‘Calling on the classical phone’: a distributional model of adjective-noun errors in learners’ English

Aurélie Herbelot
Centre for Mind/Brain Sciences
University of Trento
aurelie.herbelot@unitn.it

Ekaterina Kochmar
ALTA Institute
University of Cambridge
ek358@cl.cam.ac.uk

Abstract

In this paper we discuss three key points related to error detection (ED) in learners’ English. We focus on content word ED as one of the most challenging tasks in this area, illustrating our claims on adjective–noun (AN) combinations. In particular, we (1) investigate the role of context in accurately capturing semantic anomalies and implement a system based on distributional topic coherence, which achieves state-of-the-art accuracy on a standard test set; (2) thoroughly investigate our system’s performance across individual adjective classes, concluding that a class-dependent approach is beneficial to the task; (3) discuss the data size bottleneck in this area, and highlight the challenges of automatic error generation for content words.

1 Introduction

Error detection (ED) in the prose of ‘English as a Second Language’ (ESL) learners has recently attracted much attention (Ng et al., 2014; Ng et al., 2013; Dale et al., 2012; Dale and Kilgarriff, 2011). Earlier work on ED in ESL writing mostly focused on grammatical errors and errors in function words (Felice and Pulman, 2008; Gamon et al., 2008; Tetreault et al., 2010; Gamon, 2010; Rozovskaya and Roth, 2010a; Dahlmeier and Ng, 2011b; Ng et al., 2013). Lately, the focus has shifted to other error types, with the recent shared tasks encompassing all errors (Ng et al., 2014; Daudaravicius et al., 2016). In Ng et al. (2014), errors in content words are reported to be the second most frequent error type among 28 categories, accounting for 11.8% of all errors in the training and for about 14% in the test data, yet most teams scored poorly in this category suggesting that this is a challenging and mostly unsolved problem.

Current ED approaches can be broadly described as either *modular*, addressing one error type in particular, or as *comprehensive*, spanning all error types, as in case of the SMT-based techniques (Felice et al., 2014; Junczys-Dowmunt and Grundkiewicz, 2014). The modular approaches rely on the systematic and recurrent nature of the error patterns, and on the availability of closed confusion sets which enable casting the task as a multi-class classification problem. Since content words do not assume a finite set of confusions, it has been shown that ED for these combinations cannot be performed in a similar way (Kochmar and Briscoe, 2014; Rozovskaya et al., 2014). State-of-the-art SMT-based approaches also struggle with content word errors. We argue that ED systems for words which carry lexical meaning should necessarily involve a semantic component, which is typically not needed for other error types.

From a pedagogical point of view, detecting content word errors is an important task. Since content words carry the semantics of a sentence as well as the communicative intent of the writer, incorrect uses may lead to misunderstandings and meaning distortions: for example, *classic* and *classical* are frequently confused by language learners, yet *classic dance* and *classical dance* clearly have different denotations. The importance of content word knowledge in language learning has been demonstrated in the previous ESL studies: James (1998) points out that language learning itself is sometimes equated with mastering vocabulary. Leacock et al. (2014) mention an experiment in which teachers of English were asked to rank errors according to their gravity, and word choice errors were ranked as one of the two most serious

error categories. The study of Leacock and Chodorow (2003) also demonstrates that errors in content words have a direct impact on overall results in the *Test of English as a Foreign Language* (*TOEFL*).

In this paper, we focus on the underexplored task of content word error *detection*, independently of correction (see §2). We follow the semantically motivated approach outlined by Kochmar and Briscoe (2014) (henceforth K&B) for adjective–noun (AN) combinations,[1] building on their work by integrating context information in the classification. That is, we want to learn that although *classical dance* is more frequent than *classic dance*, the latter is correct in a context such as *They performed a classic Scottish dance*. In §3, we propose that features based on distributional topic coherence (Newman et al., 2010) can catch semantically anomalous ANs by modelling the effects of errors on the coherence of their context. A simple system based on this idea obtains state-of-the-art results. In §4, we show that the combination of the proposed in-context (*IC*) system with the out-of-context (*OOC*) ED system of K&B can further improve results, as long as the *OOC* system's error recall is sufficient. A thorough investigation of our system reveals that its performance is dependent on the adjective classes (see §5). This leads us to the conclusion that content word errors should be treated in a class-dependent way.

