Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pages 169 176
Manchester, August 2008

3 Related work

Although in the past there has been some research on determiner choice in L1 for applications such as generation and machine translation output, work to date on automatic error detection in L2 writing has been fairly limited. Izumi et al. (2004) train a maximum entropy classifier to recognise various errors using contextual features. They report results for different error types (e.g. omission - precision 75.7%, recall 45.67%; replacement - P 31.17%, R 8%), but there is no break-down of results by individual POS. Han et al. (2006) use a maximum entropy classifier to detect determiner errors, achieving 83% accuracy. Chodorow et al. (2007) present an approach to preposition error detection which also uses a model based on a maximum entropy classifier trained on a set of contextual features, together with a rule-based filter. They report 80% precision and 30% recall. Finally, Gamon et al. (2008) use a complex system including a decision tree and a language model for both preposition and determiner errors, while Yi et al. (2008) propose a web count-based system to correct determiner errors (P 62%, R 41%).

The work presented here displays some similarities to the papers mentioned above in its use of a maximum entropy classifier and a set of features. However, our feature set is more linguistically sophisticated in that it relies on a full syntactic analysis of the data. It includes some semantic components which we believe play a role in correct class assignment.

4 Contextual models for prepositions and determiners

4.1 Feature set

The approach proposed in this paper is based on the belief that although it is difficult to formulate hard and fast rules for correct preposition and determiner usage, there is enough underlying regularity of characteristic syntactic and semantic contexts to be able to predict usage to an acceptable degree of accuracy. We use a corpus of grammatically correct English to train a maximum entropy classifier on examples of correct usage. The classifier can therefore learn to associate a given preposition or determiner to particular contexts, and reliably predict a class when presented with a novel instance of a context for one or the other.

The L1 source we use is the British National Corpus (BNC) as we believe this offers a representative sample of different text types. We represent training and testing items as vectors of values for linguistically motivated contextual features. Our feature vectors include 18 feature categories for determiners and 13 for prepositions; the main ones are illustrated in Table 1 and Table 2 respectively. Further determiner features note whether the noun is modified by a predeterminer, possessive, numeral, and/or a relative clause, and whether it is part of a 'there is...' phrase. Additional preposition features refer to the grade of any adjectives or adverbs modified (base, comparative, superlative) and to whether the items modified are modified by more than one PP[1].

Head noun	'apple'
Number	singular
Noun type	count
Named entity?	no
WordNet category	food, plant
Prep modification?	yes, 'on'
Object of Prep?	no
Adj modification?	yes, 'juicy'
Adj grade	superlative
POS ± 3	VV, DT, JJS, IN, DT, NN

Table 1: Determiner feature set for *Pick **the** juiciest apple on the tree.*

POS modified	verb
Lexical item modified	'drive'
WordNet Category	motion
Subcat frame	pp_to
POS of object	noun
Object lexical item	'London'
Named entity?	yes, type = location
POS ± 3	NNP, VBD, NNP
Grammatical relation	iobj

Table 2: Preposition feature set for *John drove **to** London.*

In De Felice and Pulman (2007), we described some of the preprocessing required and offered some motivation for this approach. As for our choice of features, we aim to capture all the elements of a sentence which we believe to have an effect on preposition and determiner choice, and which can be easily extracted automatically - this is a key consideration as all the features derived rely on automatic processing of the text. Grammatical relations refer to RASP-style grammatical relations between heads and complements in which the preposition occurs (see e.g. (Briscoe et al.,

[1] A full discussion of each feature, including motivation for its inclusion and an assessment of its contribution to the model, is found in De Felice (forthcoming).

Author	Accuracy
Baseline	26.94%
Gamon et al. 08	64.93%
Chodorow et al. 07	69.00%
Our model	**70.06%**

Table 3: Classifier performance on L1 prepositions

2006)). Semantic word type information is taken from WordNet lexicographer classes, 40 broad semantic categories which all nouns and verbs in WordNet belong to[2] (e.g. 'verb of motion', 'noun denoting food'), while the POStags are from the Penn Treebank tagset - we note the POS of three words either side of the target word[3]. For each occurrence of a preposition or determiner in the corpus, we obtain a feature vector consisting of the preposition or determiner and its context, described in terms of the features noted above.

