Using Learner Data to Improve Error Correction in Adjective–Noun Combinations

Ekaterina Kochmar
Alta Institute
Computer Laboratory
University of Cambridge
ek358@cl.cam.ac.uk

Ted Briscoe
Alta Institute
Computer Laboratory
University of Cambridge
ejb@cl.cam.ac.uk

Abstract

This paper presents a novel approach to error correction in content words in learner writing focussing on adjective–noun (AN) combinations. We show how error patterns can be used to improve the performance of the error correction system, and demonstrate that our approach is capable of suggesting an appropriate correction within the top two alternatives in half of the cases and within top 10 alternatives in 71% of the cases, performing with an MRR of 0.5061. We then integrate our error correction system with a state-of-the-art content word error detection system and discuss the results.

1 Introduction

The task of error detection and correction (EDC) on non-native texts, as well as research on learner language in general, has attracted much attention recently (Leacock et al., 2014; Ng et al., 2014; Ng et al., 2013; Dale et al., 2012). The field has been dominated by EDC for grammatical errors and errors in the use of articles and prepositions (Ng et al., 2013; Rozovskaya and Roth, 2011; Chodorow et al., 2010; Gamon et al., 2008; Brockett et al., 2006; Han et al., 2006).

More recently, however, the need to address other error types has been recognised (Kochmar and Briscoe, 2014; Ng et al., 2014; Rozovskaya et al., 2014; Sawai et al., 2013; Dahlmeier and Ng, 2011). Among these, errors in content words are the third most frequent error type after errors in articles and prepositions (Leacock et al., 2014; Ng et al., 2014). The correct use of content words is notoriously hard for language learners to master, while importance of the correct word choice for successful writing has long been recognised (Leacock and Chodorow, 2003; Johnson, 2000; Santos, 1988).

The major difficulty is that correct word choice is not governed by any strictly defined rules: native speakers know that *powerful computer* is preferred over *strong computer*, while *strong tea* is preferred over *powerful tea* (Leacock et al., 2014), but language learners often find themselves unsure of how to choose an appropriate word. As a result, they often confuse words that are similar in meaning or spelling, overuse words with general meaning, or select words based on their L1s (Kochmar and Briscoe, 2014; Dahlmeier and Ng, 2011).

Previous work on EDC for content words has also demonstrated that since these error types are substantially different from errors with function words, they require different approaches. The most widely adopted approach to EDC for function words relies on availability of finite confusion sets. The task can then be cast as multi-class classification with the number of classes equal to the number of possible alternatives. Detection and correction can be done simultaneously: if the alternative chosen by the classifier is different from the original word, this is flagged as an error. However, content word errors cannot be defined in terms of a general and finite set of confusion pairs, and the set of alternatives in each case depends on the choice of original word. Moreover, it has been argued that error detection for content words should be performed independently from error correction (Kochmar and Briscoe, 2014).

Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications, 2015, pages 233–242,
Denver, Colorado, June 4, 2015.

In this work, we focus on error correction in content words and, in particular, investigate error correction in adjective–noun (AN) combinations using several publicly-available learner error datasets for this type of construction. At the same time, we believe that a similar approach can be applied to other types of content word combinations. Specifically, we make the following contributions:

1. We explore different ways to construct the correction sets and to rank the alternatives with respect to their appropriateness. We report the coverage of different resources and assess the ranked lists of suggestions.

2. We show that learner text is a useful source of possible corrections for content words. In addition, we demonstrate how error patterns extracted from learner text can be used to improve the ranking of the alternatives.

3. We present an EDC system for AN combinations which compares favourably to the previous published approaches of which we are aware.

4. We explore the usefulness of self-propagating for an error correction system.

2 Related work

Leacock *et al.* (2014) note that the usual approach to EDC in content words relies on the idea of comparing the writer's choice to possible alternatives, so that if any of the alternatives score higher than the original combination then the original combination is flagged as a possible error and one or more alternatives are suggested as possible corrections. The performance of an EDC algorithm that uses this approach depends on:

- the choice of the source of alternatives;

- the choice of the metric for ranking the alternatives.

The source of alternatives defines the *coverage* of the error correction algorithm, while the *quality* of the system suggestions depends on the choice of an appropriate metric for ranking the alternatives.

