Joint Learning and Inference for Grammatical Error Correction

Alla Rozovskaya and Dan Roth
Cognitive Computation Group
University of Illinois at Urbana-Champaign
201 N. Goodwin Avenue
Urbana, IL 61801
{rozovska,danr}@illinois.edu

Abstract

State-of-the-art systems for grammatical error correction are based on a collection of independently-trained models for specific errors. Such models ignore linguistic interactions at the sentence level and thus do poorly on mistakes that involve grammatical dependencies among several words. In this paper, we identify linguistic structures with interacting grammatical properties and propose to address such dependencies via joint inference and joint learning.

We show that it is possible to identify interactions well enough to facilitate a joint approach and, consequently, that joint methods correct incoherent predictions that independently-trained classifiers tend to produce. Furthermore, because the joint learning model considers interacting phenomena during training, it is able to identify mistakes that require making multiple changes simultaneously and that standard approaches miss. Overall, our model significantly outperforms the Illinois system that placed first in the CoNLL-2013 shared task on grammatical error correction.

1 Introduction

There has recently been a lot of work addressing errors made by English as a Second Language (ESL) learners. In the past two years, three competitions devoted to grammatical error correction for non-native writers took place: HOO-2011 (Dale and Kilgarriff, 2011), HOO-2012 (Dale et al., 2012), and the CoNLL-2013 shared task (Ng et al., 2013).

Nowadays **phone/phones* **has/have* many functionalities, **included/including* **∅/a* camera and **∅/a Wi-Fi receiver.

Figure 1: Examples of representative ESL errors.

Most of the work in the area of ESL error correction has addressed the task by building statistical models that specialize in correcting a specific type of a mistake. Figure 1 illustrates several types of errors common among non-native speakers of English: article, subject-verb agreement, noun number, and verb form. A significant proportion of research has focused on correcting mistakes in article and preposition usage (Izumi et al., 2003; Han et al., 2006; Felice and Pulman, 2008; Gamon et al., 2008; Tetreault and Chodorow, 2008; Gamon, 2010; Rozovskaya and Roth, 2010b). Several studies also consider verb-related and noun-related errors (Lee and Seneff, 2008; Gamon et al., 2008; Dahlmeier and Ng, 2012). The predictions made by individual models are then applied independently (Rozovskaya et al., 2011) or pipelined (Dahlmeier and Ng, 2012).

The standard approach of training individual classifiers considers each word independently and thus assumes that there are no interactions between errors and between grammatical phenomena. But an ESL writer may make multiple mistakes in a single sentence and these result in misleading local cues given to individual classifiers. In the example shown in Figure 1, the agreement error on the verb "have" interacts with the noun number error: a correction system that takes into account the context may infer, because of the word "phone", that the verb number is correct. For this reason, a system that consid-

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 791–802,
Seattle, Washington, USA, 18-21 October 2013.

ers noun and agreement errors separately will fail to identify and correct the interacting errors shown in Fig. 1. Furthermore, it may also produce inconsistent predictions.

Even though it is quite clear that grammatical errors interact, for various conceptual and technical reasons, this issue has not been addressed in a significant way in the literature. We believe that the reasons for that are three-fold: (1) Data: until very recently we did not have data that jointly annotates sufficiently many errors of interacting phenomena (see Sec. 2). (2) Conceptual: Correcting errors in interacting linguistic phenomena requires that one identifies those phenomena and, more importantly, can recognize reliably the interacting components (e.g., given a verb, identify the subject to enable enforcing agreement). The perception has been that this cannot be done reliably (Sec. 4). (3) Technical: The NLP community has started to better understand joint learning and inference and apply it to various phenomena (Roth and Yih, 2004; Punyakanok et al., 2008; Martins et al., 2011; Clarke and Lapata, 2007; Sutton and McCallum, 2007) (Sec. 5).

In this paper we present, for the first time, a successful approach to jointly resolving grammatical errors. Specifically:

• We identify two pairs of interacting phenomena, *subject-verb* and *article-NPhead* agreements; we show how to reliably identify these pairs in noisy ESL data, thereby facilitating the joint correction of these phenomena.

• We propose two joint approaches: (1) a *joint inference* approach implemented on top of individually learned models using an integer linear programming formulation (ILP, (Roth and Yih, 2004)), and (2) a model that *jointly learns* each pair of these phenomena. We show that each of these methods has its advantages, and that both solve the two challenges outlined above: the joint models exclude inconsistent predictions that violate linguistic constraints. The joint learning model exhibits superior performance, as it is also able to overcome the problem of the noisy context encountered by the individual models and to identify errors in contexts, where multiple changes need to be applied at the same time.

We show that our joint models produce state-of-the-art performance and, in particular, significantly outperform the University of Illinois system that placed first in the CoNLL-2013 shared task, increasing the F1 score by 2 and 4 points in different evaluation settings.

2 Task Description and Motivation

To illustrate the utility of jointly addressing interacting grammatical phenomena, we consider the corpus of the CoNLL-2013 shared task on grammatical error correction (Ng et al., 2013), which we found to be particularly well-suited for addressing interactions between grammatical phenomena. The task focuses on the following five common mistakes made by ESL writers: *article*, *preposition*, *noun number*, *subject-verb agreement*, and *verb form*, and we address two interactions: *article-NPhead* and *subject-verb*.

