Prepositions and language transfer

Improving recall in automatic error correction

D. (Dylan) Bonga
Manuscript written during his RMA Linguistics, Utrecht University, Utrecht

KEYWORDS
Automatic error correction
language transfer
prepositions
second language acquisition

ABSTRACT
In second language learning, prepositions are often a source of mistakes. These mistakes can be analysed using automatic error correction, such as the Determiner And PrePosition Error Recogniser (De Felice & Pulman, 2008; 2009). However, these models often have a low recall level. This means that the model is not able to find the targeted data. I argue that the recall level can be improved by including possible language transfer from the native language (L1) to the model. To support this claim, I propose to replicate the study done by Jarvis & Odlin (2000), who studied Swedish and Finnish learners of English as a second language. However, I propose a replication with Hungarian speakers and Hungarian speakers of German. If agglutinative L1s (Finnish and Hungarian) show a similar distribution, and a different distribution from non-agglutinative L1s (Swedish and German), this could be implemented to improve the recall of the preposition error correction models.

1. Introduction

When learning a second language, typical learners will make many mistakes. Some of these mistakes will be resolved early on. However, some aspects of grammar remain more difficult to learn, such as the correct use of prepositions. The findings of an early study on English as a Second Language (ESL) show that prepositional errors are found in 18% of sentences written by learners of English, from 15 different language backgrounds (Dalgish, 1985). Although it is hard to generalize over such a diverse group of participants, one can cautiously assume that prepositions are problematic in the context of second language acquisition (SLA), at least in language writing. This assumption is further backed by Bitchener, Young, & Cameron (2005), who performed an experiment with 53 participants from a great variety of language backgrounds. They reported the mistakes made in an informal writing task and found that around 30% of the errors involved prepositions.

One way of resolving the mistakes made with prepositions, is constructing an automatic preposition error recogniser and using it to correct the errors. Such a system could make correcting for teachers much easier, but more importantly, it could also support second language learners when they are writing in their L2. Such systems were constructed, for example, by Chodorow, Tetreault, & Han (2007), as well as De Felice & Pulman (2008, 2009). Both use a maximum entropy classifier. This type of algorithm is used because the features observed in the data set (for instance:

words) are not conditionally independent. Even if they were, the maximum entropy classifier works more reservedly, whereas the alternative (a naive Bayesian classifier) would be too greedy. Chodorow et al. (2007) also showed that a naive Bayes model performs worse than the maximum entropy classifier they used.

In order to see how successful a model is, there are different measures to report. Three important measures that are very often used are accuracy, precision, and recall. Accuracy and precision are often seen as synonyms, which is incorrect. Table 1 gives the possible outcomes of a classifier. The observed values are the values that are considered the 'correct' values, whereas the predicted values are the values that the model assigns. Accuracy, precision, and recall are calculated by combining the outcomes from Table 1.

Table 1
Classification table of predicted and observed values.

	Observed values	
Predicted values	True	False
True	True positive	False positive
False	False negative	True negative

The definitions of the quality measures mentioned above are given in (1-3), as functions of the outcomes in Table 1. The abbreviations in (1)-(3) resemble the outcomes in Table 1. Applied to preposition correction, precision would be the amount of true corrections divided by the total amount of corrections. Recall can be interpreted as the amount of true corrections divided by the amount of cases where correction was needed. Lastly, accuracy is the amount of truly analysed cases (so correction if needed and no correction if no correction needed), divided by all counted instances.

(1) Precision: TP / (TP + FP)

(2) Recall: TP / (TP + FN)

(3) Accuracy: TP + TN / (TP + FP + FN + TN)

The system used in Chodorow et al. (2007) reached a precision of 80%. They report precision, instead of accuracy, because they stress that it is most important to have as little false positives as possible. They also report a recall level around 30%. This means that the system is unable to detect a great part of all made mistakes. De Felice & Pulman (2009) reported an accuracy of 70%, and a recall of 35%.[1] In other words, both systems process the data that they find fairly well, but they do not find much of the targeted data. Chodorow et al. (2007) reported that a low level of recall is typical for the domain of error detection.

[1] Additionally, De Felice & Pulman (2008) report a high accuracy on determiner correction. However, this was only tested on L1 data and it is not relevant to this proposal.