Early work on EDC for content words (Wible et al., 2003; Shei and Pain, 2000) relied on the use of reference databases of known learner errors and their corrections. While such approaches can achieve good quality, they cannot provide good coverage.

Previous research considered semantically related confusions between content words as the most frequent type of confusion in learner writing and used WordNet (Miller, 1995), dictionaries and thesauri to search for alternatives (Östling and Knutsson, 2009; Futagi et al., 2008; Shei and Pain, 2000). Since these resources cannot cover alternatives that are not semantically related to the original words, other resources have been considered as well: for example, Dahlmeier and Ng (2011) consider spelling alternatives and homophones as possible corrections.

L1-specific confusions have been reported to cover a substantial portion of errors in content words for some groups of language learners (Chang et al., 2008; Liu, 2002), and some previous EDC approaches have considered using parallel corpora and bilingual dictionaries to generate and rank alternatives (Dahlmeier and Ng, 2011; Chang et al., 2008). L1-specific approaches have shown the best results in EDC for content words so far, but it should be noted that their success relies on availability of high-quality L1-specific resources which is hard to guarantee for the full variety of learner L1s.

At the same time, good performance demonstrated by L1-specific approaches shows the importance of taking learner background into consideration. In contrast to the other resources like WordNet and thesauri, which can only cover confusions between words in the L2, use of parallel corpora and bilingual dictionaries gives access to the types of confusions which cannot be captured by any L2 resources. Learner corpora and databases of text revisions can be used to similar effect.

For example, Rozovskaya and Roth (2011) show that performance of an EDC algorithm applied to articles and prepositions can be improved if the classifier uses L1-specific priors, with the priors being set using the distribution of confusion pairs in learner texts. Sawai *et al.* (2013) show that an EDC system that uses a large learner corpus to extract confusion sets outperforms systems that use WordNet and roundtrip translations. Madnani and Cahill (2014)

use a corpus of Wikipedia revisions containing annotated errors in the use of prepositions and their corrections to improve the ranking of the suggestions.

Finally, we note that a number of previous approaches to errors in content words have combined error detection and correction, flagging an original choice as an error if an EDC algorithm is able to find a more frequent or fluent combination (Östling and Knutsson, 2009; Chang et al., 2008; Futagi et al., 2008; Shei and Pain, 2000), while some focussed on error correction only (Dahlmeier and Ng, 2011; Liu et al., 2009). Kochmar and Briscoe (2014) argue that error detection and correction should be performed separately. They show that an EDC algorithm is prone to overcorrection, flagging originally correct combinations as errors, if error detection is dependent on the set of alternatives and if some of these alternatives are judged to be more fluent than the original combination.

We follow Kochmar and Briscoe (2014) and treat error detection and error correction in content words as separate steps. We focus on the correction step, and first implement a simple error correction algorithm that replicates previous approaches to EDC for content words. We believe that performance of this algorithm on our data reflects the state-of-the-art in content error correction. Next, we show how learner data and distribution of confusion pairs can be used to improve the performance of this algorithm.

3 Data

In our experiments, we use three publicly-available datasets of learner errors in AN combinations: the AN dataset extracted from the *Cambridge Learner Corpus* (*CLC*)[1] and annotated with respect to the learner errors in the choice of adjectives and nouns;[2] the AN dataset extracted from the CLC-FCE dataset;[3] and the set of errors in ANs that we have extracted for the purposes of this work from the training and development sets used in the CoNLL-2014 Shared Task on Grammatical Error Correction.[4] We discuss these datasets below.

3.1 Annotated dataset

We use the dataset of AN combinations released by Kochmar and Briscoe (2014). This dataset presents typical learner errors in the use of 61 adjectives that are most problematic for language learners. The examples are annotated with respect to the types of errors committed in the use of adjectives and nouns, and corrections are provided.

Kochmar and Briscoe note that learners often confuse semantically related words (e.g., synonyms, near-synonyms, hypo-/hypernyms). Examples (1) and (2) from Kochmar and Briscoe (2014) illustrate the confusion between the adjective *big* and semantically similar adjectives *large* and *great*:

(1) *big*/large* quantity

(2) *big*/great* importance

In addition, in Kochmar and Briscoe (2014) we note that the adjectives with quite general meaning like *big*, *large* and *great* are often overused by language learners instead of more specific ones, as is illustrated by examples (3) to (6):

(3) *big*/long* history

(4) *bigger*/wider* variety

(5) *greatest*/highest* revenue

(6) *large*/broad* knowledge

Words that seem to be similar in form (either related morphologically or through similar pronunciation) are also often confused by learners. Examples (7) and (8) illustrate this type of confusions:

(7) *classic*/classical* dance

(8) *economical*/economic* crisis

The dataset contains 798 annotated AN combinations, with 340 unique errors.