The training data for the task is from the NUCLE corpus (Dahlmeier et al., 2013), an error-tagged collection of essays written by non-native learners of English. The test data is an additional set of essays by learners from the same linguistic background. The training and the test data contain 1.2M and 29K words, respectively. Although the corpus contains errors of other types, the task focuses on five types of errors. Table 1 shows the number of mistakes[1] of each type and the error rates, i.e. the percentage of erroneous words by error type.

Error	**Number of errors and error rate**	
	Training	**Test**
Article	6658 (2.4%)	690 (10.0%)
Prep.	2404 (2.0%)	311 (10.7%)
Noun	3779 (1.6%)	396 (6.0%)
Verb Agr.	1527(2.0%)	124 (5.2%)
Verb Form	1453 (0.8%)	122 (2.5%)

Table 1: **Number of annotated errors in the CoNLL-2013 shared task.** Percentage denotes the error rates, i.e. the number of erroneous instances with respect to the total number of relevant instances in the data. For example, 10.7% of prepositions in the test data are used incorrectly. The numbers in the *revised* data set are slightly higher.

We note that the CoNLL-2013 data set is the first annotated collection that makes a study like ours feasible. The presence of a common test set that

[1]System performance in the shared task is evaluated on data with and without additional revisions added based on the input from participants. The number of mistakes in the revised test data is slightly higher.

contains a good number of interacting errors – article, noun, and verb agreement mistakes – makes the data set well-suited for studying which approach works best for addressing interacting phenomena. The HOO-2011 shared task collection (Dale and Kilgarriff, 2011) contains a very small number of noun and agreement errors (41 and 11 in test, respectively), while the HOO-2012 competition (Dale et al., 2012) only addresses article and preposition mistakes. Indeed, in parallel to the work presented here, Wu and Ng (2013) attempted the ILP-based approach of Roth and Yih (2004) in this domain. They were not able to show any improvement, for two reasons. First, the HOO-2011 data set which they used does not contain a good number of errors in interacting structures. Second, and most importantly, they applied constraints in an indiscriminate manner. In contrast, we show how to identify the interacting structures' components in a reliable way, and this plays a key role in the joint modeling improvements.

Lack of data hindered other earlier efforts for error correction beyond individual language phenomena. Brockett et al. (2006) applied machine-translation techniques to correct noun number errors on mass nouns and article usage but their application was restricted to a small set of constructions. Park and Levy (2011) proposed a language-modeling approach to whole sentence error correction but their model is not competitive with individually trained models. Finally, Dahlmeier and Ng (2012) proposed a decoder model, focusing on four types of errors in the data set of the HOO-2011 competition (Dale and Kilgarriff, 2011). The decoder optimized the sequence in which individual classifiers were to be applied to the sentence. However, because the decoder still corrected mistakes in a pipeline fashion, one at a time, it is unlikely that it could deal with cases that require simultaneous changes.

3 The University of Illinois System

Below, we briefly describe the University of Illinois system (henceforth *Illinois*; in the overview paper of the shared task the system is referred to as UI) that achieved the best result in the CoNLL-2013 shared task and which we use as our baseline model. For a complete description, we refer the reader to Rozovskaya et al. (2013).

The Illinois system implements five machine-learning independently-trained classifiers that follow the popular approach to ESL error correction borrowed from the context-sensitive spelling correction task (Golding and Roth, 1999; Carlson et al., 2001). A *confusion set* is defined that specifies a list of confusable words. Each occurrence of a confusable word in text is represented as a vector of features derived from a context window around the target. The problem is cast as a multi-class classification task and a classifier is trained on native or learner data. At prediction time, the model selects the most likely candidate from the confusion set.

The confusion set for prepositions includes the top 12 most frequent English prepositions. The article confusion set is as follows: {a,the,∅}[2]. The confusion sets for *noun*, *agreement*, and *form* modules depend on the target word and include its morphological variants (Table 2).

"Hence, the environmental ***factor/factors** also ***contributes/contribute** to various difficulties, ***included/including** problems in nuclear technology."	
Error type	**Confusion set**
Noun	{factor, factors}
Verb Agr.	{contribute, contributes}
Verb Form	{included, including, includes, include}

Table 2: **Confusion sets for noun number, agreement, and form classifiers.**

The *article* classifier is a discriminative model that draws on the state-of-the-art approach described in Rozovskaya et al. (2012). The model makes use of the Averaged Perceptron algorithm (Freund and Schapire, 1996) and is trained on the training data of the shared task with rich features.

The other models are trained on native English data, the Google Web 1T 5-gram corpus (henceforth, Google, (Brants and Franz, 2006)) with the Naïve Bayes (NB) algorithm. All models use word n-gram features derived from the 4-word window around the target word. In the *preposition* model, priors for preposition preferences are learned from the shared task training data (Rozovskaya and Roth, 2011).

