Modeling ESL Word Choice Similarities By Representing Word Intensions and Extensions

Huichao Xue,*Rebecca Hwa*
Department of Computer Science,
University of Pittsburgh,
210 S Bouquet St, Pittsburgh, PA 15260 USA
hux10@cs.pitt.edu, hwa@cs.pitt.edu

Abstract

Automatic error correction systems for English as a Second Language(ESL) speakers often rely on the use of a *confusion set* to limit the choices of possible correction candidates. Typically, the confusion sets are either manually constructed or extracted from a corpus of manually corrected ESL writings. Both options require the involvement of English teachers. This paper proposes a method to automatically construct confusion sets for commonly used prepositions from non-ESL corpus without manual intervention. The proposed method simulates how ESL learners learn both the intensions and extensions of English words from standard English text. Our experimental results suggest that the automatically constructed confusion sets based on the similarities between the learned words' intensions is competitive with those directly learned from an ESL corpus containing about 150K preposition usages.

Title and Abstract in another language, L_2 (optional, and on same page)

通过分析单词的内涵和外延来对用词混淆建模

针对把英语作为第二语言的人群的自动语法纠错系统，通常会需要使用“混淆集”来限制系统纠错的种类。一般来说，这些混淆集或者是由专家总结经验得出，或者是从被专家纠错过的文字当中提取的。这两种方法都需要英语专家的介入。在这篇论文当中，我们提出了一种无需专家介入，自动建立常用介词混淆集的方法。在此方法中，我们对英语单词的内涵和外延建模，并且模拟了英语学习者们学习单词内涵和外延的过程。实验表明，使用单词内涵之间的相似度来创建的混淆集，与从含15万介词的标注语料当中提取的混淆集质量是相当的。

Keywords: Confusion Sets, Lexical Semantics, Grammatical Error Correction, Distance Metric Learning, English as Second Language, Second Language Acquisition.

Keywords in L_2: 选词混淆、混淆集、词汇语义学、语法纠错、距离度量学习、英语作为第二语言、第二语言学习.

1 Introduction

A large portion of the English text (e.g., on the web) is written by people whose native language is not English. Many *English as a Second Language* (ESL) writers, even those with a high level of proficiency, make common grammatical mistakes. Researchers working on Grammar Error Correction (GEC) try to analyze the patterns of these mistakes in order to understand the underlying reasons for their occurrence and to build tools that help ESL writers to correct their errors (Leacock et al., 2010).

Many recently developed GEC systems (Chodorow et al., 2007; J. R. Tetreault & Chodorow, 2008; Gamon et al., 2009; Liu et al., 2010; Rozovskaya & Roth, 2011; Dahlmeier & Ng, 2011a) share a similar infrastructure: first, they isolate some specific types of errors (e.g., preposition errors, article errors, or word choice errors); then, they propose a correction for each instance by treating it as a classification problem. To cast the correction problem as a classification problem, the system has to know, a priori, what are the set of possible corrections for an error. That is, the system needs to pre-define a *confusion set* for each error type.

Previous work has shown the importance of the role of confusion sets. However, the construction of confusion sets requires a great deal of human involvement. English teachers are involved in Liu et al. (2010) to manually filter the initial large verb confusion sets; Rozovskaya & Roth (2010a) used annotated ESL corpus to limit their confusion sets for prepositions. They have shown that even for closed word classes such as prepositions, limiting the confusion sets help simplify the classifiers' tasks and finally lead to both a better precision and recall.

In this paper, we propose a method to automatically construct confusion sets without manual intervention or an annotated ESL corpus. Our approach is to model and simulate how ESL learners might learn words from reading English text. In the process of mastering the language, the learners are often confused about how to choose between similar words. Our goal in this work is to build a model that analyzes which words might appear similar to each other to an ESL learner and then builds up confusion sets with those words. The work presented in this paper addresses learning frequently used *prepositions*, but the idea may be generalized to open word classes.

Our simulation focuses on two main aspects of learning new words: learning their intensions and extensions. The intension of a word is often implied by its definition and its relations to other words; the extension of a word is often characterized by its usages[1]. Ultimately, ESL learners need to achieve a compatible understanding of both the word's intensions and extensions; but before that happens, they may confuse words that have either similar intensions or extensions. Our proposed model applies an algorithm called Relevance Component Analysis (Bar-Hillel et al., 2006) to describe how an ESL learner might organize the extensional representations of words onto an intensional space. We then build up confusion sets with words that have similar intensions.

We compare our model against two models that simulate how learners obtain words' intensions and extensions separately. Under the intensions-only model, word choice confusions are directly measured by the semantic similarity between words. Under the extensions-only model, word choice confusions are attributed to the learner not having completely mastered a word's usages; it can be seen as a faulty language model. In our experiments, we found that, by considering the

[1] We use the terms *intension* and *extension* following the definitions from from Linguistics literature(see, for example, Chalmers (2002)).

interaction of word intensions and extensions, our proposed model produces better confusion sets than those which consider them separately; moreover, the resulting confusion sets are competitive with those directly learned from an error-annotated ESL corpus containing 150K preposition usages.

2 Background

The mistakes made by ESL writers are not random. In their studies, Rozovskaya & Roth (2011) find that those who share the same native language tend to make similar types of mistakes. The natural question that arises is: what are the underlying causes for the mistakes? In the frame of computational linguistics research, the question might be rephrased as: Can we build a mathematical model that simulates ESL writing mistakes?