Finally, we show that availability of high-quality learner data for training the ED algorithms is of paramount importance. We note that certain error types, having recurrent error patterns, allow for straightforward artificial error data generation. However, we experimentally show that quality artificial data cannot be so easily generated for content words (see §6).

2 Related work

2.1 ED in content words

Previous approaches to error detection and correction of content words fall into two paradigms. One focuses on correction only, assuming that errors are detected by a separate ED algorithm (Liu et al., 2009; Dahlmeier and Ng, 2011a). The second performs error detection and correction through a single algorithm (Chang et al., 2008; Futagi et al., 2008; Park et al., 2008; Yi et al., 2008). The latter type relies on comparison of the learner's choice with possible alternatives: if any alternative scores higher than the original according to the chosen frequency-based metric, the original combination is flagged as an error and the alternative is suggested as a correction (Leacock et al., 2014). Approaches based on this idea have a number of weaknesses. In particular, they rely on the availability of a set of plausible alternatives and are unable to detect errors in the absence of such alternatives, even though a number of studies (Leacock et al., 2009; Chodorow et al., 2010; Andersen et al., 2013) have shown that ED alone (without correction) is useful for language learners. Crucially, K&B have also shown that some original word combinations can be felicitous even when some alternatives score higher, leading to over-corrections which can be hugely detrimental to ESL learners. Such considerations speak for considering ED as a separate task.

Prior proposals have rarely analysed content word errors from a semantic perspective. K&B have focused on ED for AN phrases and shown that approaches aimed at detecting semantic deviance can also identify errors in content words. These authors cast ED as a binary classification task and train a machine learning classifier using features derived from compositional distributional semantics. They obtain 81% accuracy, and show that their semantically motivated approach outperforms the state-of-the-art in ED so far. One drawback of their method is that they do not take context into account: features are based on the distributions of an AN's components and their composition, but not on the particular context where it is used. When evaluating their system in context by comparing the system's predictions with human, context-sensitive annotation, the authors note that accuracy drops to 65%. Our approach takes both semantic aspect *and* surrounding context into account.

2.2 Learner data

Since the standard approaches to ED rely on machine learning, availability of learner data is of paramount importance. Some error types allow for straightforward generalisation from seen examples (e.g., errors in function words or mechanical errors), but errors in content words appear to be less systematic. Therefore, it is crucial to have sufficient, thoroughly annotated learner data. To the best of our knowledge,

[1] We also believe that our approach can ultimately be applied to other types of content word combinations.

the AN dataset released by K&B[2] is the only publicly available dataset of learners' errors in content words that satisfies quality requirements. ANs are labelled with error type (semantically-related or form-related confusion, or no relation) and possible corrections are suggested. The data contains a two-level annotation, with the ANs being labelled as correct or incorrect out of their context (*OOC*), as well as in the original context of use (*IC*). For example, *classic dance* is annotated as correct *OOC*, but incorrect *IC* whenever it is used erroneously in place of *classical dance*. The dataset contains 798 ANs that are extracted from the *Cambridge Learner Corpus* (*CLC*)[3] and are unattested in the *British National Corpus* (*BNC*). This data is interesting for a linguistically-motivated investigation of learners' errors: K&B demonstrated that approaches that simply rely on frequency and collocational strength do not perform well on it.

The dataset contains 892 unique contexts, and in our experiments, we use a subset of 824 contexts (see §3). The lower bound for *IC*-annotated ANs is estimated as the majority class baseline and equals 0.55. The upper bound estimated as the average inter-annotator agreement is 0.74.

3 Topic coherence for error detection

3.1 Topic coherence

Topic coherence is a measure of the semantic relatedness of the items in a given set of words, which has mostly been studied from the perspective of 'topic modelling' techniques (Steyvers and Griffiths, 2007). Topic modelling is a text classification method which generates so-called topics from a corpus by analysing word co-occurrences, and subsequently models any new text in terms of those topics. A topic is expressed as a list of keywords supposed to be highly characteristic for a subject: for instance, {*film, actor, cinema, Hollywood*} might be the main keywords for a *film* topic. In order to obtain an intrinsic evaluation of such models, recent work has started investigating whether topics produced by standard techniques can be said to be 'coherent', i.e. whether topic keywords belong together, from a human point of view (Chang et al., 2009; AlSumait et al., 2009; Newman et al., 2010; Mimno et al., 2011).