[2] ∅ denotes noun-phrase-initial contexts where an article is likely to have been omitted. The variants "a" and "an" are conflated and are restored later.

Example	Predictions made by the Illinois system
"They believe that *such situation* must be avoided."	such situation → such a situations
"Nevertheless , electric *cars is* still regarded as a great trial innovation."	cars is → car are
"Every *students have* appointments with the head of the department."	No change

Table 3: **Examples of predictions of the Illinois system that combines independently-trained models.**

The words that are selected as input to classifiers are called *candidates*. Article and preposition candidates are identified with a closed list of words; noun-phrase-initial contexts for the article classifier are determined using a shallow parser[3] (Punyakanok and Roth, 2001). Candidates for the noun, agreement, and form classifiers are identified with a part-of-speech tagger[4], e.g. *noun* candidates are words that are tagged as NN or NNS. Table 4 shows the total number of candidates for each classifier.

	Classifier				
	Art.	**P**	**N**	**Agr.**	**F**
Train	254K	103K	240K	75K	175K
Test	6K	2.5K	2.6K	2.4K	4.8K

Table 4: **Number of candidate words by classifier type in training and test data.**

4 Interacting Mistakes

The approach of addressing each type of mistake individually is problematic when multiple phenomena interact. Consider the examples in Table 3 and the predictions made by the Illinois system. In the first and second sentences, there are two possible ways to correct the structures "such situation" and "cars is". In the former, either the article or the noun number should be changed; in the latter, either the noun number or the verb agreement marker[5]. In these examples, each of the independently-trained classifiers identifies the problem because each system makes a decision using the second error as part of its contextual cues, and thus the individual systems produce inconsistent predictions.

The second type of interaction concerns cases that require correcting more than one word at a time: the last example in Table 3 requires making changes both to the verb and the subject. Since each of the independent classifiers (for nouns and for verb agreement) takes into account the other word as part of its features, they both infer that the verb number is correct and that the grammatical subject "student" should be plural.

We refer to the words whose grammatical properties interact as *structures*. The independently-trained classifiers tend to fail to provide valid corrections in contexts where it is important to consider both words of the structure.

4.1 Structures for Joint Modeling

We address two linguistic structures that are relevant for the grammatical phenomena considered: *article-NPhead* and *subject-verb*. In the *article-NPhead* structures, the interaction is between the head of the noun phrase (NP) and the article that refers to the NP (first example in Table 3). In particular, the model should take into account that the article "a" cannot be combined with a noun in plural form. For subject-verb agreement, the subject and the verb should agree in number.

We now need to identify all pairs of candidates that form the relevant structures. *Article-NPhead* structures are pairs of words, such that the first word is a candidate of type article, while the second word is a noun candidate. Given an article candidate, the head of its NP is determined using the POS information (this information is obtained from the article feature vector because the NP head is a feature used by the article system)[6]. *Subject-verb* structures are pairs of noun-agreement candidates. Given a verb, its subject is identified with a dependency parser (Marneffe et al., 2006).

To evaluate the accuracy of subject and NP head

[3] http://cogcomp.cs.illinois.edu/page/software_view/Chunker

[4] http://cogcomp.cs.illinois.edu/page/software_view/POS

[5] Both of these solutions will result in grammatical output and the specific choice between the two depends on the wider essay context.

[6] Some heads are not identified or belong to a different part of speech.

predictions, a random sample of 500 structures of each type from the training data was examined by a human annotator with formal training in Linguistics. The human annotations were then compared against the automatic predictions. The results of the evaluation for subject-verb and article-NPhead structures are shown in Tables 5 and 6, respectively. Although the overall accuracy is above 90% for both structures, the accuracy varies by the distance between the structure components and drops significantly as the distance increases. For article-NPhead structures, *distance* indicates the position of the NP head with respect to the article, e.g. distance of 1 means that the head immediately follows the article. For subject-verb structures, *distance* is shown with respect to the verb: a distance of -1 means that the subject immediately precedes the verb. Although in most cases the subject is located to the left of the verb, in some constructions, such as existential clauses and inversions, it occurs after the verb.

Based on the accuracy results for identifying the structure components, we select those structures where the components are reliably identified. For article-NPhead, valid structures are those where the distance is at most three words. For subject-verb, we consider as valid those structures where the identified subject is located within two words to the left or three words to the right of the verb.

The valid structures are selected as input to the joint model (Sec. 5). The *joint learning* model considers only those valid structures whose components are *adjacent*. In adjacent structures the NP head immediately follows the article, and the verb immediately follows the subject. *Joint inference* is not restricted to adjacent structures.

The last column of Table 5 shows that valid subject-verb structures account for 67.5% of all verbs whose subjects are common nouns (51.7% are cases where the words are adjacent). Verbs whose subjects are common nouns account for 57.8% of all verbs that have subjects (verbs with different types of subjects, most of which are personal pronouns, are not considered here, since these subjects are not part of the noun classifier).

Valid article-NPhead structures account for 98.0% of all articles whose NP heads are common nouns (47.5% of those are adjacent structures), as shown in the last column of Table 6. 71.0% of articles in the training data belong to an NP whose head is a common noun; NPs whose heads belong to different parts of speech are not considered.