A model that builds a table of **confusion sets** whose distributions correlate well with the mistakes made by ESL writers is an important component in simulating ESL writing. For instance, Brockett et al. (2006) simulates an ESL corpus according to a set of manually constructed rules, which would not be available until confusion sets are established.

In addition to aiding our understanding of the underlying causes of ESL writing mistakes, confusion sets also have useful practical applications. Generally speaking, reducing the confusion set helps lead the classifiers in the GEC system to a better performance by prohibiting them from considering the outcomes that are both *unlikely* and *misleading*. For example, although ESL learners normally would not confuse *within* with *in*, classifiers may have difficulties telling them apart. Therefore, eliminating *within* from *in*'s confusion set may help the classifier. Generally speaking, by reducing the confusion set's size to rule out these outcomes, although the systems will be disabled from correcting certain types of mistakes, they will often increase the accuracies on more prevalent error types and finally lead to a better *overall* performance. In the past, Rozovskaya & Roth (2010a) showed that by limiting the size of the confusion set for prepositions, their GEC system's performance improved.

One challenge in building a model of confusion sets is that automatic methods typically generate huge lists of words, given the many possible factors that contribute to confound ESL writers. For instance, Dahlmeier & Ng (2011a) observed that ESL collocation errors may be due to similarities of the words' spellings, pronunciations, synonyms, and paraphrases in the writer's native language (L1). However, by including all words that are similar according to any of these factors, one would end up with a large confusion set which introduces difficulties for the classification tasks down the GEC pipeline.

A possible solution is to ask human experts using their knowledge about ESL mistakes to restrict the confusion set. This is the approach taken by Liu et al. (2010) for their GEC system for verb selection. Another alternative is to make use of an ESL corpus in which the mistakes have been corrected by an English teacher; in this case, the confusion sets can be tabulated from the annotations (Rozovskaya & Roth, 2010a; Dahlmeier & Ng, 2011a). A benefit of the corpus-driven approach is that the resulting confusion sets provide a reliable estimation of the distributions of the underlying error patterns. However, this type of annotated corpora take time and effort to develop. Moreover, even when an ESL student makes many mistakes, the proportion of the writing that contains no error is still much greater. For example, in the NUS Corpus of Learner English (NUCLE) corpus (Dahlmeier & Ng, 2011b), there are a total of 3,302 preposition mistakes out of a total of 147,087 prepositions. Therefore, to build confusion sets for open class words such as verbs, one would need a very large annotated corpus.

To address the challenge without relying on extensive human involvement, this paper proposes methods to construct confusion sets directly from standard English corpora (Section 3). We conjecture that standard English corpora contain enough information for us to infer ESL learners confusions. This is because learners' confusions are mainly caused by learners' understandings of word similarities, which is developed while studying standard English texts.

What knowledge do ESL learners learn about words? There are mainly two views. One view is that learning words mean understanding the words' meanings and their relations to one another. Another view is that learning words mainly means understanding which word to choose under which conditions.

In *lexical semantics*, people hold the first view. In this area, researchers try to find how and what words mean, denote, and their relations/similarities. This view tends to explain the cause of confusions to be the similarities between words. Dahlmeier & Ng (2011a); Liu et al. (2010) take this view in confusion set construction. They build confusion sets containing the words that are similar in semantic meanings.

In *language modeling*, people hold the second view. People consider the ability of choosing the appropriate word under each context to imply the mastery of the *language*, which include the understandings of the *words* in the language. This view tends to explain the cause of confusions to be the learners' incapability to completely manage how to use words.

3 Automatic Confusion Sets Construction

ESL writers are more likely to confuse words that they find to be similar during their language learning. In this section we present three models that simulates how ESL learners might learn words. In the first two subsections, we describe models of separately learning words' intensions and extensions, respectively. In the last subsection, we introduce a model that is optimized for learning the intensions and extensions of words all together. Within each subsection, we also develop the reason of ESL writers' confusions, and propose the corresponding way to automatically construct confusion sets.

3.1 Learning Words' Intensions – Distributional Models

Under an *intension based* perspective, a learner's primary goal is to understand word meanings, and it is the similarities between words’ intensions that cause word choice confusions. However, this is not to say that learners ignore word usages. Indeed, although dictionary entries contain direct definitions of words, researches in the past showed that learners do not learn by memorizing dictionary entries; instead, they infer words’ meaning/function from the context, and then connecting the new words to the words they are already feel familiar(Fischer, 1990). Under this perspective, learning the extensions of words is not explicit, it is a means to achieve the primary goal of understanding word meanings.

To simulates an intension based learner, we build a model of word similarity metrics from processing standard English text. Specifically, we build *distributional models* in which the similarities of words are calculated from a comparison of the contexts they appear in (Pereira et al., 1993; Lin, 1998; Lee, 1999). Then, to fill in a word's confusion set, we pick the words that are most similar according to the metric. Pantel & Lin (2002) showed this method is able to yield similarities that correlate well with the similarities of words' intensions.