5 Acquiring the models

5.1 Prepositions

At the moment, we restrict our analysis to the nine most frequent prepositions in the data: ***at, by, for, from, in, of, on, to,*** and ***with***, to ensure a sufficient amount of data for training. This gives a training dataset comprising 8,898,359 instances. We use a standard maximum entropy classifier [4] and do not omit any features, although we plan to experiment with different feature combinations to determine if, and how, this would impact the classifier's performance. Before testing our model on learner data, it is important to ascertain that it can correctly associate prepositions to a given context in grammatical, well-edited data. We therefore tested the model on a section of the BNC not used in training, section J. Our best result to date is **70.06%** accuracy (test set size: 536,193). Table 3 relates our results to others reported in the literature on comparable tasks. The baseline refers to always choosing the most frequent option, namely ***of***.

We can see that our model's performance compares favourably to the best results in the literature, although direct comparisons are hard to draw since different groups train and test on different preposition sets and on different types of data (British vs. American English, BNC vs. news reports, and so on). Furthermore, it should be noted that Gamon et al. report more than one figure in their results, as there are two components to their model: one determining whether a preposition is needed, and the other deciding what the preposition should be. The figure reported here refers to the latter task, as it is the most similar to the one we are evaluating. Additionally, Chodorow et al. also discuss some modifications to their model which can increase accuracy; the result noted here is the one more directly comparable to our own approach.

[2]No word sense disambiguation was performed at this stage.

[3]In NPs with a null determiner, the target is the head noun.

[4]Developed by James Curran.

	Proportion of training data	Precision	Recall
of	27.83% (2,501,327)	74.28%	90.47%
to	20.64% (1,855,304)	85.99%	81.73%
in	17.68% (1,589,718)	60.15%	67.60%
for	8.01% (720,369)	55.47%	43.78%
on	6.54% (587,871)	58.52%	45.81%
with	6.03% (541,696)	58.13%	46.33%
at	4.72% (424,539)	57.44%	52.12%
by	4.69% (421,430)	63.83%	56.51%
from	3.86% (347,105)	59.20%	32.07%

Table 4: L1 results - individual prepositions

5.1.1 Further discussion

To fully assess the model's performance on the L1 data, it is important to consider factors such as performance on individual prepositions, the relationship between training dataset size and accuracy, and the kinds of errors made by the model.

Table 4 shows the classifier's performance on individual prepositions together with the size of their training datasets. At first glance, a clear correlation appears between the amount of data seen in training and precision and recall, as evidenced for example by ***of*** or ***to***, for which the classifier achieves a very high score. In other cases, however, the correlation is not so clear-cut. For example ***by*** has one of the smallest data sets in training but higher scores than many of the other prepositions, while ***for*** is notable for the opposite reason, namely having a large dataset but some of the lowest scores.

The absence of a definite relation between dataset size and performance suggests that there might be a cline of 'learnability' for these prepositions: different prepositions' contexts may be more or less uniquely identifiable, or they may have more or fewer senses, leading to less confusion for the classifier. One simple way of verifying the latter case is by looking at the number of senses assigned to the prepositions by a resource

Target prep	*Confused with*								
	at	**by**	**for**	**from**	**in**	**of**	**on**	**to**	**with**
at	xx	4.65%	10.82%	2.95%	36.83%	19.46%	9.17%	10.28%	5.85%
by	6.54%	xx	8.50%	2.58%	41.38%	19.44%	5.41%	10.04%	6.10%
for	8.19%	3.93%	xx	1.91%	25.67%	36.12%	5.60%	11.29%	7.28%
from	6.19%	4.14%	6.72%	xx	26.98%	26.74%	7.70%	16.45%	5.07%
in	7.16%	9.28%	10.68%	3.01%	xx	43.40%	10.92%	8.96%	6.59%
of	3.95%	2.00%	18.81%	3.36%	40.21%	xx	9.46%	14.77%	7.43%
on	5.49%	3.85%	8.66%	2.29%	32.88%	27.92%	xx	12.20%	6.71%
to	9.77%	3.82%	11.49%	3.71%	24.86%	27.95%	9.43%	xx	8.95%
with	3.66%	4.43%	12.06%	2.24%	28.08%	26.63%	6.81%	16.10%	xx