Even though they stem from research on topic modelling, topic coherence measures can be applied to any set of words, and might for instance tell that the set {*chair, table, office, team*} is more coherent than {*chair, cold, elephant, crime*}. They are well suited to model semantic association and we hypothesise that they can tell us something about the semantic validity of a sentence.

3.2 Experimental setting

Following Newman et al. (2010), we define the coherence $\mathcal{COH}$ of a set of words $w_1...w_n$ as the mean of their pairwise similarities:

$$\mathcal{COH}(w_{1...n}) = mean\{Sim(w_i, w_j), ij \in 1...n, i < j\}$$

We estimate similarity as the cosine distance between two words in a distributional space (Turney and Pantel, 2010). In that setup, the meaning of a word is a vector that lives in a space where dimensions correspond to linguistic contexts. The vector's components reflect how characteristic a context is for the word under consideration.

Our hypothesis is that some lexical errors might result in a sharp variation of semantic coherence. Consider an example from learners' data:

(1) ... it was very difficult for my friends to call me with the *classical* *phone*...

The adjective *classical* is distributionally associated with the arts, collocating with nouns like *dance, music, style* or *literature*. Its similarity to *friend, call* or *phone* is much lower than the pairwise similarities of those words alone. We hypothesise that the inclusion of the unrelated *classical* in the sentence would thus have an adverse effect on its overall coherence.

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

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

Context size	1	2	3	4	5
SVM (low)	0.59 (± 0.02)	0.58 (± 0.02)	0.58 (± 0.01)	0.58 (± 0.02)	0.58 (± 0.02)
SVM (high)	0.59 (± 0.02)	0.59 (± 0.03)	0.58 (± 0.02)	0.59 (± 0.02)	0.59 (± 0.02)
+ adj. (low)	0.62 (± 0.04)	0.62 (± 0.04)	0.62 (± 0.03)	0.62 (± 0.02)	0.62 (± 0.04)
+ adj. (high)	0.64 (± 0.04)	**0.66** (± 0.06)	0.64 (± 0.04)	0.63 (± 0.02)	0.64 (± 0.02)

Table 1: SVM classification accuracy over different context sizes with three $\mathcal{COH}$ features, and with added *adjective* feature. The 'low/high' scores are the lowest/highest across all values of the C parameter.

Our learners' data consists of the AN dataset (see §2), spell-checked to correct orthographic errors. We build a distributional semantics space from the BNC,[4] using lemmatised word windows as context (size=10), the top 2000 most frequent content words as dimensions, and Positive Pointwise Mutual Information (PPMI) as weighting measure.[5] For each instance in the learners' data, we define a context window W as the AN under consideration and n words on each side of that AN. Allowable context words are nouns, verbs, adjectives and adverbs for which a BNC distribution is available. When the AN contains a word not found in the BNC, we discard the corresponding instances, ending up with 824 items out of the original 892 in the AN dataset. We are interested in three measures:

- the topic coherence $\mathcal{COH}$ of context W;
- the topic coherence $\mathcal{COH}_{\text{-}adj}$ of the context without the *adjective*;
- the topic coherence $\mathcal{COH}_{\text{-}noun}$ of the context without the *noun*.

Our starting hypothesis is that when the AN is erroneous, omitting either the adjective or the noun in the calculation results in a significant variation of the original coherence score.

3.3 Topic coherence results

We perform SVM classification with 5-fold cross-validation, using the coherence figures $\mathcal{COH}$, $\mathcal{COH}_{\text{-}adj}$ and $\mathcal{COH}_{\text{-}noun}$ as features. The order of the data is randomised before creating the folds, and the ratio of correct/incorrect instances kept equal between folds (55% correct to 45% incorrect). We use *SVM*^{light} (Joachims, 1999) with an RBF kernel to classify the data. We tune penalty parameter C, experimenting with the values of C in the range 10-200. Since the size of the AN dataset does not allow for the use of a development set, we report the lowest and highest system performance across all folds.

Table 1 shows that our simple 3-feature system reaches 59% accuracy. We attempt to improve on this by specifying which adjective occurs in the AN: we add 61 binary features to the SVM, corresponding to the 61 different adjectives in the data, and 'turn on' the feature matching the adjective under consideration for each data point. This step results in a further increase in accuracy, reaching 66%, which is on a par with the result reported by K&B when taking the *OOC* annotation to an in-context setting. This result is highly encouraging since our system is overall much simpler: given an available distributional semantic space, coherence values can be computed very straightforwardly and the SVM classifier relies on few features. We also note that the system is stable across various values of C: the differences between lowest and highest scores are not significant given the variability observed across all 5 folds. In the rest of this paper, we only report our highest scores under the understanding that varying C does not significantly affect results.