Table 1 presents the statistics on the error types detected in this dataset. The majority of the errors

[1] http://www.cup.cam.ac.uk/gb/elt/catalogue/subject/custom/item3646603/Cambridge-International-Corpus-Cambridge-Learner-Corpus/

[2] http://ilexir.co.uk/media/an-dataset.xml

[3] http://ilexir.co.uk/applications/adjective-noun-dataset/

[4] http://www.comp.nus.edu.sg/~nlp/conll14st.html

Error type	Distribution
S	56.18%
F	25.88%
N	17.94%

Table 1: Distribution of error types in the annotated dataset.

Word	**Ann. data**	**CLC-FCE**	**NUCLE**
A	63.24%	43.20%	34.15%
N	30.29%	52.63%	60.16%
Both	6.47%	4.17%	5.69%

Table 2: Distribution of errors in the choice of adjectives (A), nouns (N) or both words in the datasets.

involve semantically related words (type S). Form-related confusions occur in 25.88% of the cases (type F); while 17.94% are annotated as errors committed due to other reasons (type N), possibly related to learners' L1s.

3.2 CLC-FCE dataset

The CLC-FCE AN dataset is extracted from the publicly-available CLC-FCE subset of the CLC released by Yannakoudakis *et al.* (2011). The CLC error coding (Nicholls, 2003) has been used to extract the correctly used ANs and those that are annotated as errors due to inappropriate choice of an adjective or/and noun, but the error subtypes for the AN errors are not further specified. We have extracted 456 combinations that have adjective–noun combinations as corrections.

3.3 NUCLE dataset

We have also used the training and development sets from the CoNLL-2014 Shared Task on Grammatical Error Correction (Ng et al., 2014) to extract the incorrect AN combinations. The data for the shared task has been extracted from the *NUCLE* corpus, the *NUS Corpus of Learner English* (Dahlmeier et al., 2013). Unlike the other two datasets it represents a smaller range of L1s, and similarly to the CLC-FCE dataset the errors are not further annotated with respect to their subtypes.

We have preprocessed the data using the RASP parser (Briscoe et al., 2006), and used the error annotation provided to extract the AN combinations that contain errors in the choice of either one or both words. Additionally, we have also checked that the suggested corrections are represented by AN combinations. The extracted dataset contains 369 ANs.

Table 2 reports the distribution of the errors with respect to the incorrect choice of an adjective, noun or both words within AN combinations in all three datasets.

4 Error Correction Algorithm

First, we implement a basic error correction algorithm that replicates the previous approaches to error correction overviewed in §2, and investigate the following aspects of the algorithm:

1. We explore different resources to retrieve alternatives for the adjectives and nouns within incorrect ANs and report the coverage of these resources;

2. The alternative ANs are generated by crossing the sets of alternatives for the individual words, and ranked using a metric assessing AN frequency or fluency in native English. We assess the quality of the ranking using *mean reciprocal rank* (*MRR*) by comparing the system suggestions to the gold standard corrections;

3. Finally, we also show how the confusion sets extracted from the learner data can help improve the ranking and the quality of the suggested corrections.

When reporting the results, we specifically focus on two aspects of the error correction algorithm: the *coverage* estimated as the proportion of gold standard corrections that can be found in any of the resources considered, and the ability of the algorithm to rank the more appropriate corrections higher than the less appropriate ones measured by *MRR* of the gold standard corrections in the system output.

4.1 Word alternatives

We extract word alternatives using three resources:

1. We use the notion of Levenshtein distance (henceforth, Lv) (Levenshtein, 1966) to find the words that learners might have accidentally confused or misspelled. These alternatives can cover errors annotated as form related. To avoid introducing too much change

to the original words, we only consider alternatives that differ from the original words by no more than $1/3$ of the characters in the original word and that start with the same letter as the original word. The generated alternatives are checked against the *British National Corpus* (*BNC*)[5] and the *ukWaC* corpus[6] to avoid generating non-words. This allows the algorithm to find alternatives like *customer* for *costumer* (in **important costumer*), *metropolis* for *metropole* (in **whole metropole*), or *electronic* for *electric* (in **electric society*).