Table 5: Confusion matrix for L1 data - prepositions

such as the Oxford English Dictionary. However, we find no good correlation between the two as the preposition with the most senses is ***of*** (16), and that with the fewest is ***from*** (1), thus negating the idea that fewer senses make a preposition easier to learn. The reason may therefore be found elsewhere, e.g. in the lexical properties of the contexts.

A good picture of the model's errors can be had by looking at the confusion matrix in Table 5, which reports, for each preposition, what the classifier's incorrect decision was. Analysis of these errors may establish whether they are related to the dataset size issue noted above, or have a more linguistically grounded explanation.

From the table, the frequency effect appears evident: in almost every case, the three most frequent wrong choices are the three most frequent prepositions, ***to***, ***of***, and ***in***, although interestingly not in that order, ***in*** usually being the first choice. Conversely, the less frequent prepositions are less often suggested as the classifier's choice. This effect precludes the possibility at the moment of drawing any linguistic conclusions. These may only be gleaned by looking at the errors for the three more frequent prepositions. We see for example that there seems to be a strong relation between ***of*** and ***for***, the cause of which is not immediately clear: perhaps they both often occur within noun phrases (e.g. *book of recipes*, *book for recipes*). More predictable is the confusion between ***to*** and ***from***, and between locative prepositions such as ***to*** and ***at***, although the effect is less strong for other potentially confusable pairs such as ***in*** and ***at*** or ***on***.

Table 6 gives some examples of instances where the classifier's chosen preposition differs from that found in the original text. In most cases, the classifier's suggestion is also grammatically correct,

Classifier choice	**Correct phrase**
demands **of** the sector	demands **for**...
condition **for** development	condition **of**...
travel **to** speed	travel **at**...
look **at** the USA	look **to**...

Table 6: Examples of classifier errors on preposition L1 task

Author	**Accuracy**
Baseline	59.83%
Han et al. 06	83.00%
Gamon et al. 08	86.07%
Turner and Charniak 07	86.74%
Our model	**92.15%**

Table 7: Classifier performance - L1 determiners

but the overall meaning of the phrases changes somewhat. For example, while the *demands **of** the sector* are usually made by the sector itself, the *demands **for** the sector* suggest that someone else may be making them. These are subtle differences which it may be impossible to capture without a more sophisticated understanding of the wider context.

The example with *travel*, on the other hand, yields an ungrammatical result. We assume that the classifier has acquired a very strong link between the lexical item *travel* and the preposition ***to*** that directs it towards this choice (cf. also the example of *look at/to*). This suggests that individual lexical items play an important role in preposition choice along with other more general syntactic and semantic properties of the context.

	%of training data	Prec.	Recall
a	9.61% (388,476)	70.52%	53.50%
the	29.19% (1,180,435)	85.17%	91.51%
null	61.20% (2,475,014)	98.63%	98.79%

Table 8: L1 results - individual determiners

Target det	*Confused with*		
	a	**the**	**null**
a	xx	92.92%	7.08%
the	80.66%	xx	19.34%
null	14.51%	85.49%	xx

Table 9: Confusion matrix for L1 determiners

5.2 Determiners

For the determiner task, we also consider only the three most frequent cases (***a, the, null***), which gives us a training dataset consisting of 4,043,925 instances. We achieve accuracy of 92.15% on the L1 data (test set size: 305,264), as shown in Table 7. Again, the baseline refers to the most frequent class, ***null***.

The best reported results to date on determiner selection are those in Turner and Charniak (2007). Our model outperforms their n-gram language model approach by over 5%. Since the two approaches are not tested on the same data this comparison is not conclusive, but we are optimistic that there is a real difference in accuracy since the type of texts used are not dissimilar. As in the case of the prepositions, it is interesting to see whether this high performance is equally distributed across the three classes; this information is reported in Table 8. Here we can see that there is a very strong correlation between amount of data seen in training and precision and recall. The indefinite article's lower 'learnability', and its lower frequency appears not to be peculiar to our data, as it is also found by Gamon et al. among others.