The best accuracy is obtained for a context size of 2, but the differences in performance between various context sizes are not statistically significant either. Most likely, the ideal context window for the task depends on the sentence. In some cases, larger context is actually harmful to ED, as in Ex. 2 below, where the context words are mostly about shopping/buying and do not have a straightforward association with either *cat* or *funny*. In contrast, Ex. 3 needs a larger window to catch that the sentence is not about different *bears* drinking but rather about a restaurant with *beers*.

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

[5] The space was built using the DISSECT toolkit (Dinu et al., 2013), available at http://clic.cimec.unitn.it/composes/toolkit/index.html.

	System description	Classifier	Accuracy
COH	§3.3	SVM	0.66
+ COMPDIST	feat. combination (§4.1)	SVM	0.68
+ COMPDIST	pipeline (§4.2)	SVM	0.67
+ *OOC* gold	§4.2	SVM	0.76
COMPDIST	K&B	DT	0.64
+ COH	feat. combination (§4.1)	DT	0.66

Table 2: Classifier combination accuracies

(2) I went shopping yesterday, and I've bought a new shirt. I had to buy it because it had a *funny cat* on it. It was quite cheap, it costs just £4.

(3) In the second one you can eat some easy food as salads, but you also can drink a great number of *different bears*.

An analysis of our results shows that the classifier is well-balanced, achieving 0.66 and 0.65 precision for correct and incorrect instances, as well as 0.65 and 0.66 recall. This is an improvement over the context-insensitive system from K&B, which scored much better recall on correct than incorrect instances (72% vs 58%).

4 Combining classifiers

K&B showed that it is possible to classify *OOC*-annotated content word errors with high accuracy: the authors reported an accuracy of 81% on this task. This means that regardless of context, we can learn that e.g. **big quantity* is incorrect. In the next set of experiments, we investigate the benefits of including context-insensitive classifier in our system: since the ANs that are annotated as errors *OOC* are also errors *IC*, we would expect the system to benefit overall from context-insensitive annotation.

We consider combinations of the following two systems and their respective semantic information:

- COMPDIST is the *context-insensitive* system of K&B, which we ran on our subset of 824 contexts. The AN vectors are built using the *multiplicative* model of semantic composition (Mitchell and Lapata, 2008). A set of measures that can distinguish between the representations of the correct and incorrect ANs is applied, including, for example, *vector length*, *cosine similarity* between the AN vector and the input noun, and so forth.[6] The values of these measures are then used by an ED algorithm running over a *Decision Tree* (DT) classifier.[7]

- COH is the *context-sensitive* system presented in this paper, with three coherence-based features and a feature representing the adjective in the AN (see §3.3).

We examine two architectures: in a first experiment, we simply concatenate the features from COMPDIST and COH and input the resulting vector into (a) an SVM classifier, as used in this paper; (b) a DT classifier, as used in K&B. In a second experiment, we design a pipeline system, where the classification of COMPDIST is fed into the topic coherence model. All results reported in this section are for a context size of 2. A summary can be found in Table 2.

4.1 Direct feature combination

We first run COMPDIST in isolation over the *IC* annotation, to get a baseline accuracy. This results in a performance of 64%, just below our COH system accuracy of 66% (see Table 2).

We then proceed with feature concatenation, starting with the full feature set and then applying ablation tests to identify the best-performing features. The features are fed into the DT classifier of K&B on one hand (line COMPDIST+COH in Table 2) and our SVM classifier on the other hand (line COH+COMPDIST).

[6] For the full feature set, please consult K&B.

[7] We use the *NLTK* implementation (Bird et al., 2009).