In our work, we calculate the words' intension similarity by using a distributional model (Pereira et al., 1993; Lee, 1999), in which each preposition is represented as a distributional vector of its context features. Examples of usage contexts that have been shown to be relevant for the task of preposition selection in previous work (De Felice, 2008; J. Tetreault et al., 2010; Dahlmeier & Ng, 2011a) include:

Gov: the syntactic dependency governors of the preposition

Obj: the dependency objects of the preposition

GovTag, ObjTag: the part-of-speech tags of the dependency governors and objects

L1-Trans: L1 translations of the preposition

We employ **Gov, Obj, GovTag, ObjTag** features to capture the grammatical context of the preposition selection. We also employ **L1-Trans** to capture both the intended semantic meaning of the preposition and the **L1** background information which was shown to be relevant to confusions(Rozovskaya & Roth, 2010a; Dahlmeier & Ng, 2011a).

The distribution of each preposition's usage context can be estimated from a standard English corpus. Then the similarity between any pair of preposition vectors can be computed using common distance metrics such as: KL-Divergence, Euclidean distance, and cosine similarity.

This approach, however, may not be appropriate for our problem for the following two reasons:

Firstly, under a distributional model, two prepositions are considered similar only if the distribution of all their usages are similar. This is a strong restriction in the sense that two prepositions might only be similar under certain specific usage contexts but are not generally similar. For example, the prepositions *of* and *for* typically have fairly distinctive usages; however, ESL writers often confuse the two if the previous word was *need*.

Secondly, even if two words have similar usages under certain usage context, i.e. have similar probabilities of being used(e.g. both with 0.2 probability), people still may not be likely to confuse them with each other – instead, they are more likely to confuse them with a third word which have higher probabilities(e.g. 0.5). This is because the learner is more likely to pick the word that seems most plausible in the context, if without further information.

3.2 Learning Words' Extensions – Preposition Selector

Under an *extension-only* model, it is assumed that the learners' main goal is to understand how to choose words in a given context, and that they learn about such knowledge from standard English text. Because classifiers can also be trained to choose words, we simulate ESL learners' learning process as training a classifier for the word selections task(J. R. Tetreault & Chodorow, 2008; J. Tetreault et al., 2010) on standard English text. The trained classifier can be seen as a type of language model: given a context, it predicts the most likely word in that context.

Under this model, it is expected that the word choice confusions are mainly caused by the learners' incapability to completely master the word usages. Therefore, to see what confusions an ESL learner may have, we then rerun the trained classifier on the training data to collect the mistakes it makes.

3.3 Learning Both Intensions and Extensions – RCA

We believe the knowledge of word intensions and extensions build on top of each other while learners learn English words. Therefore we propose a model that reflects the interactions between the understandings of words' intensions and extensions; it works toward making the intensions and extensions compatible with each other. Similar with the model in section 3.1, in the end, we build words' confusion sets by filling them in with words that are most similar in their intensions.

Our new model describes a two step process when a learner makes word selection choices: by examining the context, he/she will first think about an intension to convey; then he/she chooses a word that conveys as similar an intension as possible. We will refer to the first step as making *intension decisions,* and the second step as making *word choice decisions*. The goal of their learning is to become more comfortable about the word choices in standard English texts.

We formalize the learning process described above mathematically, to facilitate further analysis:

Intension Space We firstly assume that all possible intensions may be embedded in an Euclidean space S. Two intensions are similar when their locations in S are close to each other. The n prepositions $w_1, \ldots, w_n$ have corresponding intensions $\vec{v}_1, \ldots, \vec{v}_n \in S$. Because we mainly focus on these n prepositions, we assume that all intensions during learners' learning process can be described by a linear interpolation of the n prepositions' intension vectors $v_1, \ldots, v_n$. That is, the subspace containing all intensions learners consider has at most n dimensions. We therefore may assume $S = \mathbb{R}^n$, without loss of generality. Further, we denote $V = (\vec{v}_1, \ldots, \vec{v}_n)$, $I_i = (\underbrace{0, \ldots, 0}_{i-1}, 1, \underbrace{0, \ldots, 0}_{n-i})^T$, so that $\vec{v}_i = VI_i$. Because the n prepositions cannot be too similar to each other, their intensions $\vec{v}_1, \ldots, \vec{v}_n$ cannot clutter. We ensure this by forcing $|\det(V)| \geq 1$[2].

Intension Decisions We model the learners' ability to make *intension decisions* as a function $\vec{f}$ which maps a context C to a point in the intension space $\vec{f}(C) \in S$, which points to the intension the learner would like to choose under context C. C is in the format of a set of relevant contextual features for the preposition choice decisions. Following the discussion in section 3.1, we consider the relevant contextual features **Gov, Obj, GovTag, ObjTag, L1-Trans**.

The Uncomfortness of Word Choice Decisions For one word usage sample (C, w_i), where C is the context, and w_i is the actual preposition choice, we define the "uncomfortness" of the ESL learner by $||\vec{f}(C) - \vec{v}_i||^2$. This means: the more difference between the learner's expected word choice $\vec{f}(C)$ and the actual word choice $\vec{v}_i$, the more "uncomfortable" the learners are.

Learning Goal We assume that the learners learn about the word usages from some standard English corpus[3] D, containing word usage samples in the format (C, w_i). The learners' learning goal is to find V and $\vec{f}$ which get them most "comfortable" with the word usages in D. Therefore, mathematically, the learners' objective is to minimize the overall uncomfortness on the English text D: $\min_{\vec{f},V} \sum_{(C,w_i)\in D} ||\vec{f}(C) - v_i||^2$.

[2] Because $|\det(V)|$ is the area circled by the word vectors in V, forcing it to be higher than 1 can be interpreted as assuming the learners know beforehand that the prepositions cannot be too similar to the others.