The disparity in training is a reflection of the distribution of determiners in the English language. Perhaps if this imbalance were addressed, the model would more confidently learn contexts of use for ***a***, too, which would be desirable in view of using this information for error correction. On the other hand, this would create a distorted representation of the composition of English, which may not be what we want in a statistical model of language. We plan to experiment with smaller scale, more similar datasets to ascertain whether the issue is one of training size or of inherent difficulty in learning about the indefinite article's occurrence.

In looking at the confusion matrix for determiners (Table 9), it is interesting to note that for the classifier's mistakes involving ***a*** or ***the***, the erroneous choice is in the almost always the other determiner rather than the null case. This suggests that the frequency effect is not so strong as to override any true linguistic information the model has acquired, otherwise the predominant choice would always be the null case. On the contrary, these results show that the model is indeed capable of distinguishing between contexts which require a determiner and those which do not, but requires further fine tuning to perform better in knowing which of the two determiner options to choose. Perhaps the introduction of a discourse dimension might assist in this respect. We plan to experiment with some simple heuristics: for example, given a sequence 'Determiner Noun', has the noun appeared in the preceding few sentences? If so, we might expect ***the*** to be the correct choice rather than ***a***.

6 Testing the model

6.1 Working with L2 text

To evaluate the model's performance on learner data, we use a subsection of the Cambridge Learner Corpus (CLC)[5]. We envisage our model to eventually be of assistance to learners in analysing their writing and identifying instances of preposition or determiner usage which do not correspond to what it has been trained to expect; the more probable instance would be suggested as a more appropriate alternative. In using NLP tools and techniques which have been developed with and for L1 language, a loss of performance on L2 data is to be expected. These methods usually expect grammatically well-formed input; learner text is often ungrammatical, misspelled, and different in content and structure from typical L1 resources such as the WSJ and the BNC.

6.2 Prepositions

For the preposition task, we extract 2523 instances of preposition use from the CLC (1282 correct, 1241 incorrect) and ask the classifier to mark them

[5]The CLC is a computerised database of contemporary written learner English (currently over 25m words). It was developed jointly by Cambridge ESOL and Cambridge University Press. The Cambridge Error Coding System has been developed and applied manually to the data by Cambridge University Press.

Instance type	Accuracy
Correct	66.7%
Incorrect	70%

Table 10: Accuracy on L2 data - prepositions. Accuracy on incorrect instances refers to the classifier successfully identifying the preposition in the text as not appropriate for that context.

as correct or incorrect. The results from this task are presented in Table 10. These first results suggest that the model is fairly robust: the accuracy rate on the correct data, for example, is not much lower than that on the L1 data. In an application designed to assist learners, it is important to aim to reduce the rate of false alarms - cases where the original is correct, but the model flags an error - to a minimum, so it is positive that this result is comparatively high. Accuracy on error identification is at first glance even more encouraging. However, if we look at the suggestions the model makes to replace the erroneous preposition, we find that these are correct only 51.5% of the time, greatly reducing its usefulness.

6.2.1 Further discussion

A first analysis of the classifier's decisions and its errors points to various factors which could be impairing its performance. Spelling mistakes in the input are one of the most immediate ones. For example, in the sentence *I'm Franch, responsable **on** the computer services*, the classifier is not able to suggest a correct alternative to the erroneous ***on***: since it does not recognise the adjective as a misspelling of *responsible*, it loses the information associated with this lexical feature, which could potentially determine the preposition choice.