2. We look for further alternatives in WordNet (henceforth, WN) (Miller, 1995), which has previously been widely used to find semantically related words. For each original noun, we extract a set of synonyms and hypo-/hypernyms. For each original adjective, we extract synonyms and the adjectives related via the WN relation *similar-to*. This allows us to cover semantically related confusions, and find alternatives such as *luck* for *fate* (in **good fate*) and *steep* for *heavy* (in **heavy decline*).

3. Both Lv and WN cover confusions that occur in L2, but none of them can cover confusions that occur due to L1-transfer. Therefore, we extract the corrections provided by the annotators in the *Cambridge Learner Corpus* (henceforth, CLC). This approach is similar to that of Madnani and Cahill (2014), but it uses learner data as the database. We believe that the confusion pairs extracted this way cover a substantial portion of errors committed due to L1-transfer, while, computationally, it is much less expensive than the use of bilingual dictionaries or parallel corpora as in Dahlmeier and Ng (2011) or Chang *et al.* (2008). This approach allows us to extract confusion pairs that are covered by the CLC only, for example, *novel* for *roman* (in **historical roman*), *narrow*, *short* and *brief* for *small* (in **small interruption*) or *big*, *high* and *loud* for *strong* (in **strong noise*).

[5] http://www.natcorp.ox.ac.uk

[6] http://wacky.sslmit.unibo.it/doku.php?id=corpora

Setting	Ann. data	CLC-FCE	NUCLE
Lv	0.1588	0.0833	0.0897
WN	0.4353	0.3904	0.2880
CLC	0.7912	0.8684	0.5625
CLC+Lv	0.7971	0.8706	0.5951
CLC+WN	0.8558	0.8904	0.6141
All	**0.8618**	**0.8925**	**0.6467**

Table 3: Coverage of different sets of alternatives.

We assess how many of the gold standard corrections can be found in each of these confusion sets as well as in different combinations of these sets. Coverage of the different resources is reported in Table 3. We note that the CLC as a single source of corrections provides the highest coverage: for example, 79% of erroneous ANs from the annotated dataset and 87% of erroneous ANs in the CLC-FCE can potentially be corrected using only the previous corrections for the content words from the CLC. We note that although the ANs in the annotated dataset have been extracted from the CLC, they have been error-annotated independently. The lower figure of 56% on the NUCLE dataset can be explained by the difference between the CLC and NUCLE corpora since the distribution of errors in these corpora is also different (see Table 2). Nevertheless, we note that the corrections extracted from the CLC still cover a substantial amount of the errors in the NUCLE dataset. A combination of the corrections from the CLC and semantically related words from WordNet covers an additional 6% of ANs in the annotated dataset, 5% in the NUCLE dataset, and 2% in the CLC-FCE dataset, which demonstrates that the majority of the semantically related confusions are already covered by the corrections extracted from the CLC, so WordNet improves the coverage of this resource only marginally. Addition of the form related words (Lv) does not improve coverage significantly.

4.2 Alternative ANs ranking

Once the alternatives for the words within the combinations are collected, the alternative AN combinations are generated by the Cartesian product of the sets of alternatives for the adjectives and the nouns. The alternatives then need to be ranked with respect to their appropriateness.

We apply two simple methods to rank the alternatives: we use the frequency of the generated ANs in a combined BNC and ukWaC corpus, and we also measure collocational strength of the alternative combinations using *normalised pointwise mutual information* (*NPMI*) since PMI-based metrics have been widely used before (see §2):

$$NPMI(AN) = \frac{PMI(AN)}{-log_2(P(AN))} \quad (1)$$

where

$$PMI(AN) = log_2 \frac{P(AN)}{P(A)P(N)} \quad (2)$$

We have noticed that when the full sets of alternatives for the adjectives and nouns are used to generate the AN alternatives, the resulting sets of ANs contain many combinations, with both original words changed to alternative suggestions, that are dissimilar in meaning to the original ANs while often being quite frequent or fluent. As a result, such alternatives are ranked higher than the appropriate corrections. To avoid this, we only consider the alternative ANs where one of the original words is kept unchanged, i.e.:

$\{alternative\ ANs\} = (\{alternative\ adjs\} \times noun) \cup (adj \times \{alternative\ nouns\})$

We evaluate the ranking using the *mean reciprocal rank* (*MRR*):

$$MRR = \frac{1}{|N|} \sum_{i=1}^{|N|} \frac{1}{rank_i} \quad (3)$$

where N is the total number of erroneous ANs considered by our algorithm. *MRR* shows how high the gold standard alternative is ranked in the whole set of alternatives provided.