1 usual (4/4)	.83 strong (15/18)	.63 fast (7/11)	.41 historical (5/12)
1 rapid (1/1)	.83 clear (5/6)	.62 small (15/24)	.40 economic (2/5)
1 magic (1/1)	.80 actual (4/5)	.62 nice (39/62)	.33 deep (1/3)
1 incorrect (3/3)	.75 bad (24/32)	.62 important (18/29)	.30 whole (3/10)
1 elder (16/16)	.72 good (39/54)	.60 unique (6/10)	.25 heavy (2/8)
1 economical (33/33)	.71 hard (5/7)	.60 high (3/5)	.20 true (1/5)
1 classical (5/5)	.70 main (7/10)	.60 electric (3/5)	.18 certain (2/11)
1 classic (3/3)	.70 different (29/41)	.60 correct (3/5)	.14 precious (1/7)
.90 funny (10/11)	.69 best (36/52)	.57 near (4/7)	.14 particular (1/7)
.89 suitable (8/9)	.68 typical (11/16)	.53 wrong (7/13)	.14 ancient (1/7)
.89 soft (8/9)	.67 big (63/94)	.50 short (6/12)	0 far (0/2)
.89 full (8/9)	.66 various (6/9)	.50 present (2/4)	0 false (0/2)
.89 convenient (8/9)	.64 proper (9/14)	.47 common (8/17)	0 electrical (0/3)
.87 large (7/8)	.63 great (26/41)	.42 appropriate (3/7)	

Table 3: Per-adjective precision values for SVM classification, sorted from highest to lowest

Using the DT classifier, the best accuracy of the direct combination of features is 66% with the feature set including *cosine similarity* to the input noun, *ranked density*, *adjective* and a *coherence-based* feature based on $(\mathcal{COH} - \mathcal{COH}_{\text{-}adj})$. The improvement over baseline is however not statistically significant.

Adding the COMPDIST features to the SVM COH system results in a similar improvement, reaching 68% from a 66% baseline, using the *cosine similarity* to the noun, and semantic neighbourhood features.

4.2 Pipeline system

To test the actual effect of the topic coherence features at in-context classification stage, we first attempt to add the *OOC* gold annotation to our system, in the form of a new feature. The baseline created by the *OOC* gold annotation is very high: running the classifier over that one feature results in 73% accuracy. Nevertheless, performance increases when the gold annotation is combined with COH. The best result is 76% – a 3% improvement (statistically significant at $p = 0.03$). This is over the human upper bound of 74%, and it shows that the topic coherence features perform well in contextualising the *OOC* annotation.

However, since a 'real-world' pipeline system does not have access to the gold annotations, we replace the gold annotations with the output of COMPDIST. In this setting, the combination only produces minimal improvement, reaching 67% accuracy with the SVM and 66% with the DT classifier. The reason for this result is low error recall of the *OOC* system which is tuned towards high precision because of the strong negative impact on the learners when wrongly reporting an error. While this is sensible from an educational point of view, it means that we are only recalling 17% of erroneous ANs at the *OOC* stage. We conclude that improving *OOC* detection can hugely benefit the overall system.

5 Adjective-dependent classification analysis

5.1 COH analysis

Our experiments with the coherence-based system showed that it is particularly accurate in classifying form-related errors: the accuracy on *classic*, *classical*, *economical*, *elder*, *electric*, *electrical* and *historical* – which are responsible for 80% of the form-related errors in our data – is 77%. Otherwise, the accuracy of the system is generally dependent on the adjective being classified. Table 3 shows precision values for each adjective in our data.[8] While *economical* (33 instances) and *funny* (11 instances) achieve 100% and 90% precision respectively, *certain* (11 instances) and *ancient* (7 instances) only reach 18% and 14%. Roughly speaking, adjectives expressing a sentiment (*funny*, *suitable*, *convenient*, *bad*, *good*, *best*, *great*, *nice*) are to be found at the top of the table, while no such consistency is to be found for the quantity adjectives: *large*, *big*, *small*, *high*, *deep*, *short* and *heavy* span a whole range of precision values, from 87% down to 25%. We conclude that the adjectives in our dataset can behave very differently with respect to the types of errors they attract, and a single classifier may not be able to model all cases equally well. In the next set of experiments, we thoroughly investigate our system's performance

[8] Morphologically related forms, for example *big* and *biggest*, are collapsed together.

Adjective	Best training elements	Accuracy
appropriate	{nice, good, best, different, bad, short, fast}	71.43%
bad	{unique}	78.12%
best	{nice, good, different, fast, funny, unique}	71.70%
big	{proper}	68.09%
correct	{nice, good, best, different, bad, short, fast, unique}	80.00%
economic	{strong, typical, elder, certain}	80.00%
economical	{small, strong, typical, elder, proper, certain}	100.00%
elder	{economical, small, strong, typical, proper, certain}	100.00%
funny	{big}	90.91%
good	{nice, best, different, fast}	70.91%
great	{wrong, main}	69.05%
nice	{good, best, different, fast}	67.74%
precious	{funny}	71.43%
small	{big, proper, funny}	68.00%

Table 4: Best training elements for a subset of adjectives, together with accuracy

across individual adjective classes. Since we lack data for a separate development set, these experiments present the analysis of the data rather than actual classification results but we believe that these results can inform future studies.

