.95</td>
<td>0.51</td>
</tr>
<tr>
<td>Erroneous verbs</td>
<td>0.88</td>
<td>0.81</td>
<td>0.41</td>
</tr>
</tbody>
</table>

2For instance, the missing verb errors (MV, 11.7%) require an additional step to identify contexts for missing verbs, and then appropriate verb properties need to be determined based on verb finiteness.
ing paradigm commonly used for correcting other ESL errors (Sec. 2), with the exception that the verb model includes additional components. All of the components are listed below:

1. Candidate selection (5.1)
2. Verb finiteness prediction (5.2)
3. Feature generation (5.3)
4. Error identification (5.4)
5. Error correction (5.5)

After verb candidates are selected, verb finiteness is determined and features are generated for each candidate. The finiteness prediction is used in the error identification component. Given the output of the error identification stage, the corresponding classifiers for each error type are invoked to propose an appropriate correction.

We split the corpus documents into two equal parts – training and test. We chose a train-test split and not cross-validation, since the FCE data set is quite large to allow for such a split. The training data is also used to develop the components for candidate selection and verb finiteness prediction.

### 5.1 Candidate Selection

This stage selects the set of verb instances that are presented as input to the classifier. A verb instance refers to the verb, including its auxiliaries or the infinitive marker (e.g. “found”, “will find”, “to find”). Candidate selection is a crucial step for models that correct mistakes on open-class words because those errors that are missed at this stage have no chance of being detected. We implement four candidate selection methods. Method (1) extracts all verbs heading a verb phrase, as identified by a shallow parser (Punyakanok and Roth, 2001). Method (2) also includes words tagged with one of the verb tags: {VB, VBN, VBG, VBD, VBP, VBZ} predicted by the POS tagger. However, relying on the POS information is not good enough, since the POS tagger performance on ESL data is known to be suboptimal (Nagata et al., 2011). For example, verbs lacking agreement markers are likely to be mistagged as nouns (Lee and Seneff, 2008). Methods (3) and (4) address the problem of pre-processing errors. Method (3) adds words that are on the list of valid English verb lemmas; the lemma list is constructed using a POS-tagged version of the NYT section of the Gigaword corpus and contains about 2,600 of frequently-occurring words tagged as VB; for example, (3) will add shop but not shopping, but (4) will add both.

For methods (3) and (4), we developed verb-Morph, a tool that performs morphological analysis on verbs and is used to lemmatize verbs and to generate morphological variants. The module makes uses of (1) the verb lemma list and (2) a list of irregular English verbs.

The quality of the candidate selection methods is evaluated in Table 5 on the gold subset by computing the recall, i.e. the percentage of erroneous verbs that have been selected as candidates. Methods that address pre-processing mistakes are able to recover more erroneous verb candidates in text. It is also interesting to note that across all methods, the highest recall is obtained for tense errors. This suggests that the POS tagger is more prone to fail-

---

1. Examples of verb instances:
   - "He discussed this with me last week"
   - "He discusses this with me every week"
   - "They let him discuss this with me"
   - "To discuss this now would be ill-advised."

2. Table 2: Contexts that license finite and non-finite verbs and the corresponding active properties.

<table>
<thead>
<tr>
<th>Error on Verb Type</th>
<th>Subcategory</th>
<th>Example</th>
</tr>
</thead>
</table>
| Finite (67.7%)     | Agreement (90%) | "We discuss this every time."
|                    | Tense (80%)     | "If you buy something, you would be happy."
| Non-finite (25.3%) | Passive/Active (42.3%) | "If one is famous he has to accept the disadvantages of being famous."
|                    | Compound (40.7%) | "They arrived early to organize everything."
| Other errors (7.0%) | Other (16.8%) | "Our end-of-conference party (is included) includes dinner and dancing."

3. Table 4: Verb error classification based on 4864 mistakes marked as TV, AGV, and FV errors in the FCE corpus.

<table>
<thead>
<tr>
<th>Error on Verb Type</th>
<th>Subcategory</th>
<th>Example</th>
</tr>
</thead>
</table>
| Finite (67.7%)     | Agreement (90%) | "We discuss this every time."
|                    | Tense (80%)     | "If you buy something, you would be happy."
| Non-finite (25.3%) | Passive/Active (42.3%) | "If one is famous he has to accept the disadvantages of being famous."
|                    | Compound (40.7%) | "They arrived early to organize everything."
| Other errors (7.0%) | Other (16.8%) | "Our end-of-conference party (is included) includes dinner and dancing."