Note also that because a noun may belong both to an article-NPhead and a subject-verb structure, the structures contain an overlap.

Distance	**Accuracy**	**% of all subj. predictions**	**Cumul.**
-1	97.6%	51.7%	51.7%
1,2,3	100.0%	8.9%	60.6%
-2	88.2%	6.9%	67.5%
Other	80.8%	32.5%	100.0%

Table 5: **Accuracy of subject identification on a random sample of subject-verb structures from the training data.** The overall accuracy is 91.52%. For each distance, the following are shown: accuracy based on comparison with human evaluation; the percentage of all predictions that have this distance; the cumulative percentage.

Distance	**Accuracy**	**% of all head predictions**	**Cumul.**
1	94.8%	47.5%	47.5%
2	94.4%	44.0%	91.5%
3	92.3%	6.5%	98.0%
Other	89.1%	2.0%	100%

Table 6: **Accuracy of NP head identification on a random sample of article-NPhead structures from training data.** The overall accuracy is 94.45%. For each distance, the following are shown: accuracy based on comparison with human evaluation; the percentage of all predictions that have this distance; the cumulative percentage.

5 The Joint Model

In this section, we present the *joint inference* and the *joint learning* approaches. In the joint inference approach, we use the independently-learned models from the Illinois system, and the interacting target words identified earlier are considered only at inference stage. In the joint learning method, we jointly learn a model for the interacting phenomena.

The label space in the joint models corresponds to sequences of labels from the confusion sets of the individual classifiers: $\{a-sing, a-pl, the-sing, the-pl, \varnothing-sing, \varnothing-pl\}$ and $\{sing-sing, sing{-}pl, pl{-}sing, pl{-}pl\}$ for article-NPhead and subject-verb structures, respectively[7]. Invalid

[7]"sing" and "pl" refer to the grammatical number of noun

structures, such as *pl-sing* are excluded via hard constraints (when we run joint inference) or via implicit soft constraints (when we use joint learning).

5.1 Joint Inference

In the individual model approach, decisions are made for each word independently, ignoring the interactions among linguistic phenomena. The purpose of joint inference is to include linguistic (i.e. structural) knowledge, such as "plural nouns do not take an indefinite article", and "agreement consistency between the verb and the subject that controls it". This knowledge should be useful for resolving inconsistencies produced by individual classifiers.

The inference approach we develop in this paper follows the one proposed by Roth and Yih (2004) of training individual models and combining them at decision time via joint inference. The advantage of this method is that it allows us to build upon any existing independently-learned models that provide a distribution over their outcome, and produce a coherent global output that respects our declarative constraints. We formulate our component inference problems as integer linear program (ILP) instances as in Roth and Yih (2004).

The inference takes as input the individual classifiers' confidence scores for each prediction, along with a list of constraints. The output is the optimal solution that maximizes the linear sum of the confidence scores, subject to the constraints that encode the interactions. The joint model thus selects a hypothesis that both obtains the best score according to the individual models and satisfies the constraints that reflect the interactions among the grammatical phenomena at the level of linguistic structures, as defined in Sec. 4.

Inference The joint inference is enforced at the level of structures, and each structure corresponds to one ILP instance. All structures consist of two or three words: when an article-NPhead structure and a subject-verb structure include the same noun, the structure input to the ILP consists of an article-noun-verb triple. We formulate the inference problem as follows: Given a structure s that consists of n words, let w_i correspond to the i^{th} word in the structure. Let h denote a hypothesis from the hypothesis space H for s, and $score(w_i, h, l^i)$ denote the score assigned by the appropriate error-specific model to w_i under h for label l from the confusion set of word w_i. We denote by $e_{w,l}$ the Boolean variable that indicates whether the prediction on word w is assigned the value l ($e_{w,l} = 1$) or not ($e_{w,l} = 0$).

We assume that each independent classifier returns a score that corresponds to the likelihood of word w_i under h being labeled l^i. The softmax function (Bishop, 1995) is used to convert raw activation scores to conditional probabilities for the discriminative article model. The NB scores are also normalized and correspond to probabilities. Then the inference task is solved by maximizing the overall score of a candidate assignment of labels l to words w (this set of feasible assignments is denoted H here) subject to the constraints C for the structure s:

$$\hat{h} = \arg\max_{h \in H} score(h) =$$

$$= \arg\max_{h \in H} \sum_{i=1}^{n} score(w_i, h, l^i) e_{w_i, l^i}$$

subject to $C(s)$

Constraints In the $\{0, 1\}$ linear programming formulation described above, we can encode linguistic constraints that reflect the interactions among the linguistic phenomena. The inference enforces the following structural and linguistic constraints:

1. The indefinite article "a" cannot refer to an NP headed by a plural noun.
2. Subject and verb must agree in number.

In addition, we encode "legitimacy" constraints, that make sure that each w is assigned a single label. All constraints are encoded as hard constraints.