[3] Although there may be other sources where the learners may obtain English knowledge from, such as dictionaries, the learners would learn word usages better from texts(Fischer, 1990).

Following the discussion above, we formalize the learning process of ESL learners as finding the best *uncluttered* word vectors V and word usage patterns $\vec{f}$ which together minimize the "uncomfortness" function over some standard English corpus D:

$$\min_{\vec{f},V} \sum_{(C,w_i)\in D} ||\vec{f}(C) - VI_i||^2 \quad s.t. |\det(V)| \geq 1 \tag{1}$$

We calculate the optimal set of word vectors V in the optimization problem above by firstly reducing the problem into a Minimization of Within Class Distances problem, as shown in Appendix A, and then solving it using the Relevance Component Analysis(RCA) algorithm(Bar-Hillel et al., 2006).

In the end, we will be able to obtain the word vectors for prepositions $VI_1, \ldots, VI_n$, and therefore also their similarities by calculating the distance between them (the distance between prepositions w_i, w_j is $||VI_i - VI_j||$). According to our model's assumption, after ESL learners' learning, the similarities of intensions of two words' will highly correlate to this distance. We can therefore fill in the confusion set for every preposition with the prepositions that have the least distances to it.

This approach is similar to the approach described in Section 3.2 in that it also focuses on the similarities of preposition usages under specific contexts. The two approaches differ, however, in their treatments of the degrees to which words are considered to be similar. For example, consider a corpus where under some certain context C, prepositions p_a, p_b and p_c occur 101, 100, 100 times, respectively. Using RCA, the system would consider all three to be mutually confusable because they appear almost equally frequently in the same context. On the other hand, while the preposition selector considers p_b and p_c to be confusable with p_a, it does not conclude that p_b and p_c are also mutually confusable under context C.

Thus, if most usage contexts contain only one or two preposition types, the preposition selector and RCA may produce similar confusion sets; but if the data also include usage contexts that contain three or more preposition types, RCA may offer confusion sets based on a more globally optimized similarity metric.

4 Experimental Setup

We conduct experiments to compare different methods for constructing confusion sets. To evaluate the confusion sets' qualities, we examine how they impact the performance of an end-to-end grammar error correction (GEC) system. In particular, we train a separate classifier for each preposition using only training examples that are covered by the confusion set, a setup similar to the **NegL1** system as described in (Rozovskaya & Roth, 2010a). Additionally, we also compare the confusion sets with an intrinsic evaluation; we measure how well each method's confusion sets match real ESL mistakes by calculating their *coverage* on an annotated ESL corpus.

4.1 Data

As the ground-truth for our experiments, we use the NUS Corpus of Learner English(NUCLE) (Dahlmeier & Ng, 2011b). This is an error-annotated ESL corpus; that is, the writers' mistakes have been identified and corrected by an English teacher. In this collection, many writers' native (L1) language is Chinese. Following the methodologies established in other studies on the preposition selection problem, we focus on the 36 most frequent prepositions[4]. We used 80% of the full corpus for training, 10% for development and 10% for testing.

We use the NUCLE corpus in several ways. First, it is used to establish upper-bound confusion sets. We constructed these "gold" confusion sets by tabulating the observed preposition errors in the corpus. Second, it is used as a source of training data for the end-to-end GEC system[5]. For each confusion set construction method, we extract from the training portion of NUCLE those instances that are consistent with the proposed confusion sets to train the GEC system. The trained systems are then tested on the unfiltered test set. Third, it is used as the ground truth for computing the coverage metric.

The non-ESL corpus used for constructing confusion sets is the Foreign Broadcast Information Service (FBIS) corpus, which is a Chinese-English bilingual corpus. For most experiments, only the English portion is used. For experiments that make use of L1 translations, we extracted the Chinese translations for English prepositions using the GIZA++ (Och & Ney, 2004) implementation of the IBM word alignment model (Brown et al., 1993). Of the FBIS corpus, we used its first 32,000 sentences, which contain 151,767 prepositions.

4.2 Metrics

4.2.1 Extrinsic Evaluation

We use F_1-measure to evaluate the confusion sets' effects on the GEC system

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

where **precision** is the number of suggested corrections that agree with the human annotator divided by the total number of proposed corrections by the system, and **recall** is the number of suggested corrections that agree with the human annotator divided by the total number of errors annotated by the human annotator.

A challenge faced by automatic GEC system is that ESL writers do not make mistakes on most of the usual cases. In NUCLE, 1.3% of the preposition instances contain an error. To reduce the class imbalance for the underlying classifiers during training, we follow the methodology used by Dahlmeier & Ng (2011b) to keep all instances that contain an error and retain a random sample of q percent of the correct instances in the training data. In our experiments, the value of q ($20\% \leq q \leq 40\%$) is tuned on development data. We keep the test data as it is. That is, the filtering we discussed above is only applied on the training data.

4.2.2 Intrinsic Evaluation: Coverage

When an ESL student mistakenly uses some preposition instead of the correct one, the wrong preposition is not necessarily in the proposed confusion set list. We refer to the proportion of

[4] These preposition words include *about, along, among, around, as, at, beside, besides, between, by, down, during, except, for, from, in, inside, into, of, off, on, onto, outside, over, through, to, toward, towards, under, underneath, until, up, upon, with, within, without*

[5] While using NUCLE to train the GEC system seems in contradiction with our overall aim of reducing our reliance on error-annotated corpus, we argue that the usage is appropriate here because we need to compare different approaches of constructing confusion sets without interference from other factors. We do not pursue alternatives such as injecting noise into standard English as training data (Rozovskaya & Roth, 2010a,b) to avoid unintended interactions between the confusion sets and the error generation methods.