---

http://cogcomp.cs.illinois.edu/page/software
http://cogcomp.cs.illinois.edu/demo/shallowparse
http://cogcomp.cs.illinois.edu/page/publication_view/POS
http://cogcomp.cs.illinois.edu/page/software_view/POS
Table 5: Candidate selection methods performance.

<table>
<thead>
<tr>
<th>Method</th>
<th>Recall (%)</th>
<th>Recall by error group (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Agr.</td>
</tr>
<tr>
<td>(1) All verb phrases</td>
<td>83.00</td>
<td>86.62</td>
</tr>
<tr>
<td>(2) + tokens tagged as verbs</td>
<td>91.96</td>
<td>90.30</td>
</tr>
<tr>
<td>(3) + tokens that are valid verb lemmas</td>
<td>95.50</td>
<td>95.99</td>
</tr>
<tr>
<td>(4) + tokens with inflections that are valid verb lemmas</td>
<td>96.09</td>
<td>96.32</td>
</tr>
</tbody>
</table>

5.4 Error Identification

The goal of this stage is to identify errors and to predict their type. We define a linear model where, given a verb, a weight vector \( w \) assigns a score to each label in the label space \{Correct, Form, Agreement, Tense\}. The prediction of the classifier is the label with the highest score.

The baseline error identification model, called combined, is agnostic to the type of the verb. In the combined model, for each verb \( v \) and label \( l \), we generate a feature vector, \( \phi(v, l) \) and the best label is predicted as

\[
\text{arg max}_l w^T \phi(v, l).
\]

The combined model makes use of all the features we have defined earlier for each verb.

The type-based model uses the verb finiteness prediction made by the verb finiteness classifier. A soft way to use the finiteness prediction is to add the predicted finiteness value as a feature. The other – hard-decision approach – is to use only a subset of the features depending on the predicted finiteness: Agreement and Tense for the finite verbs, and Form features for non-finite. The hard-decision type-driven approach defines a feature vector for a verb based on its type. Thus, given the verb \( v \) and its type \( t \), we define features \( \phi(v, t, l) \) for each label \( l \). Thus, the label is predicted as

\[
\text{arg max}_l w^T \phi(v, t, l).
\]

5.5 Error Correction

The correction module consists of three components, one for each type of mistake. Given the output of the error identification model, the appropriate correction component is run for each instance predicted to be a mistake.\(^6\) The verb finiteness prediction is used to select finite instances for training the Agreement and Tense components and non-finite – for the Form component. The label space for Tense specifies tense and aspect properties of the English verbs (see Tajiri et al., 2012 for more detail), the Agreement component specifies the person and number properties, while the Form component includes the commonly confusable non-finite English forms (see Table 2). These components are trained as multiclass classifiers.

\(^6\)We assume that each verb contains at most one mistake. Less than 1% of all erroneous verbs have more than one error present.
A verb is Non-Finite if any of the following hold:

- \( \text{numTokens} = 2 \) \& \( \text{firstToken} = \text{to} \)
- \( \text{firstToken} = \text{be} \)
- \( \text{pos} = \text{V BG} \)

A verb is Finite if any of the following hold:

- All verbs identified by shallow parser
- can, could
- \( \text{pos} \in \{ \text{V BD}, \text{V BP}, \text{V BZ} \} \)
- \( \text{numTokens} = 2 \) \& \( \text{firstToken} = \text{to} \)
- \( \text{numTokens} \geq 2 \)

<table>
<thead>
<tr>
<th>Agreement</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>subjHead, subjPOS</td>
<td>The surface form and the POS tag of the subject head</td>
</tr>
<tr>
<td>subjHead (those, these, him, her, his, her, it)</td>
<td>Determinant of the subject phrase</td>
</tr>
<tr>
<td>subjDistance</td>
<td>Distance between the verb and the subject head</td>
</tr>
<tr>
<td>subjNumber [Sing, Pl]</td>
<td>Sing – singular pronouns and nouns; Pl – plural pronouns and nouns</td>
</tr>
<tr>
<td>subjPerson {3rdSing, Not3rdSing, 1stSing}</td>
<td>3rdSing – she, he, it, singular nouns; Not3rdSing – we, you, they, plural nouns; 1stSing – I</td>
</tr>
<tr>
<td>conjunctions</td>
<td>(1) &amp; (3)(4)(5)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tense</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>verb phrase (VP)</td>
<td>verb lemma, negation, surface forms and POS tags of all words in the verb phrase</td>
</tr>
<tr>
<td>verbs in sentence(3 features)</td>
<td>tenses and lemmas of the finite verbs preceding and following the verb instance</td>
</tr>
<tr>
<td>tense adverbs (2 features)</td>
<td>temporal adverb before and after the verb instance</td>
</tr>
<tr>
<td>bag-of-words (BOW) (8 features)</td>
<td>Includes the following words in the sentence: {it, when, since, then, when, how, when, since, after}</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Form</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>closest word</td>
<td>surface form, lemma, POS tag, and distance of the closest open-class word to the left of the verb</td>
</tr>
<tr>
<td>governor</td>
<td>surface form, POS tag and dependency type of the target</td>
</tr>
<tr>
<td>preposition</td>
<td>if the verb is preceded by a preposition: preposition itself and the surface form, POS tag and dependency of the governor of the preposition</td>
</tr>
<tr>
<td>pos and lemma</td>
<td>POS tag and lemma of the verb and their conjunctions with features in (2) and (3) and word ngrams</td>
</tr>
</tbody>
</table>