5.2 Joint Learning

We now describe how we *learn* the subject-verb and article-NPhead structures jointly. The joint model is implemented as a NB classifier and is trained in the same way as the independent models on the Google corpus with word n-gram features. Unlike the independent models, where the target corresponds to one

and verb agreement candidates. The candidates themselves are the surface forms of specific words that realize these grammatical properties. Note that a subject in subject-verb structures is always third person, since all subjects in subject-verb structures are common nouns; other subjects, including pronouns, are excluded. Thus the agreement distinction is singular vs. plural.

System	Adjacent structures		All distances	
	F1 (Orig.)	F1 (Revised)	F1 (Orig.)	F1 (Revised)
Illinois	31.20	42.14	31.20	42.14
NaïveVerb	31.19	42.20	31.13	42.16
NaïveNoun	31.03	41.87	30.91	41.70
This paper joint systems	**Joint Inference** (adjacent)		**Joint Inference** (all distances)	
	F1 (Orig.)	F1 (Revised)	F1 (Orig.)	F1 (Revised)
Subject-verb	31.90	42.94	31.97	42.86
Article-NPhead	31.63	42.48	31.79	42.59
Subject-verb + article-NPhead	**32.35**	**43.16**	**32.51**	**43.19**

Table 7: **Joint Inference Results.** All results are on the CoNLL-2013 test data using the original and revised gold annotations. *Adjacent* denotes a setting, where the joint inference is applied to structures with consecutive components (article-NPhead or subject-verb). *All distances* denotes a setting, where the constraints are applied to all valid structures, as described in Sec. 4.1. *Illinois* denotes the result obtained by the top CoNLL-2013 shared task system. In all cases, the candidates that are not part of the structures are handled by the respective components of the Illinois system. *NaïveVerb* and *NaïveNoun* denote heuristics, where a verb or subject are changed to ensure agreement. All improvements over the Illinois system are statistically significant (McNemar's test, $p < 0.01$).

word, here the target corresponds to two words that are part of the structure and the label space of the model is modified accordingly. Since we use features that can be computed from the small windows in the Google corpus, the joint learning model handles only adjacent structures (Sec. 4.1). Because the target consists of two words and the Google corpus contains counts for n-grams of length at most five, the features are collected in the three word window around the target.[8]

Unlike with the joint inference, here we do not explicitly encode linguistic constraints. One reason for this is that the NP head and subject predictions are not 100% accurate, so input structures will have noise. However, the joint model learns these constraints through the evidence seen in training.

6 Experiments

In this section, we describe our experimental setup and evaluate the performance of the joint approach. In the joint approach, the joint components presented in Sec. 5 handle the interacting structures described in Sec. 4. The individual classifiers of the Illinois system make predictions for the remaining words. The research question addressed by the experiments is the following: Given independently-trained systems for different types of errors, can we improve the performance by considering the phenomena that interact jointly? To address this, we report the results in the following settings:

1. *Joint Inference*: we compare the Illinois system that is a collection of individually-trained models that are applied independently with a model that uses joint inference encoded as declarative constraints in the ILP formulation and show that using joint inference results in a strong performance gain.
2. *Joint Learning*: we compare the Illinois system with a model that incorporates jointly-trained components for the two linguistic structures that we described in Sec. 4. We show that joint training produces an even stronger gain in performance compared to the Illinois model.
2. *Joint Learning and Inference*: we apply joint inference to the output of the joint learning system to account for dependencies not covered by the joint learning model.

We report F1 performance scored using the official scorer from the shared task (Dahlmeier and Ng, 2012). The task reports two types of evaluation: on the original gold data and on gold data with additional corrections. We refer to the results as *Original* and *Revised*.

6.1 Joint Inference Results

Table 7 shows the results of applying joint inference to the Illinois system. Both the article-NPhead and the subject-verb constraints improve the performance. The results for the joint inference are shown in two settings, adjacent and all structures, so that later we can compare joint inference with the joint learning model that handles only adjacent structures.

[8] Also note that when the article is ∅, the surface form of the structure corresponds to the NP head alone; this does not present a problem because in the NB model the context counts are normalized with the prior counts.

	Illinois system		Illinois-NBArticle	
	F1 (Orig.)	**F1 (Revised)**	**F1 (Orig.)**	**F1 (Revised)**
Illinois	31.20	42.14	31.71	41.38
This paper joint systems	**Joint Learning** (adjacent)		**Joint Learning** (adjacent)	
	F1 (Orig.)	**F1 (Revised)**	**F1 (Orig.)**	**F1 (Revised)**
Subject-verb	32.64*	43.37*	33.09*	**42.78***
Article-NPhead	33.89*	42.57*	33.16*	41.51
Subject-verb + article-NPhead	**35.12***	**43.73***	**34.41***	42.76*

Table 8: **Joint Learning Results.** All results are on the CoNLL-2013 test data using the original and revised gold annotations. *Illinois-NBArticle* denotes the Illinois system, where the discriminative article model is replaced with a NB classifier. *Adjacent* denotes a setting, where the structure components are consecutive (article-NPhead or subject-verb), as described in Sec. 4.1. *Illinois* denotes the result obtained by the top CoNLL-2013 shared task system. In all cases, the candidates that are not part of the structures are handled by the respective components of the Illinois system. Statistically significant improvements (McNemar's test, $p < 0.01$) over the Illinois system are marked with an asterisk (*).