5.2 Modelling the AN data

We hypothesise that some adjectives behave in a similar way with respect to their interaction with topic coherence, and may be classifiable under a joint category. Since there are no obvious confusion sets for content words to guide category formation (see §1), we attempt to model the AN set in a purely data-driven way. We first train a classifier over each adjective with frequency ≥ 10, thus obtaining 27 individual classifiers. We then apply each classifier to every single other adjective and record which one(s) perform(s) best for that item. For example, we verify how well *ancient* is classified by each of the 26 models and record the classifier trained on *unique* as the one performing best. We take this as evidence that *ancient* and *unique* share some properties with respect to the task.

Table 4 shows accuracy for some adjectives, with the best recorded training set(s). The overall accuracy, averaged over all adjectives, is 75%. This result is on a par with human performance estimated at 74% (see §2). Per-class precision is 80% on errors and 73% on correct instances, while recall is 59% for errors and 88% for correct ANs.

We note two trends: (a) adjectives of judgement (*appropriate*, *bad*, *correct*, *nice*, *precious*) tend to be best trained by other judgement adjectives (*best*, *good*, *nice*); (b) adjectives for which form-related errors are frequent (*classic*/*classical*, *economic*/*economical*, *elder*) tend to get their best accuracy when trained on the same set (*strong*, *typical*, *elder*, *certain*).

These results suggest that training adjective classes separately could have a very positive impact. For instance, let's consider the set {*nice*, *good*, *best*, *bad*, *convenient*, *suitable*, *appropriate*}. Training each adjective over the other members of the set results in an increase in performance for those adjectives: e.g., accuracy for *nice* increases by 5 points, for *appropriate* by 14 points. Similarly, training {*small*, *big*, *large*} as a set gives a 5-point improvement on our best results.

However, due to the relatively small size of the dataset, it is impossible to have a development phase to choose the best training sets, and a separate test phase to verify robustness: confirming that *bad* is indeed best trained on *unique* would overall require more data than the 32 and 10 instances currently available.

5.3 Adjective-specific system combination

We have shown that COH performs differently on adjective-specific subsets and we have assumed a complex interaction between topic coherence and error patterns specific for particular adjectives. Our tests using COMPDIST and COH with the DT classifier confirm that the performance of the two systems on the adjective-specific subsets in the data is also different: for example, COH performs well on the adjectives *large*, *bad* and *good* (see Table 3), while COMPDIST achieves better results on *short* and *heavy*. In the next experiment, we combine the two systems in an integrated adjective-specific ED algorithm.

We implement an oracle system via a voting step based on the COMPDIST and COH adjective-specific performance. That is, for each test item, we use the prediction of the system which has best accuracy for the particular adjective under consideration. This oracle system achieves an accuracy of 71% which compares favourably to the results of COMPDIST (64%), COH (66%) alone, as well as the direct combination of the features used by the two systems (68%). Although this result is an upper bound since we lack data to set up a separate development set, we can reliably make two observations. First, we confirm that using adjective-specific information in ED improves the algorithm performance. Second, we note that the voting system's upper bound is comparable to human performance, indicating that the combination of a strong *OOC* baseline and a relatively simple semantic model of context provides the necessary conditions for ideal results: the combined system covers all relevant information for the task, and training on an expanded dataset can be expected to drastically improve performance.

6 Generating errors

Earlier we noted that high-quality learner data is crucial for ED and that, due to the size of the K&B dataset, we could not verify the results of our experiments on a separate development set (§3 and §5). Since annotated content word error data is expensive and time-consuming to produce, in this final set of experiments we attempt to generate more data in an automated way. Artificial error generation has previously been demonstrated to be useful for function word ED (Foster and Andersen, 2009; Rozovskaya and Roth, 2010b; Felice and Yuan, 2014). Following this line of work, we attempt to produce data by random substitution of adjectives.