Table 6: Algorithm for determining verb type. \( \text{numTokens} \) denotes the number of tokens in the verb instance, e.g., for the verb instance “to go”, \( \text{numTokens} = 2 \). Verbs not covered by the rules, e.g., those that are not tagged with a verb-related POS in methods (3) and (4), are not assigned any verb type. The last column shows algorithm accuracy on the gold subset separately for correct and incorrect verbs.

Table 7: Features used, grouped by error type.

6 Experiments

The main goal of this work is to propose a unified framework for correcting verb mistakes and to address the specific challenges of the problem. We thus do not focus on features or on the specific learning algorithm. Our experimental study addresses the following research questions:

I. Linguistic questions: (i) candidate selection methods; (ii) verb finiteness contribution to error identification

II. Computational Framework: error identification vs. correction

III. Gold annotation: (i) using gold candidates and verb type vs. automatic; (ii) performance comparison by error type

Learning Framework There is a lot of understanding for which algorithmic methods work best for ESL correction tasks, how they compare among themselves, and how they compare to n-gram based methods. Specifically, despite their intuitive appeal, language models were shown to not work well on these tasks, while the discriminative learning framework has been shown to be superior to other approaches and thus is commonly used for error correction tasks (see Sec. 2). Since we do not address the algorithmic aspect of the problem, we refer the reader to Rozovskaya and Roth (2011) for a discussion of these issues. We train all our models with the SVM learning algorithm implemented in JLIS (M. Chang and Roth, 2010).

Evaluation We report both Precision/Recall curves and AAUC (as a summary). Error correction is generally evaluated using F1 (Dale et al., 2012); Precision and Recall (Gamon, 2010; Tajiri et al., 2012); or Average Area Under Curve (AAUC) (Rozovskaya and Roth, 2011). For a discussion on these metrics with respect to error correction tasks, we refer the reader to Rozovskaya (2013). AAUC (Hanley and McNeil, 1983)) is a measure commonly used to generate a summary statistic, computed as an average precision value over a range of recall points. In this paper, AAUC is computed over the first 15 recall points:

\[
\text{AAUC} = \frac{1}{15} \cdot \sum_{i=1}^{15} \text{Precision}(i).
\]

6.1 Linguistic Questions

Candidate Selection Methods The contribution of the candidate selection component with respect to error identification is evaluated in Table 8, using the methods presented in Sec. 5.1. Overall,
**Table 8:** Impact of candidate selection methods on error identification performance. The first column shows the percentage of erroneous verbs selected by each method. Type-based models are discussed in Sec. 6.1.

<table>
<thead>
<tr>
<th>Recall of candidate selection method (%)</th>
<th>Combined</th>
<th>Type-based</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) (83.00)</td>
<td>73.38</td>
<td>79.49</td>
</tr>
<tr>
<td>(2) (91.96)</td>
<td>80.36</td>
<td>86.48</td>
</tr>
<tr>
<td>(3) (95.50)</td>
<td><strong>81.39</strong></td>
<td><strong>87.05</strong></td>
</tr>
<tr>
<td>(4) (96.09)</td>
<td>81.27</td>
<td>86.81</td>
</tr>
</tbody>
</table>

**Table 9:** Training and test data statistics. Candidates are selected using method (3).

<table>
<thead>
<tr>
<th></th>
<th>Correct verbs</th>
<th>Erroneous verbs</th>
<th>Error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>41721</td>
<td>1981</td>
<td>4.75%</td>
</tr>
<tr>
<td>Test</td>
<td>41836</td>
<td>2014</td>
<td>4.81%</td>
</tr>
</tbody>
</table>

Table 9 shows the number of verb instances in training and test selected with this method.

**Verb Finiteness** Sec. 5.4 presented two ways of adding verb finiteness: (1) adding the predicted verb type as a feature and (2) selecting only the relevant features depending on the finiteness of the verb. Table 10 shows the results of using verb type in the error identification stage. While the first approach does not provide improvement over the combined model, the second method is very effective. We conjecture that because verb type prediction is quite accurate, the second, hard-decision approach is preferred, as it provides knowledge in a direct way. Henceforth, we will use the second method in the type-based model.