Alternatively, one could use unsupervised learning. This approach is taken by Islam & Inkpen (2010) for L1 data and Islam & Inkpen (2011) for L2 data. Islam & Inkpen (2011) reported that this method performs well, as both accuracy and recall are quite high. However, this study does not only consider prepositional errors, making this model incomparable to the maximum entropy classifier. Therefore, this proposal will not consider unsupervised learning as an alternative approach.

In this proposal, I shall further discuss the systems made by Chodorow et al. (2007) and De Felice & Pulman (2008, 2009). I will describe the training data and the classification process in both algorithms. Further, I shall discuss the quality measures that both systems achieve on written L2 data and how they can be interpreted. Both systems display a decent level of accuracy or precision, but a low level of recall. I will propose a follow-up study that further examines the effects of language transfer on prepositional use in L2 writing. Additionally, I will argue that taking language transfer into account in the classifier model may improve the level of recall for the maximum entropy classifier. The main question that this proposal will focus on, is formulated as follows: to what extend can the level of recall be improved when using a maximum entropy classifier for automatic preposition error correction?

2. Models for automatic error correction

This section looks into two models that are used for automatic error correction, namely Chodorow et al. (2007), in section 2.1, and De Felice and Pulman (2008, 2009), in section 2.2. Section 2.3 will briefly discuss how these models compare.

2.1 Classifier with rule-based filters

The classifier used by Chodorow et al. (2007) is a maximum entropy classifier combined with rule-based filters. These filters function as predefined rules of errors that are easily recognisable, such as redundantly repeated prepositions (‘friends with with the same interests’). This classifier was trained on a part of the MetaMetrics corpus and issues of the Sant Jose Mercury newspaper. The MetaMetrics corpus contains texts especially adapted to different reading levels. The part that was used in Chodorow et al. (2007) was aimed at 11th and 12th grade. This training data was tagged for Part-of-Speech (POS) and chunked. The test data was a random set of ESL essays. Here, the classifier had to assign a new value (an occurrence of a preposition) to one of 34 classes, with each class resembling one preposition. The assignment was done mainly on basis of the adjacent words, i.e., two or three words to the left and to the right of the target word. The classifier had a confidence parameter, which made the algorithm skip a correction if the top ranked candidates to replace the error were ranked too close to each other. For the testing phase, the threshold for the confidence parameter was set on 0.9. This means that if the algorithm is less than 90 percent certain about a correction, the error is skipped.

The test data were essays written by ESL students from a Chinese, Russian or Japanese background. Chodorow et al. (2007) reported several problems with the test data. Firstly, they found that there were many spelling errors and punctuation errors. This does not only influence the classifier, but also the POS-tagger, leading to a lower performance. The sentences that contained spelling and punctuation errors were omitted for that reason. The classifier was tested on a total of 2000 preposition contexts. This leads to a decent performance, with a precision measure of around 70%. They added a naive Bayesian classifier, which performs worse when used by itself. This naive Bayesian classifier was allowed to veto a decision made by the maximum entropy classifier, which raised to precision to 88% and lowered the recall even further to about 16%.

2.2 DAPPER

The maximum entropy classifier presented by De Felice & Pulman (2008, 2009) is DAPPER, or Determiner And PrePosition Error Recogniser. DAPPER is a classifier trained to find and evaluate the use of determiners and prepositions, for the purpose of error correction. For this classifier, the determiner error recogniser is a separate module that functions fully without the preposition error recogniser, and vice versa. Since this proposal only focusses on prepositions, the module that evaluates determiners is left aside.

De Felice & Pulman (2009) explained that their work is different than previous research, because they train the classifier on a wider range of syntactic and semantic components, which includes a full syntactic analysis of the context of a preposition. Instead of using a great number of different prepositions, like Chodorow et al. (2007), De Felice & Pulman (2009) restricted their search to the nine most frequent prepositions in the British National Corpus. They use about 9 million ‘contexts’, that contain one of these nine most frequent prepositions. All this data is processed with a POS-tagger, a stemmer, a lexicographer for verb and noun classes, a syntactic parser, and a named entity recogniser. Processing the data provides a feature vector, which is a representation of the context and all the extra information gained during processing. Furthermore, they only considered the contexts in which a preposition was needed.