It is also interesting to note that the key improvement comes from considering structures whose components are adjacent. This is not surprising given that the accuracy for subject and NP head identification drops as the distance increases.

For subject-verb constraints, we also implement a naïve approach that looks for contradictions and changes either the verb or the subject if they do not satisfy the number agreement. These two heuristics are denoted as *NaïveVerb* and *NaïveNoun*. The heuristics differ from the joint inference in that they enforce agreement by always changing either the noun (*NaïveNoun*) or the verb (*NaïveVerb*), while the joint inference does this using the scores produced by the independent models. In other words, the key is the objective function, while the components of the objective function are the same in the heuristics and the joint inference. The results in Table 7 show that simply enforcing agreement does not work well and that the ILP formulation is indeed effective and improves over the independently-trained models in all cases.

Recall that valid structures include only those whose components can be identified in a reliable way (Sec. 4.1). To evaluate the impact of that filtering, we perform two experiments with subject-verb structures (long-distance dependencies are more common in those constructions than in the article-NPhead structures): first, we apply joint inference to all subject-verb structures. We obtain F1 scores of 31.61 and 42.28, on original and revised gold data, respectively, which is significantly worse than the results on subject-verb structures in Table 7 (31.97 and 42.86, respectively) and only slightly better than the baseline performance of the Illinois system. Furthermore, when we apply joint inference to those structures which were excluded by filtering in Sec. 4.1, we find that the performance degrades compared to the Illinois system (30.85 and 41.58). These results demonstrate that the joint inference improvements are due to structures whose components can be identified with high accuracy and that it is essential to identify these structures; bad structures, on the other hand, hurt performance.

6.2 Joint Learning Results

Now we show experimental results of the joint learning (Table 8). Note that the joint learning component considers only those structures where the words are adjacent. Because the Illinois system presented in Sec. 3 makes use of a discriminative article model, while the joint model uses NB, we also show results, where the article model is replaced by a NB classifier trained on the Google corpus. In all cases, joint learning demonstrates a strong performance gain.

6.3 Joint Learning and Inference Results

Finally, we apply joint inference to the output of the joint learning system in Sec. 6.2. Table 9 shows the results of the Illinois model, the model that applies joint inference and joint learning separately, and both. Even though the joint learning performs better than the joint inference, the joint learning covers only adjacent structures. Furthermore, joint learning does not address overlapping structures of triples that consist of article, subject, and verb (6% of all structures). Joint inference allows us to ensure consistent predictions in cases not addressed by the

Example	Illinois system	JL and JI
"Moreover, the increased **technologies help** people to overcome different natural disasters.	No change	technology helps
"At that time,... **there are surveillances** in everyone's heart and criminals are more difficult to hide."	there are* surveillance*	there is surveillance
"In **such situation**, individuals will lose their basic privacy."	such a* situations*	such a situation
"In supermarket **monitor is** needed because we have to track thieves."	No change	monitors are

Table 10: **Examples of mistakes that are corrected by the joint model but not by the Illinois model.** *Illinois* denotes the result obtained by the top CoNLL-2013 shared task system from the University of Illinois. *JL* and *JI* stand for joint learning and joint inference, respectively. Inconsistent predictions are starred.

	F1 (Orig.)	F1 (Revised)
Illinois	31.20	42.14
Joint Inference	32.51	43.19
Joint Learning	35.12	43.73
Joint Learn. + Inf.	**35.21**	**43.74**

Table 9: **Joint Learning and Inference.** All results are on the CoNLL-2013 test data using the original and revised gold annotations. Results of the joint models that include the joint inference component are shown for structures of all distances. *Illinois* denotes the result obtained by the top CoNLL-2013 shared task system. All joint systems demonstrate a statistically significant improvement over the Illinois system; joint learning improvements are also statistically significant compared to the joint inference results (McNemar's test, $p < 0.01$).

joint learning model. Indeed, we can get a small improvement by adding joint inference on top of the joint learning on original annotations. Since the revised corrections are based on the participants' input and are most likely biased towards system predictions for corrections missed by the original annotators (Ng et al., 2013), it is more difficult to show improvement on revised data.

7 Discussion and Error Analysis

In the previous section, we evaluated the proposed joint inference and joint learning models that handle interacting grammatical phenomena. We showed that the joint models produce significant improvements over the highest-scoring CoNLL-2013 shared task system that consists of independently-trained classifiers: the joint approaches increase the F1 score by 4 F1 points on the original gold data and almost 2 points on the revised data (Table 9).

These results are interesting from the point of view of developing a practical error correction system. However, recall that the errors in the interacting structures are only a subset of mistakes in the CoNLL-2013 data set but the evaluation in Sec. 6 is performed with respect to all of these errors. From a scientific point of view, it is interesting to evaluate the impact of the joint models more precisely by considering the improvements on the relevant structures only. Table 11 shows how much the joint learning approach improves on the subset of relevant mistakes.