ESL students' mistakes in a corpus that fall into the proposed confusion set list as the *coverage* of the confusion set list on that corpus.

The coverage metric can be seen as measuring *recall*: how well does the proposed confusion set table cover the mistakes in some ESL corpus? If each confusion set includes all the prepositions, then the coverage would be 100%. As discussed earlier, in order for the confusion sets to be useful, they cannot be too large. A high quality confusion set table is one whose confusion sets are small in their sizes but cover the majority of the mistakes seen in the ESL corpus.

4.3 Confusion Set Construction Methods

Our experiments compare the following confusion set construction methods:

The Trivial Confusion Sets(all preps) To show the confusion sets' effect in general from comparison, we establish a baseline by using the trivial confusion sets, in which all prepositions are considered to be confusable to each other.

Construction from NUCLE(gold) We establish the upper-bound of the confusion set table by tabulating the preposition mistakes in NUCLE. This confusion set table contains the most prepositions, and therefore is the one with the highest coverage of ESL mistakes.

Construction by Distributional Similarity Metrics As described in Section 3.1, this model represents a preposition as a feature vector and directly computes the distance between pairs of prepositions to construct confusion sets. The values of the feature vectors are computed from the FBIS corpus. Three standard distance/similarity measures are used: KL-Divergence(kl div), Euclidean Distance(euc dist) and Cosine Similarity(cos sim).

Construction from Preposition Selector Errors(selector) Section 3.2 proposes generating confusion sets from classification errors. Here, we train a Maximum Entropy classifier[6] for the preposition selection task on the FBIS corpus, and rerun the classifier on the same data to collect the mistakes it still makes.

Construction by Word Usage Similarity Modeling(RCA) In Section 3.3 we proposed to simulate ESL learners' learning of both words' intensions and extensions. We formalize their learning as an optimization problem and then calculate words' intensions and extensions using the RCA algorithm(Bar-Hillel et al., 2006). The final confusion sets contain words which have similar intensions.

4.3.1 Fixing Sizes of Confusion Sets

Our evaluation fixes the size of the confusion sets in the final confusion set tables to be N, where $3 \leq N \leq 7$. This is mainly because confusion sets tables with sizes greater than 7 are able to cover over 90% of the ESL mistakes, and increasing confusion sets' sizes from there start to hurt the GEC systems' performance. On the other hand, when the sizes are too small, the confusion set lists prevents the GEC system from making reasonable corrections.

5 Experiments

We compare the proposed methods of constructing confusion sets by using the resulting confusion sets in an end-to-end GEC system as described in Section 4. The experiments aim

[6] We used the package downloaded from `http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html`

to address the following questions: (1) How does the proposed method for automatically constructing confusion sets from non-ESL corpus compare against those developed from error-annotated ESL corpus? (2) Regarding the models for ESL learners' word learning, does considering the interactions between their learning of words' intensions and extensions help to capture the learners' confusions? (3) How are these models affected by the choices of different context feature groups? (4) How would the quality be affected by the choice of confusion set sizes?

Figure 1 shows a summary of the results. Each plot shows the GEC system's *performance* versus the *size* of the confusion sets for each confusion sets construction method under a different set of context feature choices [7]. In the baseline all preps, because we always fix the confusion sets' size to be a constant number 36 to contain all prepositions, the resulting curves are displayed as horizontal lines in the figures.

We make four observations:

First, regarding the use of non-ESL corpus, the experimental results suggest that confusion sets that are automatically constructed from non-ESL corpus is competitive with those constructed from an error-corrected ESL corpus. When picking the best feature sets **Gov,Obj,L1-Trans** in RCA, the GEC system can perform as well as if it were using the gold confusion sets constructed from a corpus containing 150K preposition usages.

Second, regarding the models for ESL learners' word learning, our experiments suggest that the learners' confusions are better captured when we model their learning of both words' intensions and extensions altogether. In our experimental results, confusion sets constructed by RCA model, which considers the interaction of words' intensions and extensions, consistently outperforms the other automatic methods selector, kl div, euc dist, cos sim, which only consider the learning of either words' intensions or their extensions.