The results are reported in the upper half of the Table 4. We note that often the wider sets of alternatives for the individual words yield lower ranks for the gold standard corrections since some other frequent AN alternatives are ranked higher by the algorithm.

4.3 Exploitation of confusion probabilities

Next, we consider a novel approach to ranking the alternative ANs. Since we are using the CLC corrections for the adjectives and nouns within the ANs, in addition to the possible corrections themselves we can also use the confusion probabilities – probabilities associated with the words used as corrections given the incorrect word choice – for the pairs of words that we extract from the CLC.

Setting	Ann. set	CLC-FCE	NUCLE
$\texttt{CLC}_{freq}$	**0.3806**	0.3121	0.2275
$\texttt{CLC}_{NPMI}$	0.3752	0.2904	0.1961
$(\texttt{CLC+Lv})_{freq}$	0.3686	**0.3146**	**0.2510**
$(\texttt{CLC+Lv})_{NPMI}$	0.3409	0.2695	0.1977
$(\texttt{CLC+WN})_{freq}$	0.3500	0.2873	0.2267
$(\texttt{CLC+WN})_{NPMI}$	0.3286	0.2552	0.1908
$\texttt{All}_{freq}$	0.3441	0.2881	0.2468
$\texttt{All}_{NPMI}$	0.3032	0.2407	0.1943
$\texttt{All}_{\mathbf{freq'}}$	**0.5061**	**0.4509**	**0.2913**
$\texttt{All}_{\mathbf{NPMI'}}$	0.4843	0.4316	0.2118

Table 4: *MRR* for the alternatives ranking.

We use a refined formula to rank the possible corrections:

$$\begin{aligned} M' &= M \times CP(a_{orig} \rightarrow a_{alt}) \\ &\quad \times CP(n_{orig} \rightarrow n_{alt}) \end{aligned} \quad (4)$$

where M is the measure for ranking the alternatives (frequency or *NPMI*, as before), and CP is the confusion probability of using the alternative word (possible correction) instead of the original one (error) estimated from the examples in the CLC. We set $CP(a/n_{orig} \rightarrow a/n_{orig})$ to 1.0.

For instance, consider an incorrect AN **big enjoyment* and its gold standard correction *great pleasure*. Table 5 shows some alternatives for the words *big* and *enjoyment* with the corresponding corrections and their probabilities extracted from the CLC. If we use these sets of confusion pairs to generate the alternative ANs and rank them with raw frequency, the algorithm will choose *great fun* (7759 in the native corpus) over the gold standard correction *great pleasure* (2829 in the native corpus). However, if we use the confusion probabilities with the new measure (4) the gold standard correction *great pleasure* ($Freq' = 3.8212$) will be ranked higher than *great fun* ($Freq' = 1.1620$). The new measure helps take into account not only the fluency of the correction in the native data but also the appropriateness of a

Original	Alternatives	CP(orig → alt)
big	*great*	0.0144
	large	0.0141
	wide	0.0043
	...	...
	significant	$5.1122 * 10^{-5}$
enjoyment	*pleasure*	0.0938
	entertainment	0.0313
	fun	0.0104
	happiness	0.0052

Table 5: CLC confusion pairs

Top N	Ann. data	CLC-FCE	NUCLE
1	41.18	34.21	21.20
2	49.12	45.18	27.99
3	56.77	50.88	33.70
4	61.77	55.04	38.04
5	65.29	58.55	40.49
6	66.18	61.40	42.39
7	67.35	62.28	43.21
8	68.53	63.60	44.29
9	69.71	65.35	45.38
10	71.18	66.45	46.20
Not found	25.29	19.96	48.64

Table 6: Results breakdown: % of errors covered.

Type	S	F	N
MRR_{found}	0.6007	0.8486	0.6507
Not found	0.1990	0.1705	0.5410

Table 7: Subtype error analysis for the annotated dataset.

particular correction given a learner error.