Structure	Performance (F1)	
	Illinois	Joint Learning
Subject-verb	39.64	**52.25**
Article-NPhead	30.65	**35.90**

Table 11: **Evaluation of the joint learning performance on the subset of the data containing interacting errors.** All results are on the CoNLL-2013 test data using the original annotations. *Illinois* denotes the result obtained by the top CoNLL-2013 shared task system. All improvements are statistically significant over the Illinois system (McNemar's test, $p < 0.01$).

Error Analysis To better understand where the joint models have an advantage over the independently-trained classifiers, we analyze the output produced by each of the approaches. In Table 10 we show examples of mistakes that the model that uses joint learning and inference is able to identify correctly, along with the original predictions made by the Illinois system.

Joint Inference vs. Joint Learning We wish to stress that the joint approaches do not simply perform better but also make coherent decisions by disallowing illegitimate outputs. The joint inference approach does this by enforcing linguistic constraints on the output. The joint learning model, while not explicitly encoding these constraints, learns them from the distribution of the training data.

Joint inference is a less expensive model, since it uses the scores produced by the individual classifiers and thus does not require additional training. Joint learning, on the other hand, is superior to joint inference, since it is better at modeling interactions where multiple errors occur simultaneously – it eliminates the noisy context present when learning the independent classifiers. Consider the first example from Table 10, where both the noun and the agreement classifiers receive noisy input: the verb "help" and the noun "technologies" act as part of input features for the noun and agreement classifiers, respectively. The noisy features prevent both modules from identifying the two errors.

Finally, an important distinction of the joint learning method is that it considers all possible output sequences *in training*, and thus it is able to better identify errors that require multiple changes, such as the last example in Table 10, where the Illinois system proposes no changes.

7.1 Error Correction: Challenges

We finalize our discussion with a few comments on the challenges of the error correction task.

Task Difficulty As shown in Table 1 in Sec. 2, only a small percentage of words have mistakes, while over 90% (about 98% in training) are used correctly. The low error rates are the key reason the error correction task is so difficult: it is quite challenging for a system to improve over a writer that already performs at the level of over 90%. Indeed, very few NLP tasks already have systems that perform at that level, even when the data is not as noisy as the ESL data.

Evaluation Metrics In the CoNLL-2013 competition, as well as the competitions alluded to earlier, systems were compared on F1 performance, and, consequently, this is the metric we optimize in this paper. Practical error correction systems, however, should be tuned to minimize recall to guarantee that the overall quality of the text does not go down. Indeed, the error sparsity makes it very challenging to identify mistakes accurately, and no system in the shared task achieves a precision over 50%. However, once the precision drops below 50%, the system introduces more mistakes than it identifies.

Clearly, optimizing the F1 measure does not ensure that the quality of the text improves as a result of running the system. Thus, it can be argued that the F1 measure is not the right measure for error correction. A different evaluation metric based on the *accuracy* of the data before and after running the system was proposed in Rozovskaya and Roth (2010c). When optimizing for this metric, the noun module, for instance, at recall point 20%, achieves a precision of 63.93%. This translates into accuracy of 94.46%, while the baseline on noun errors in the test data (i.e. the accuracy of the data before running the system) is 94.0% (Table 1). This means that the system improves the quality of the data.

Annotation Lastly, we believe that it is important to provide alternative corrections, as the agreement on what constitutes a mistake even among native English speakers can be quite low (Madnani et al., 2011).

8 Conclusion

This work presented the first successful study that jointly corrects grammatical mistakes. We addressed two pairs of interacting phenomena and showed that it is possible to reliably identify their components, thereby facilitating the joint approach.

We described two joint methods: a *joint inference* approach implemented via ILP and a *joint learning* model. The joint inference enforces constraints using the scores produced by the independently-trained models. The joint learning model learns the interacting phenomena as structures. The joint methods produce a significant improvement over a state-of-the-art system that combines independently-trained models and, importantly, produce linguistically legitimate output.

Acknowledgments

The authors thank Peter Chew, Jennifer Cole, Mark Sammons, and the anonymous reviewers for their helpful feedback. The authors thank Josh Gioja for the code that performs phonetic disambiguation of the indefinite article. This material is based on research sponsored by DARPA under agreement number FA8750-13-2-0008. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA or the U.S. Government.

References

C. Bishop. 1995. *Neural Networks for Pattern Recognition, chapter 6.4: Modelling conditional distributions*. Oxford University Press.

T. Brants and A. Franz. 2006. *Web 1T 5-gram Version 1*. Linguistic Data Consortium, Philadelphia, PA.

C. Brockett, D. B. William, and M. Gamon. 2006. Correcting ESL errors using phrasal SMT techniques. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 249–256, Sydney, Australia, July. Association for Computational Linguistics.

A. J. Carlson, J. Rosen, and D. Roth. 2001. Scaling up context sensitive text correction. In *IAAI*.

J. Clarke and M. Lapata. 2007. Modelling compression with discourse constraints. In *Proceedings of the 2007 Joint Conference of EMNLP-CoNLL*.

D. Dahlmeier and H.T Ng. 2012. A beam-search decoder for grammatical error correction. In *EMNLP-CoNLL*, Jeju Island, Korea, July. Association for Computational Linguistics.