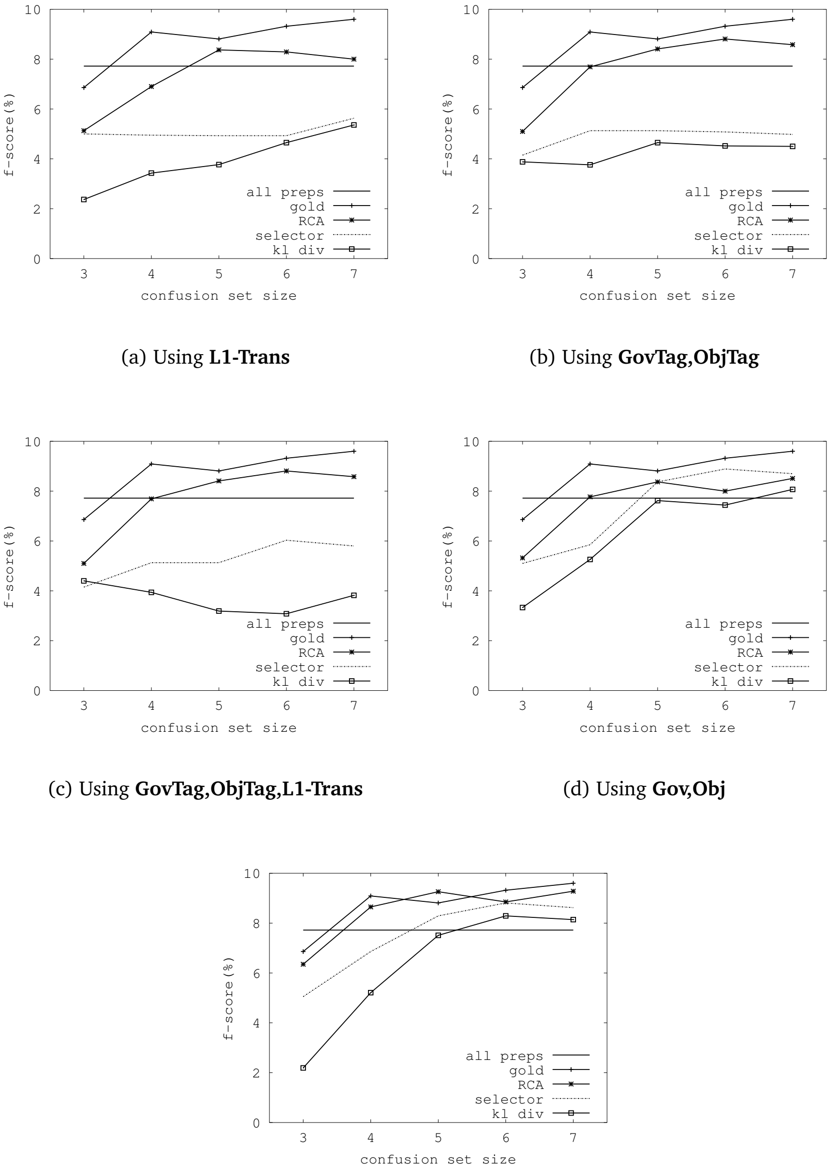
Third, regarding the feature sets used in constructing confusion sets, we find that in general all the models tend to perform better when they use more features. For example, by using **Gov,Obj** in addition to **L1-Trans**, selector raises the GEC system's F-score from 5.00% to 8.81%. RCA, however, is more stable with respect to the feature set changes. We separately show, for these two models, a comparison of the features' effects on them in Figure 2.

Fourth, our evaluation confirms that, in general, using confusion sets helps improving the GEC system's performance. This is because by limiting the confusion set's sizes, one can greatly reduce the underlying classifiers' mis-classification errors, at the cost of reducing their coverage a little. These two factors together lead to positive changes overall. To further demonstrate this effect, we show in Table 1 statistics of the decomposition of GEC systems' errors on the testing dataset. Also worth noticing is that our proposed approach (RCA), although having a slightly less coverage compared to the (gold), reduces mis-classification errors even further.

5.1 Discussions

The experiments above demonstrated RCA's strength over other methods. In this section we provide more in-depth analysis on the differences between RCA and other methods, by comparing those methods' effects on the GEC system's precision and recall separately.

[7] Note that among the standard similarity metrics, we only plot kl div's F_1-scores because it performs better or similar to the other two methods in most of the cases. In later experiments, we will also only demonstrate the best of the three when all of them are performing similarly.



(a) Using **L1-Trans**

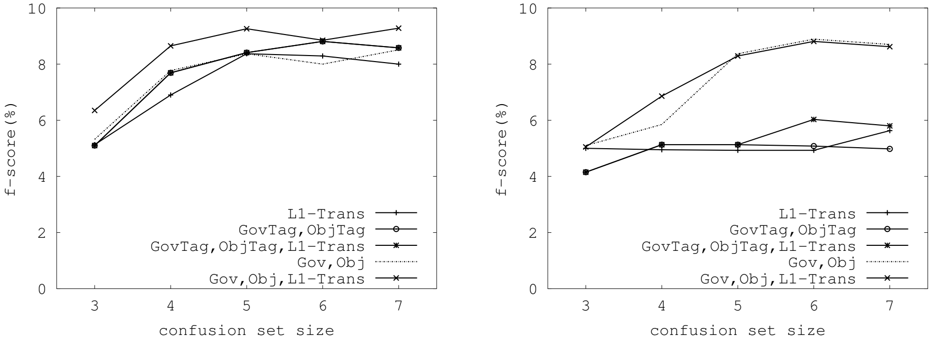
(b) Using **GovTag,ObjTag**

(c) Using **GovTag,ObjTag,L1-Trans**

(d) Using **Gov,Obj**

(e) Using **Gov,Obj,L1-Trans**

Figure 1: F_1-Scores of different confusion set construction methods. For each of the five feature combinations, a plot demonstrates the performance of different methods using that feature combination. We display every method's performance as a curve in which each point represents the GEC system's F_1-Score when using that method to construct confusion set list of a particular size for the 36 prepositions.