In addition, this algorithm allows us to consider both words as possibly incorrectly chosen: equation (4) ensures that the alternative combinations where both original words are changed are only ranked higher if they are both very frequent in the native corpus and very likely as a confusion pair since $CP(a/n_{orig} \rightarrow a/n_{orig})$ is set to 1.0.

Finally, if no confusion pairs are found for either an adjective or a noun in the CLC, the algorithm considers the alternatives from other resources and uses standard measures to rank them.

The lower half of Table 4 presents the results of this novel algorithm and compares them to the previous results from §4.2. The new metric consistently improves performance across all three datasets, with the difference in the results being significant at the 0.05 level.

5 Discussion

5.1 Analysis of the results

An *MRR* of 0.4509 and 0.5061 reported in §4.3 implies that for a high number of the ANs from the CLC-FCE and annotated dataset the gold standard correction is ranked first or second in the list of all possible corrections considered by the system. Table 6 presents the breakdown of the results and reports the proportion of ANs for which the gold standard correction is covered by the top N alternatives.

We note the small difference between the number of cases covered by the top 10 system alternatives for the annotated dataset (71.18%) and the upper bound – the total number of corrections that can potentially be found by the system (74.71%) – which shows that the system reaches its potential around the top 10 suggestions. These results also compare favourably to those reported in previous research (Chang et al., 2008; Dahlmeier and Ng, 2011), although direct comparison is not possible due to the differences in the data used.

We also further investigate the performance of the error correction algorithm on the different error subtypes in the annotated dataset (see Table 1). Table 7 presents the proportion of the gold standard corrections for each subtype that are not found by the algorithm, as well as the *MRR* for those corrections that are identified. We see that the highest proportion of gold standard corrections that are not found by the algorithm are the corrections that are not related to the originally used words (type N). This result is not surprising: if the original words and their corrections are not related semantically or in form, it is hard to find the appropriate suggestions. The results also suggest that the system performs best on the errors of type F: a possible reason for this is that errors of this type are more systematic and have smaller confusion sets. For example, the average *MRR* on the set of ANs involving errors in the use of the adjective *elder* in the annotated dataset is 0.875 since most often such ANs require changing

Corpus	MRR_{adj}	MRR_{noun}
Ann	0.5188	0.4312
CLC	0.3986	0.4665
NUCLE	0.3191	0.2608

Table 8: Average *MRR* on the sets of ANs with the errors in the choice of adjectives and nouns.

Setting	Ann. set	CLC-FCE	NUCLE
CLC	0.3806	0.3121	0.2275
CLC+Lv	0.3686	0.3146	0.2510
Augm	**0.4420**	**0.3533**	**0.2614**

Table 9: Augmented sets of alternatives.

the adjective for form related alternatives *elderly* or *older*.

At the same time, we note that the results on the NUCLE dataset are lower than on the two other datasets. In Table 3 we report that about 35% of the gold standard corrections from this dataset are not covered by any of the available sets of alternatives for adjectives and nouns, while the confusion sets extracted from the CLC can only cover about 56% of the cases. We conclude that there might be a substantial difference between the two learner corpora in terms of topics, vocabulary used, learner levels and the distribution of the L1s. We assume that a high number of errors in NUCLE dataset can be caused by reasons other than semantic or form similarity of the words in L2. For example, our system does not suggest the gold standard correction *bill* for **debt* in **medical debt*, or *infrastructural* for **architectural* in **architectural development* because these suggestions are not originally covered by any of the sets of alternatives, including the set of confusion pairs extracted from the CLC.

Table 8 reports the average *MRR* on the sets of ANs involving errors in the choice of adjectives and nouns separately. The NUCLE dataset contains ANs with 105 adjectives and 185 nouns, with 76 adjectives and 145 nouns occurring in the NUCLE ANs only. The low overlap between the sets of individual words explains the differences in performance. Since the annotated dataset contains ANs within a set of frequent adjectives, the algorithm achieves highest performance in correcting adjective-specific errors in this dataset.

5.2 Augmenting sets of alternatives

We investigate whether self-propagation of the system can mitigate the problem of gold standard suggestions not covered by the original sets of alternatives. Some previous research (Shei and Pain, 2000; Chang et al., 2008) has suggested that if an error correction system is implemented in an interactive way, learners can be asked to accept the suggested corrections so that the error–correction pairs can be added to the error database for future reference. We add the gold standard suggestions for the adjectives and nouns from all three datasets to the sets of alternatives and run our error correction system using the augmented sets. For example, we add *bill* to the set of alternatives for *debt* and *infrastructural* to the set of alternatives for *architectural* and check whether the results of the error correction system improve.