For each adjective, we extract new data from a section of *ukWaC* (Baroni et al., 2009) totalling 1M tokens, POS-tagged and parsed with the RASP parser (Briscoe et al., 2006). We collect word windows containing the adjective under consideration, using the pattern `[word`$_{-2}$`] [word`$_{-1}$`] [ADJ] [noun] [word`$_{+1}$`] [word`$_{+2}$`]`, thus using a 2-word context around the AN. Next, we randomly shuffle the adjectives and their contexts so that for a particular context window W_k associated with an adjective a_k, we replace a_k with a_m – an adjective linked to another context window – assuming that in most cases, such substitutions will produce incorrect instances. Then, we collect all incorrect instances for a given adjective and concatenate them with an equal number of correct uses, giving us a balanced training set for that adjective. The size of each training set is dependent on the overall frequency of the adjective, ranging from 20 to 2600 instances. Around half of the adjectives have a training set with over 1000 instances, the vast majority (93%) have at least 100 training examples.

Training and testing on this data unfortunately does not produce the expected improvements, with the accuracy falling to 56%. We conclude that the nature of the training data is vital to the performance of the system: in complex tasks like content word ED, automatically generated examples are no substitute for real human errors, and the subtle semantic phenomena occurring in learners' writing cannot be easily reproduced. The absence of clear confusion sets for content word errors makes the task of error generation particularly arduous. The erroneous *calling on the classical phone* is a case in point, showing that a wrong use of *classical* does not necessarily derive from a confusion with *classic*: the speaker probably meant *landline*. Such cases show that the diversity of content words errors makes artificial error generation less viable than for function words, and illustrates the value of 'real', annotated learner data.

7 Conclusion

We have investigated the linguistic felicity of AN phrases through the lens of distributional topic coherence and conclude by showing how this work can inform future research on content word ED.

First, we showed that using topic coherence features to model context leads to accuracy figures that are competitive with previously reported results (K&B). Framing ED in terms of coherence is linguistically sensible and computationally efficient. There are benefits in using *OOC* and *IC* systems in a pipeline architecture, but this relies on the *OOC* system having good enough recall.

Second, we found that per-adjective classification could in principle approach human-like performance. However, proper training and evaluation requires a larger dataset than is currently available.

Thirdly, we investigated the automatic creation of training data. Our experiments demonstrated that real learners' data cannot be easily substituted: in contrast with function words, content words are not naturally associated with clear confusion sets which might guide data generation.

In further work, we would like to pursue our investigation of the linguistic factors that govern certain types of errors. One interesting avenue would be to research the influence of the learner's L1 language on the observed semantic mistakes: we could imagine, for instance, that some systematicity could be captured in the effects of polysemy across languages (e.g. *heavy* might be more—or at least differently—polysemous in English compared to the learner's L1).

We will also concentrate on the expansion of the available dataset in a controlled fashion, ensuring that enough data is supplied for individual adjective training and testing. Whilst our results on error generation indicate that automatic methods may not be suited to the task, more sophisticated procedures could be tried out. For instance, it is conceivable that a distributional analysis of a (non-annotated) learners' corpus would highlight certain systematic errors which would be replicable on a larger scale. With more training data available, another interesting avenue would be to further explore adjective-dependent classification approaches and adjective category formation.

Acknowledgments

We are grateful to the COLING reviewers for their helpful feedback, and for their interesting suggestions for further work. Ekaterina Kochmar's research is supported by Cambridge English Language Assessment via the ALTA Institute. Aurélie Herbelot's contribution to this paper was similarly supported by ALTA.

References

Loulwah AlSumait, Daniel Barbará, James Gentle, and Carlotta Domeniconi. 2009. Topic significance ranking of LDA generative models. In *Machine Learning and Knowledge Discovery in Databases*, pages 67–82. Springer.

Øistein E. Andersen, Helen Yannakoudakis, Fiona Barker, and Tim Parish. 2013. Developing and testing a self-assessment and tutoring system. In *Proceedings of the BEA-2013*, pages 32–41.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit.* O'Reilly Media.

Ted Briscoe, John Carroll, and Rebecca Watson. 2006. The second release of the RASP system. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 77–80.

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):283–299.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-graber, and David M Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, pages 288–296.

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

Daniel Dahlmeier and Hwee Tou Ng. 2011b. Grammatical error correction with alternating structure optimization. In *Proceedings of the ACL-HLT 2011*, pages 915–923.

Robert Dale and Adam Kilgarriff. 2011. Helping Our Own: The HOO 2011 Pilot Shared Task. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 242–249.

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 BEA-2012*, pages 54–62.