A more complex problem arises when poor grammar in the input misleads the parser so that the information it gives for a sentence is incorrect, especially as regards PP attachment. In this example, *I wold like following equipment **to** my speech: computer, modem socket and microphone*, the missing *the* leads the parser to treat *following* as a verb, and believes it to be the verb to which the preposition is attached. It therefore suggests ***from*** as a correction, which is a reasonable choice given the frequency of phrases such as *to follow from*. However, this was not what the PP was meant to modify: impaired performance from the parser could be a significant negative factor in the model's performance. It would be interesting to test the model on texts written by students of different levels of proficiency, as their grammar may be more error-free and more likely to be parsed correctly. Alternatively, we could modify the parser so as to skip cases where it requires several attempts before producing a parse, as these more challenging cases could be indicative of very poorly structured sentences in which misused prepositions are dependent on more complex errors.

A different kind of problem impacting our accuracy scores derives from those instances where the classifier selects a preposition which can be correct in the given context, but is not the correct one in that particular case. In the example *I received a beautiful present **at** my birthday*, the classifier identifies the presence of the error, and suggests the grammatically and pragmatically appropriate correction ***for***. The corpus annotators, however, indicate ***on*** as the correct choice. Since we use their annotations as the benchmark against which to evaluate the model, this instance is counted as the classifier being wrong because it disagrees with the annotators. A better indication of the model's performance may be to independently judge its decisions, to avoid being subject to the annotators' bias. Finally, we are beginning to look at the relations between preposition errors and other types of error such as verb choice, and how these are annotated in the data.

An overview of the classifier's error patterns for the data in this task shows that they are largely similar to those observed in the L1 data. This suggests that the gap in performance between L1 and L2 is due more to the challenges posed by learner text than by inherent shortcomings in the model, and therefore that the key to better performance is likely to lie in overcoming these problems. In future work we plan to use L2 data where some of the spelling errors and non-preposition or determiner errors have been corrected so that we can see which of the other errors are worth focussing on first.

6.3 Determiners

Our work on determiner error correction is still in the early stages. We follow a similar procedure to the prepositions task, selecting a number of both correct and incorrect instances. On the former (set size 2000) accuracy is comparable to that on L1 data: **92.2%**. The danger of false alarms, then, appears not to be as significant as for the prepositions

task. On the incorrect instances (set size ca. 1200), however, accuracy is less than 10%.

Preliminary error analysis shows that the model is successful at identifying cases of misused determiner, e.g. ***a*** for ***the*** or vice versa, doing so in over two-thirds of cases. However, by far the most frequent error type for determiners is not confusion between indefinite and definite article, but omitting an article where one is needed. At the moment, the model detects very few of these errors, no doubt influenced by the preponderance of ***null*** cases seen in training. Furthermore, some of the issues raised earlier in discussing the application of NLP tools to L2 language hold for this task, too.

In addition to those, though, in this task more than for prepositions we believe that differences in text type between the training texts - the BNC - and the testing material - learner essays - has a significant negative effect on the model. In this task, the lexical items play a crucial role in class assignment. If the noun in question has not been seen in training, the classifier may be unable to make an informed choice. Although the BNC comprises a wide variety of texts, there may not be a sufficient number covering topics typical of learner essays, such as 'business letters' or 'postcards to penpals'. Also, the BNC was created with material from almost 20 years ago, and learners writing in contemporary English may use lexical items which are not very frequently seen in the BNC. A clear example of this discrepancy is the noun *internet*, which requires the definite article in English, but not in several other languages, leading to countless sentences such as *I saw it in internet*, *I booked it on internet*, and so on. This is one of the errors the model never detects: a fact which is not surprising when we consider that this noun occurs only four times in the whole of the training data. It may be therefore necessary to consider using alternative sources of training data to overcome this problem and improve the classifier's performance.

7 Comparison to human learners

In developing this model, our first aim was not to create something which learns like a human, but something that works in the best and most efficient possible way. However, it is interesting to see whether human learners and classifiers display similar patterns of errors in preposition choice. This information has twofold value: as well as being of pedagogical assistance to instructors of English L2, were the classifier to display student-like error patterns, insights into 'error triggers' could be derived from the L2 pedagogical literature to improve the classifier. The analysis of the types of errors made by human learners yields some insights which might be worthy of further investigation. A clear one is the confusion between the three locative and temporal prepositions ***at***, ***in***, and ***on*** (typical sentence: *The training programme will start* ***at*** *the 1st August*). This type of error is made often by both learners and the model on both types of data, suggesting that perhaps further attention to features might be necessary to improve discrimination between these three prepositions.