Table 9 reports the results. Since we focus on the effect of the sets of alternatives, we run the experiments using one setting of the system only. We note that, since the datasets contain only a few examples for each adjective and noun, we cannot expect to see a significant change in the results if we updated the confusion probabilities and used the refined measure from §4.3. Therefore, we rank the AN alternatives using frequency of occurrence in the corpus of native English. For ease of comparison, we copy the relevant results from Table 4.

The best results obtained in experiments in §4.2 with the original sets of alternatives are underlined, while the results obtained with the augmented sets of alternatives are marked in bold. We note that the results improve, although the difference is not statistically significant across the three datasets.

5.3 Error Detection and Correction System

Finally, we combine the error correction algorithm from §4.3 with the error detection algorithm from Kochmar and Briscoe (2014): the error correction algorithm is applied to the set of erroneous ANs correctly detected by the error detection algorithm.

In Kochmar and Briscoe (2014) we report precision of 0.6850 and recall of 0.5849 on the incorrect examples in the annotated dataset. Some of the errors identified cannot be further corrected by our al-

gorithm since the corrections are longer than two words. MRR of the error correction system applied to the set of detected errors is 0.2532, while for 24.28% of the cases the system does not find a gold standard correction. If these cases are not considered, $MRR_{found} = 0.6831$. We believe that these results reflect state-of-the-art performance for the combined EDC system for AN combinations.

6 Conclusion

In this paper, we have addressed error correction in adjective–noun combinations in learner writing using three publicly available datasets. In particular, we have explored different ways to construct the correction sets and to rank the suggested corrections, and showed that the confusion patterns extracted directly from the learner data not only provide the highest coverage for the system, but can also be used to derive confusion probabilities and improve the overall ranking of the suggestions. We have shown that an error correction system can reach an *MRR* of 0.5061 which compares favourably to the results reported previously.

Further analysis shows that the majority of errors not covered by the algorithm involve confusion between words that are not related semantically or in form and, therefore, cannot be found in L2 resources like WordNet. Our experiments with the augmented sets of alternatives, where we use known learner confusion pairs to further extend the sets of correction candidates, show improvement in the results and suggest that extension of the learner corpus can help system find appropriate corrections. At the same time, the difference in the results obtained on the datasets extracted from the CLC and the NUCLE corpora can be explained by the difference in the topics, learner levels and L1s represented by the two learner corpora. Future research should explore further ways to extend the learner data.

We also note that in the current work we do not consider the wider context for error detection and correction in ANs. In future work we plan to investigate the use of surrounding context for EDC for ANs.

Finally, we have integrated our error correction system with a state-of-the-art content word error detection system. To the best of our knowledge, this is the first attempt to combine two such systems, and we believe that the results obtained – an *MRR* of 0.2532 on the set of errors identified by the error detection algorithm – reflect state-of-the-art performance on the EDC task for AN combinations. Our future work will also extend this approach to other types of content word combinations.

Acknowledgments

We are grateful to Cambridge English Language Assessment and Cambridge University Press for supporting this research and for granting us access to the CLC for research purposes. We also thank the anonymous reviewers for their valuable comments.

References

Ted Briscoe, John Carroll, and Rebecca Watson. 2006. *The second release of the RASP system*. In ACL-Coling06 Interactive Presentation Session, pp. 77–80.

Chris Brockett, William B. Dolan, and Michael Gamon. 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.

Yu-Chia Chang, Jason S. Chang, Hao-Jan Chen, and Hsien-Chin Liou. 2008. *An automatic collocation writing assistant for Taiwanese EFL learners: A case of corpus-based NLP technology*. Computer Assisted Language Learning, 21(3), pp. 283–299.

Martin Chodorow, Michael Gamon, and Joel Tetreault. 2010. *The utility of grammatical error detection systems for English language learners: Feedback and Assessment*. Language Testing, 27(3):335–353.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2012. *Building a large annotated corpus of learner English: The NUS Corpus of Learner English*. In Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 22–31.

Daniel Dahlmeier and Hwee Tou Ng. 2011. *Correcting Semantic Collocation Errors with L1-induced Paraphrases*. In Proceedings of the EMNLP-2011, pp. 107–117.