D. Dahlmeier, H.T. Ng, and S.M. Wu. 2013. Building a large annotated corpus of learner english: The nus corpus of learner english. In *Proc. of the NAACL HLT 2013 Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, Atlanta, Georgia, June. Association for Computational Linguistics.

R. Dale and A. Kilgarriff. 2011. Helping Our Own: The HOO 2011 pilot shared task. In *Proceedings of the 13th European Workshop on Natural Language Generation*.

R. Dale, I. Anisimoff, and G. Narroway. 2012. A report on the preposition and determiner error correction shared task. In *Proc. of the NAACL HLT 2012 Seventh Workshop on Innovative Use of NLP for Building Educational Applications*, Montreal, Canada, June. Association for Computational Linguistics.

R. De Felice and S. Pulman. 2008. A classifier-based approach to preposition and determiner error correction in L2 English. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 169–176, Manchester, UK, August.

Y. Freund and R. E. Schapire. 1996. Experiments with a new boosting algorithm. In *Proc. 13th International Conference on Machine Learning*.

M. Gamon, J. Gao, C. Brockett, A. Klementiev, W. Dolan, D. Belenko, and L. Vanderwende. 2008. Using contextual speller techniques and language modeling for ESL error correction. In *Proceedings of IJCNLP*.

M. Gamon. 2010. Using mostly native data to correct errors in learners' writing. In *NAACL*, pages 163–171, Los Angeles, California, June.

A. R. Golding and D. Roth. 1999. A Winnow based approach to context-sensitive spelling correction. *Machine Learning*.

N. Han, M. Chodorow, and C. Leacock. 2006. Detecting errors in English article usage by non-native speakers. *Journal of Natural Language Engineering*, 12(2):115–129.

E. Izumi, K. Uchimoto, T. Saiga, T. Supnithi, and H. Isahara. 2003. Automatic error detection in the Japanese learners' English spoken data. In *The Companion Volume to the Proceedings of 41st Annual Meeting of the Association for Computational Linguistics*, pages 145–148, Sapporo, Japan, July.

J. Lee and S. Seneff. 2008. Correcting misuse of verb forms. In *ACL*, pages 174–182, Columbus, Ohio, June. Association for Computational Linguistics.

N. Madnani, M. Chodorow, J. Tetreault, and A. Rozovskaya. 2011. They can help: Using crowdsourcing to improve the evaluation of grammatical error detection systems. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 508–513, Portland, Oregon, USA, June. Association for Computational Linguistics.

M. Marneffe, B. MacCartney, and Ch. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *LREC*.

A. Martins, Noah N. Smith, M. Figueiredo, and P. Aguiar. 2011. Dual decomposition with many overlapping components. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 238–249, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

H. T. Ng, S. M. Wu, Y. Wu, Ch. Hadiwinoto, and J. Tetreault. 2013. The conll-2013 shared task on grammatical error correction. In *Proc. of the Seventeenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics.

A. Park and R. Levy. 2011. Automated whole sentence grammar correction using a noisy channel model. In *ACL*, Portland, Oregon, USA, June. Association for Computational Linguistics.

V. Punyakanok and D. Roth. 2001. The use of classifiers in sequential inference. In *NIPS*.

V. Punyakanok, D. Roth, and W. Yih. 2008. The importance of syntactic parsing and inference in semantic role labeling. *Computational Linguistics*, 34(2).

D. Roth and W. Yih. 2004. A linear programming formulation for global inference in natural language tasks. In Hwee Tou Ng and Ellen Riloff, editors, *CoNLL*.

A. Rozovskaya and D. Roth. 2010a. Annotating ESL errors: Challenges and rewards. In *Proceedings of the*

NAACL Workshop on Innovative Use of NLP for Building Educational Applications.

A. Rozovskaya and D. Roth. 2010b. Generating confusion sets for context-sensitive error correction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

A. Rozovskaya and D. Roth. 2010c. Training paradigms for correcting errors in grammar and usage. In *NAACL*.

A. Rozovskaya and D. Roth. 2011. Algorithm selection and model adaptation for ESL correction tasks. In *ACL*.

A. Rozovskaya, M. Sammons, J. Gioja, and D. Roth. 2011. University of Illinois system in HOO text correction shared task.

A. Rozovskaya, M. Sammons, and D. Roth. 2012. The UI system in the HOO 2012 shared task on error correction.

A. Rozovskaya, K.-W. Chang, M. Sammons, and D. Roth. 2013. The University of Illinois system in the CoNLL-2013 shared task. In *CoNLL Shared Task*.

C. Sutton and A. McCallum. 2007. Piecewise pseudolikelihood for efficient training of conditional random fields. In Zoubin Ghahramani, editor, *ICML*.

J. Tetreault and M. Chodorow. 2008. The ups and downs of preposition error detection in ESL writing. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 865–872, Manchester, UK, August.

Y. Wu and H.T. Ng. 2013. Grammatical error correction using integer linear programming. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1456–1465, Sofia, Bulgaria, August. Association for Computational Linguistics.