(a) RCA methods' F_1 scores

(b) Selector methods' F_1 scores

Figure 2: F_1-scores of models using different feature sets to build confusion sets for all 36 prepositions.

	all	size=3		size=4		size=5		size=6		size=7	
	preps	gold	RCA	gold	RCA	gold	RCA	gold	RCA	gold	RCA
Out of Coverage	0	54	58	37	42	33	37	24	29	22	19
Mis-classification	284	135	119	162	147	173	158	189	176	203	195

Table 1: Confusion sets help reducing mis-classification errors. Here we categorize the GEC system's mistakes by whether they are caused by the confusion sets. *Out of Coverage* represents the cases where confusion sets precluded the right correction to be made, while *Mis-classification* includes all the other cases where the underlying classifiers are responsible for the prediction mistakes. The RCA we demonstrate here uses **Gov,Obj,L1-Trans** features.

We fix the feature set that all methods use to be **Gov,Obj,L1-Trans** in the discussion, because it allows all models to perform their best.

5.1.1 Precision

In Figure 3, comparing with the all prep baseline, we see that by limiting the classifiers' choices, confusion sets are indeed able to raise up GEC systems' precision. The confusion sets computed by RCA and euc dist are more helpful in raising the GEC system's precision, in contrast with selector. The difference is more significant when the confusion sets are small.

5.1.2 Confusion Set Coverage

Furthermore, we would like to provide an analysis of the GEC system's recalls, which is, in our setup, mainly *affected* by the number of ESL mistakes that are precluded from classifiers' consideration by the confusion sets. We measure this by calculating the proportion of ESL mistakes they cover using the metrics developed in 4.2.2. The coverage also reflects one confusion set's match to ESL students' real mistakes.

Shown in Figure 4 are the coverage of confusion sets constructed by different models, of different sizes. RCA and the selector greatly outperform other automatic approaches.

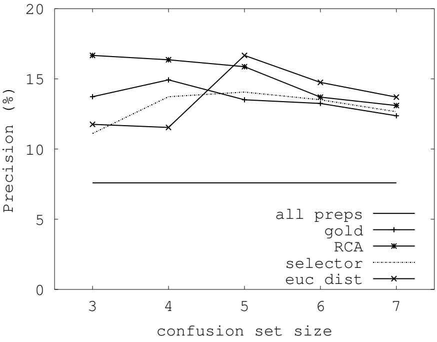


Figure 3: Precision of different methods on NUCLE

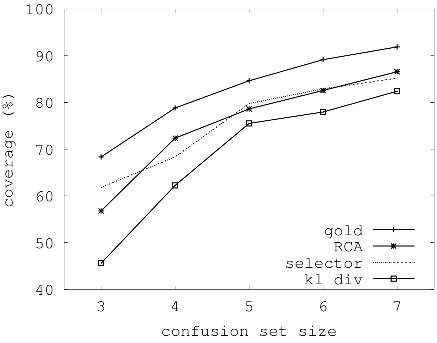


Figure 4: Coverage of confusion sets in different models using features **Gov,Obj,L1-Trans**

6 Conclusions

We proposed a method to automatically construct confusion sets for preposition errors without relying on any annotated ESL corpus or human post-processing. Based on the notion that ESL word selection errors are mainly because ESL learners are not able to choose between similar words, we build a model that analyzes which words might appear similar to each other to an ESL learner. Our model applies an algorithm called Relevance Component Analysis (Bar-Hillel et al., 2006) to describe how an ESL learner might learn both words' intensions and extensions from reading English text. The resulting confusion sets have been shown to both improve GEC system's performance, and correlate well with real ESL mistakes. Also, by modeling the interaction between the intensional and extensional knowledge in ESL learners' learning, our model ends up with better confusion sets than the models considering the development of *only* intensional or extensional knowledge. One key strength of our proposed technique is that because it only relies on standard English corpora, it is more scalable. Although this paper focuses on prepositions, the proposed approach may be applicable to other word classes.

Acknowledgments

This work is supported by U.S. National Science Foundation Grant IIS-0745914. We thank the anonymous reviewers for their suggestions; we also thank Janyce Wiebe, Joel Tetreault, Yafei Wei and Lingjia Deng for helpful discussions.

A Solving the Optimization Problem for Word Usage Similarity

To solve the minimization problem in formula 1, we will first cast it into a Minimization of Within Class Distances problem.

Firstly, suppose there are N unique contexts $C_1, \ldots, C_N$ in the corpus, note that by grouping the samples with same contexts together, we may rewrite formula 1 as:

$$\min_{f,V} \sum_{1\le k\le N} \sum_{(C_k,w_i)\in D_k} ||\vec{f}(C_k) - VI_i||^2 \quad s.t. |\det(V)| \ge 1$$

where $D_k = \{(C, w_i) \in D \mid C = C_k\}$.