Robert Dale, Ilya Anisimoff, and George Narroway. 2012. *HOO 2012: A Report on the Preposition and Determiner Error Correction Shared Task*. In Proceedings of the 7th Workshop on the Innovative Use of NLP for Building Educational Applications, pp. 54–62.

Yoko Futagi, Paul Deane, Martin Chodorow and Joel Tetreault. 2009. *A computational approach to detecting collocation errors in the writing of non-native speakers of English*. Computer Assisted Language Learning, 21(4), pp. 353–367.

Michael Gamon, Jianfeng Gao, Chris Brockett, Alexander Klementiev, William Dolan, Dmitriy Be- lenko, and Lucy Vanderwende. 2008. *Using contextual speller techniques and language modeling for ESL error correction*. In Proceedings of IJCNLP, pp. 491–511.

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

Dale D. Johnson. 2000. *Just the Right Word: Vocabulary and Writing*. In R. Indrisano & J. Squire (Eds.), *Perspectives on Writing: Research, Theory, and Practice*, pp. 162–186.

Ekaterina Kochmar and Ted Briscoe. 2014. *Detecting Learner Errors in the Choice of Content Words Using Compositional Distributional Semantics*. In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pp. 1740–1751.

Claudia Leacock and Martin Chodorow. 2003. *Automated Grammatical Error Detection*. In M. D. Shermis and J. C. Burstein (eds.), *Automated Essay Scoring: A Cross-Disciplinary Perspective*, pp. 195–207.

Claudia Leacock, Martin Chodorow, Michael Gamon and Joel Tetreault. 2014. *Automated Grammatical Error Detection for Language Learners*, Second Edition. Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers.

Vladimir I. Levenshtein. 1966. *Binary codes capable of correcting deletions, insertions, and reversals*. Soviet Physics Doklady, 10(8):707–710.

Anne Li-E Liu. 2002. *A corpus-based lexical semantic investigation of verb-noun miscollocations in Taiwan learners English*. Masters thesis, Tamkang University, Taipei.

Anne Li-E Liu, David Wible and Nai-Lung Tsao. 2009. *Automated suggestions for miscollocations*. In Proceedings of the 4th Workshop on Innovative Use of NLP for Building Educational Applications, pp. 47–50.

Nitin Madnani and Aoife Cahill. 2014. *An Explicit Feedback System for Preposition Errors based on Wikipedia Revisions*. In Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 79–88.

George A. Miller. 1995. *WordNet: A Lexical Database for English*. Communications of the ACM, 38(11):39–41.

Hwee Tou Ng, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault. 2013. *The CoNLL- 2013 Shared Task on Grammatical Error Correction*. In Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task (CoNLL-2013 Shared Task), pp. 1–12.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. *The CoNLL-2014 Shared Task on Grammatical Error Correction*. In Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task, pp. 1–14.

Diane Nicholls. 2003. *The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT*. In Proceedings of the Corpus Linguistics 2003 conference, pp. 572–581.

Robert Östling and Ola Knutsson. 2009. *A corpus-based tool for helping writers with Swedish collocations.* In Proceedings of the Workshop on Extracting and Using Constructions in NLP, pp. 28–33.

Alla Rozovskaya and Dan Roth. 2011. *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 - Volume 1, pp. 924–933.

Alla Rozovskaya, Kai-Wei Chang, Mark Sammons, Dan Roth, and Nizar Habash. 2014. *Correcting Grammatical Verb Errors*. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pp. 358–367.

Terry Santos. 1988. *Professors' reaction to the academic writing of nonnative speaking students*. TESOL Quarterly, 22(1):69–90.

Yu Sawai, Mamoru Komachi, and Yuji Matsumoto. 2013. *A Learner Corpus-based Approach to Verb Suggestion for ESL.* In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pp. 708–713.

Chi-Chiang Shei and Helen Pain. 2000. *An ESL Writer's Collocation Aid.* Computer Assisted Language Learning, 13(2), pp. 167–182.

David Wible, Chin-Hwa Kuo, Nai-Lung Tsao, Anne Liu and H.-L. Lin. 2003. *Bootstrapping in a language-learning environment*. Journal of Computer Assisted Learning, 19(4), pp. 90–102.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. *A New Dataset and Method for Automatically Grading ESOL Texts*. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 180–189.