Vidas Daudaravicius, Rafael E. Banchs, Elena Volodina, and Courtney Napoles. 2016. A Report on the Automatic Evaluation of Scientific Writing Shared Task. In *Proceedings of the BEA-2016*.

Georgiana Dinu, Nghia The Pham, and Marco Baroni. 2013. General estimation and evaluation of compositional distributional semantic models. In *Workshop on Continuous Vector Space Models and their Compositionality*, pages 50–58, Sofia, Bulgaria.

Rachele De Felice and Stephen G. Pulman. 2008. A classifier-based approach to preposition and determiner error correction in L2 English. In *Proceedings of the COLING-2008*, pages 169–176.

Mariano Felice and Zheng Yuan. 2014. Generating artificial errors for grammatical error correction. In *Proceedings of the Student Research Workshop at the ACL-2014*, pages 116–126.

Mariano Felice, Zheng Yuan, Øistein E. Andersen, Helen Yannakoudakis, and Ekaterina Kochmar. 2014. Grammatical error correction using hybrid systems and type filtering. In *Proceedings of the CoNLL 2014: Shared Task*.

Jennifer Foster and Øistein E Andersen. 2009. Generrate: generating errors for use in grammatical error detection. In *Proceedings of the BEA-2009*, pages 82–90.

Yoko Futagi, Paul Deane, Martin Chodorow, and Joel Tetreault. 2008. A computational approach to detecting collocation errors in the writing of non-native speakers of English. *Computer Assisted Language Learning*, 21(4):353–367.

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

Michael Gamon. 2010. Using mostly native data to correct errors in learners' writing. In *Proceedings of the NAACL-2010*, pages 163–171.

Carl James. 1998. *Errors in Language Learning and Use: Exploring Error Analysis.* London: Longman.

Thorsten Joachims, 1999. *Making Large-Scale SVM Learning Practical.* Advances in Kernel Methods - Support Vector Learning, B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2014. The AMU system in the CoNLL-2014 shared task: Grammatical error correction by data-intensive and feature-rich statistical machine translation. In *Proceedings of the CoNLL-2014: Shared Task*.

Ekaterina Kochmar and Ted Briscoe. 2014. Detecting learner errors in the choice of content words using compositional distributional semantics. In *Proceedings of the COLING-2014*, pages 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*, pages 195–207. Mahwah, NJ: Lawrence Erlbaum Associates.

Claudia Leacock, Michael Gamon, and Chris Brockett. 2009. User Input and Interactions on Microsoft Research ESL Assistant. In *Proceedings of the BEA-2009*, pages 73–81.

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 and Claypool Publishers.

Anne Li-E Liu, David Wible, and Nai-Lung Tsao. 2009. Automated suggestions for miscollocations. In *Proceedings of the BEA-2009*, pages 47–50.

David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the EMNLP-2011*, pages 262–272.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of ACL-HLT 2008*, pages 236–244.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Proceedings of the NAACL-HLT 2010*, pages 100–108.

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 CoNLL-2013: Shared Task*, pages 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 CoNLL-2014: Shared Task*, pages 1–14.

Taehyun Park, Edward Lank, Pascal Poupart, and Michael Terry. 2008. Is the sky pure today? AwkChecker: an assistive tool for detecting and correcting collocation errors. In *Proceedings of the 21st annual ACM symposium on User interface software and technology*, pages 121–130.

Alla Rozovskaya and Dan Roth. 2010a. Annotating ESL errors: Challenges and rewards. In *Proceedings of the BEA-2010*, pages 28–36.

Alla Rozovskaya and Dan Roth. 2010b. Training paradigms for correcting errors in grammar and usage. In *Proceedings of the NAACL-HLT 2010*.

Alla Rozovskaya, Dan Roth, and Vivek Srikumar. 2014. Correcting Grammatical Verb Errors. In *Proceedings of EACL-2014*, pages 358–367.

Mark Steyvers and Tom Griffiths. 2007. Probabilistic topic models. In T. Landauer, D. McNamara, S. Dennis, and W. Kintsch, editors, *Latent Semantic Analysis: A Road to Meaning*.

Joel Tetreault, Jennifer Foster, and Martin Chodorow. 2010. Using parse features for preposition selection and error detection. In *Proceedings of the ACL-2010: Short Papers*, pages 353–358.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Xing Yi, Jianfeng Gao, and William B. Dolan. 2008. A Web-based English Proofing System for English as a Second Language Users. In *Proceedings of the IJCNLP-2008*, pages 619–624.