Secondly, for a certain V, the optimal function $\vec{f}$ which minimizes the cost function should satisfy: $\vec{f}(C_k) = \frac{\sum_{(C_k,w_i)\in D_k} VI_i}{|D_k|} = V\vec{m}_k$, where $\vec{m}_k = \frac{\sum_{(C_k,w_i)\in D_k} I_i}{|D_k|}$. That is, $\vec{f}$ should map context C_k to the centroid of the word choice vectors in group D_k. We may therefore rewrite the formula above as:

$$\begin{aligned} &\min_{V} \sum_{1\le k\le N} \sum_{(C_k,w_i)\in D_k} ||V\vec{m}_k - VI_i||^2 \quad s.t. |\det(V)| \ge 1 \\ \Leftrightarrow \quad &\min_{V} \sum_{1\le k\le N} \sum_{(C_k,w_i)\in D_k} ||\vec{m}_k - I_i||^2_{V^TV} \quad s.t. \det(V^TV) \ge 1 \end{aligned}$$

where the notation $||\vec{t}||_B$ is the Mahalanobis distance: $||\vec{t}||_B = \sqrt{\vec{t}^T B \vec{t}}$. Together, this gives us the exact equation for the minimization of within class distances problem that the RCA algorithm may solve(Bar-Hillel et al., 2006, p. 945). We therefore directly apply the RCA algorithm to calculate the optimal V: $V = T\hat{R}^{-\frac{1}{2}}$ where T is a constant number and

$$\hat{R} = \sum_{1\le k\le N} \sum_{(C_k,w_i)\in D_k} (I_i - \vec{m}_k)(I_i - \vec{m}_k)^T$$

References

Bar-Hillel, A., Hertz, T., Shental, N., & Weinshall, D. (2006). Learning a mahalanobis metric from equivalence constraints. *Journal of Machine Learning Research*, *6*(1), 937.

Brockett, C., Dolan, W. B., & Gamon, M. (2006). Correcting esl errors using phrasal smt techniques. In *Proceedings of the 21st international conference on computational linguistics and the 44th annual meeting of the association for computational linguistics* (pp. 249–256). Sydney, Australia: Association for Computational Linguistics.

Brown, P., Pietra, V., Pietra, S., & Mercer, R. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, *19*(2), 263–311.

Chalmers, D. J. (2002). On sense and intension. *Noûs*, *36*, 135–182.

Chodorow, M., Tetreault, J. R., & Han, N.-R. (2007). Detection of grammatical errors involving prepositions. In *Proceedings of the fourth acl-sigsem workshop on prepositions* (pp. 25–30). Prague, Czech Republic: Association for Computational Linguistics.

Dahlmeier, D., & Ng, H. (2011a, July). Correcting semantic collocation errors with l1-induced paraphrases. In *Proceedings of the 2011 conference on empirical methods in natural language processing* (pp. 107–117). Edinburgh, Scotland, UK: Association for Computational Linguistics.

Dahlmeier, D., & Ng, H. T. (2011b). Grammatical error correction with alternating structure optimization. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies - volume 1* (pp. 915–923). Portland, Oregon, USA: Association for Computational Linguistics.

De Felice, R. (2008). *Automatic error detection in non-native english*. Unpublished doctoral dissertation, University of Oxford.

Fischer, U. (1990). *How students learn words from a dictionary and in context*. Unpublished doctoral dissertation, Princeton University.

Gamon, M., Gao, J., Brockett, C., Klementiev, A., Dolan, W., Belenko, D., & Vanderwende, L. (2009). Using contextual speller techniques and language modeling for esl error correction. *Urbana*, *51*, 61801.

Leacock, C., Chodorow, M., Gamon, M., & Tetreault, J. (2010). Automated grammatical error detection for language learners. *Synthesis lectures on human language technologies*, *3*(1), 1–134.

Lee, L. (1999). Measures of distributional similarity. In *Proceedings of the 37th annual meeting of the association for computational linguistics on computational linguistics* (pp. 25–32). College Park, Maryland: Association for Computational Linguistics.

Lin, D. (1998). An information-theoretic definition of similarity. In *Proceedings of the fifteenth international conference on machine learning* (pp. 296–304). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Liu, X., Han, B., Li, K., Stiller, S. H., & Zhou, M. (2010). Srl-based verb selection for esl. In *Proceedings of the 2010 conference on empirical methods in natural language processing* (pp. 1068–1076). Cambridge, Massachusetts: Association for Computational Linguistics.

Och, F., & Ney, H. (2004). The alignment template approach to statistical machine translation. *Computational Linguistics*, *30*(4), 417–449.

Pantel, P., & Lin, D. (2002). Discovering word senses from text. In *Proceedings of the eighth acm sigkdd international conference on knowledge discovery and data mining* (pp. 613–619). Edmonton, Alberta, Canada: ACM.

Pereira, F., Tishby, N., & Lee, L. (1993). Distributional clustering of english words. In *Proceedings of the 31st annual meeting on association for computational linguistics* (pp. 183–190). Columbus, Ohio: Association for Computational Linguistics.

Rozovskaya, A., & Roth, D. (2010a). Generating confusion sets for context-sensitive error correction. In *Proceedings of the 2010 conference on empirical methods in natural language processing* (pp. 961–970). Cambridge, Massachusetts: Association for Computational Linguistics.

Rozovskaya, A., & Roth, D. (2010b). Training paradigms for correcting errors in grammar and usage. In *Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics* (pp. 154–162).

Rozovskaya, A., & Roth, D. (2011, June). Algorithm selection and model adaptation for esl correction tasks. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies* (pp. 924–933). Portland, Oregon, USA: Association for Computational Linguistics.

Tetreault, J., Foster, J., & Chodorow, M. (2010). Using parse features for preposition selection and error detection. In *Proceedings of the acl 2010 conference short papers* (pp. 353–358). Uppsala, Sweden: Association for Computational Linguistics.

Tetreault, J. R., & Chodorow, M. (2008). The ups and downs of preposition error detection in esl writing. In *Proceedings of the 22nd international conference on computational linguistics - volume 1* (pp. 865–872). Manchester, United Kingdom: Association for Computational Linguistics.