Fig. 1 compares the performance of the combined and the hard-decision type-based models shown in Table 10. Precision/Recall curves are generated by varying the threshold on the confidence of the classifier. This graph reveals the behavior of the systems at multiple recall points: we observe that at every recall point the type-based classifier has higher precision.

**Table 10:** Verb finiteness contribution to error identification.

<table>
<thead>
<tr>
<th>Feature set</th>
<th>AAUC Combined</th>
<th>Type-based</th>
<th>AAUC Combined</th>
<th>Type-based</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>46.62</td>
<td>49.72</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All−Syntax</td>
<td>79.47</td>
<td>84.88</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full feature set</td>
<td><strong>81.39</strong></td>
<td><strong>87.05</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 11: Verb finiteness contribution to error identification for different features.

6.2 Identification vs. Correction

After running the error identification component, we apply the appropriate correction models to those instances identified as errors. The results for identification and correction are shown in Table 12. The correction models are also finiteness-aware models trained on the relevant verb instances (finite or non-finite), as predicted by the verb finiteness classifier.

We evaluate the correction components by fixing a recall point in the error identification stage. We observe the relatively low recall obtained by the models. Error correction models tend to have low recall (see, for example, the recent shared tasks on ESL error correction (Dale and Kilgariff, 2011; Dale et al., 2012; Ng et al., 2013)). The key reason for the low recall is the error sparsity: over 95% of verbs are correct, as shown in Table 9.

---

We can increase recall using a different threshold but higher precision is preferred in error correction tasks.
Table 12: Performance of the complete model after the correction stage. The results on Agreement mistakes are the same, since Agreement errors are always binary decisions, unlike Tense and Form mistakes.

The only way to improve over this 95% baseline is by forcing the system to have very good precision (at the expense of recall). The performance shown in Table 12 corresponds to an accuracy of 95.60% in identification (error reduction of 8.7%) and 95.40% in correction (error reduction of 4.5%) over the baseline of 95.19%.

6.3 Analysis on Gold Data

To further study the impact of each step of the system, we analyze our model on the gold subset of the data. The gold subset contains two additional pieces of information not available for the rest of the corpus: gold verb candidates and gold verb finiteness (Sec. 4). The set contains 7784 gold verbs, including 464 errors. Experiments are run in 10-fold cross-validation where on each run 90% of the documents are used for training and the remaining 10% are used for evaluation. The gold annotation can be used instead of automatic predictions in two system components: (1) candidate selection and (2) verb finiteness.

Table 13 shows the performance on error identification when gold vs. automatic settings are used. As expected, using the gold verb type is more effective than using the automatic one, both with automatic and gold candidates. The same is true for candidate selection. For instance, the combined model improves by 14 AAUC points (from 55.90 to 69.86) with gold candidates. These results indicate that candidate selection is an important component of the verb error correction system.

Note that compared to the performance on the entire data set (Table 10), the performance of the models shown here that use automatic components is lower, since the training size is smaller. On the other hand, because of the smaller training size, the gain due to the type-based approach is larger on the gold subset (19 vs. 6 AAUC points).

Finally, in Table 14, we evaluate the contribution of verb finiteness to error identification by error type. While performance varies by error, it is clear that all errors benefit from verb typing.

7 Conclusion

Verb errors are commonly made by ESL writers but difficult to address due to their diversity and the fact that identifying verbs in (noisy) text may itself be difficult. We develop a linguistically-inspired approach that first identifies verb candidates in noisy learner text and then makes use of verb finiteness to identify errors and characterize the type of mistake. This is important, since most errors made by non-native speakers cannot be identified by considering only closed classes (e.g., prepositions and articles). Our model integrates a statistical machine learning approach with a rule-based system that encodes linguistic knowledge to yield the first general correction approach to verb errors (that is, one that does not assume prior knowledge of which mistake was made). This work thus provides a first step in considering more general algorithmic paradigms for correcting grammatical errors and paves the way for developing models to address other “open-class” mistakes.

Acknowledgments

The authors thank Graeme Hirst, Julia Hockenmaier, Mark Sammons, and the anonymous reviewers for their helpful feedback. This work was done while the first and the third authors were at the University of Illinois. This material is based on research sponsored by DARPA under agreement number FA8750-13-2-0008 and by the Army Research Laboratory (ARL) under agreement W911NF-09-2-0053. Any opinions, findings, conclusions or recommendations are those of the authors and do not necessarily reflect the view of the agencies.
References