There are also interesting divergences. For example, a common source of confusion in learners is between ***by*** and ***from***, as in *I like it because it's* ***from*** *my favourite band.* However, this confusion is not very frequent in the model, a difference which could be explained either by the fact that, as noted above, performance on ***from*** is very low and so the classifier is unlikely to suggest it, or that in training the contexts seen for ***by*** are sufficiently distinctive that the classifier is not misled like the learners.

Finally, a surprising difference comes from looking at what ***to*** is confused with. The model often suggests ***at*** where ***to*** would be correct. This is perhaps not entirely unusual as both can occur with locative complements (one can *go to a place* or *be at a place*) and this similarity could be confusing the classifier. Learners, however, although they do make this kind of mistake, are much more hampered by the confusion between ***for*** and ***to***, as in *She was helpful* ***for*** *me* or *This is interesting* ***for*** *you*. In other words, for learners it seems that the abstract use of this preposition, its benefactive sense, is much more problematic than the spatial sense. We can hypothesise that the classifier is less distracted by these cases because the effect of the lexical features is stronger.

A more detailed discussion of the issues arising from the comparison of confusion pairs cannot be had here. However, in noting both divergences and similarities between the two learners, human and machine, we may be able to derive useful insights into the way the learning processes operate, and what factors could be more or less important for them.

8 Conclusions and future directions

This paper discussed a contextual feature based approach to the automatic acquisition of models of use for prepositions and determiners, which achieve an accuracy of 70.06% and 92.15% respectively, and showed how it can be applied to an error correction task for L2 writing, with promising early results. There are several directions that can be pursued to improve accuracy on both types of data. The classifier can be further fine-tuned to acquire more reliable models of use for the two POS. We can also experiment with its confidence thresholds, for example allowing it to make another suggestion when its confidence in its first choice is low. Furthermore, issues relating to the use of NLP tools with L2 data must be addressed, such as factoring out spelling or other errors in the data, and perhaps training on text types which are more similar to the CLC. In the longer term, we also envisage mining the information implicit in our training data to create a lexical resource describing the statistical tendencies observed.

Acknowledgements

We wish to thank Stephen Clark and Laura Rimell for stimulating discussions and the anonymous reviewers for their helpful comments. We acknowledge Cambridge University Press's assistance in accessing the Cambridge Learner Corpus data. Rachele De Felice was supported by an AHRC scholarship for the duration of her studies.

References

Briscoe, Ted, John Carroll, and Rebecca Watson. 2006. The second release of the RASP system. In *COLING-ACL 06 Demo Session*.

Chodorow, Martin, Joel Tetreault, and Na-Rae Han. 2007. Detection of grammatical errors involving prepositions. In *Proceedings of the 4th ACL-SIGSEM Workshop on Prepositions*.

De Felice, Rachele and Stephen Pulman. 2007. Automatically acquiring models of preposition use. In *Proceedings of the 4th ACL-SIGSEM Workshop on Prepositions*.

De Felice, Rachele. forthcoming. *Recognising preposition and determiner errors in learner English*. Ph.D. thesis, Oxford University Computing Laboratory.

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

Han, Na-Rae, Martin Chodorow, and Claudia Leacock. 2006. Detecting errors in English article usage by non-native speakers. *Natural Language Engineering*, 12(1):115–129.

Izumi, Emi, Kiyotaka Uchimoto, and Hitoshi Isahara. 2004. SST speech corpus of Japanese learners' English and automatic detection of learners' errors. *ICAME*, 28:31–48.

Turner, Jenine and Eugen Charniak. 2007. Language modeling for determiner selection. In *NAACL-HLT Companion volume*.

Yi, Xing, Jianfeng Gao, and William Dolan. 2008. A web-based English proofing system for ESL users. In *Proceedings of IJCNLP*.

Corrections(current). Submissions(current). A Classifier-Based Approach to Preposition and Determiner Error Correction in L2 English. Rachele De Felice, Stephen G. Pulman. Anthology ID