De Felice & Pulman (2009) tested their model on a subset of the Cambridge Learner Corpus, which had been manually tagged for errors. They stripped the corpus of all error tags. The corpus contains data from learners from several different language backgrounds, but De Felice & Pulman (2008, 2009) did not report which backgrounds are represented in the test data. However, Nicholls (2003) reported that the complete corpus has data of 86 mother tongues and that about one third of the corpus is tagged for errors. Like in Chodorow et al. (2007), the corpus contains spelling errors. However, De Felice & Pulman (2009) did not omit the sentences containing these errors. The advantage of this approach is that there is more training data and more test data. They do explain that certain mistakes may lead to a worse performance, but also mention that this only concerns 3%

of the data. They do adjust the final level of performance to balance out these, and other, errors. This leads to an accuracy of 69% where the given preposition was used correctly and an accuracy of 42% where it needed to be corrected. They report a recall of 35%.

2.3 Comparison

Both systems mentioned above perform quite well on a difficult task. They both mention different types of difficulties and how they are coped with. One problem that can be resolved quite easily, is given by De Felice & Pulman (2009). While analysing the data, it came to their attention that the ESL students tend to rely on a small set of lexical chunks, which they overuse. Implementing the use of lexical chunks into the model would improve both precision and accuracy. However, one problem remains unsolved for both systems, which is the low level of recall. The source of this problem is unclear, although De Felice & Pulman (2009) did report that the recall levels vary greatly for the nine different prepositions they used. Even though this is not a priority for error correction algorithms, raising the recall would improve the applicability of the system.

3. Language transfer

One possible solution that will boost the recall is to implement L1 influence. Much research has been done on the topic of crosslinguistic influence (CLI), or language transfer. This influence can go from the L1 to the L2, but also the other way around. If there are more languages involved, like an L3 or L4, there can be influence from all the languages spoken by the learner (De Angelis, 2007; Odlin, 2013). The transfer can be positive, which can reinforce the acquisition process, or negative. In this proposal, I will focus mainly on negative transfer from the L1 into the L2. This form of transfer can be seen as interference of the L1, when using the L2.

CLI can be accounted for by most recent linguistic frameworks, such as Feature (re)assembly (Lardiere, 2008; Domínguez, Arche, & Myles, 2011), Usage-Based theory (Tomasello, 2000), and Dynamic System theory (De Bot, Lowie, & Verspoor, 2007). These frameworks make similar predictions on when language transfer happens. Both Feature (re)assembly and Usage-based theory predict that language transfer happens when the language learner has not fully acquired the new language and is still reassembling features (according to Feature (re)assembly) or adjusting the language mapping (according to Usage-based theory). This does imply that the learner should make no more mistakes when the language is fully acquired (and vice versa, that the language is only fully acquired when the learner makes no more mistakes).

Dynamic System theory rather approaches language learning as a linear function. Thus, the language acquisition grows by 1 plus the level of available resources. This principle covers the difference between learners from different L1s by stating that the resources available (from the L1) are different between learners. It also accounts for interlearner differences, since all learners have different resources and thus will learn differently.

As mentioned before, the use of prepositions remains a problem for L2 learners (Dalgish, 1985; Bitchener et al. 2005). Unsurprisingly, language transfer also happens in the use of prepositions, as is shown by Jarvis & Odlin (2000). They show that ESL learners with a Finnish L1 experience more difficulties using spatial-reference prepositions than ESL learners with a Swedish L1. They reason that Finnish learners perform worse, because they agglutinate spatial-reference in their L1, rather than using prepositions like Swedish and English. Examples are given in (4) for Swedish and (5) for Finnish[2]. Jarvis & Odlin (2000) reported that similar observations have been made by Sulkala (1996) about Estonian learners of English. Estonian, like Finnish, is an agglutinative language and expresses spatial reference by means of bound morphology.

(4) i huset
in house.specific
'In the house'

(5) talo -ssa
house-Inessive
'In the house'

In their study, Jarvis & Odlin (2000) let the participants watch a silent film and afterwards made them write a narrative for that film. They tested almost 400 participants, which results in a great amount of data. Because the writing task does not contain many restrictions for the participants, the resulting data is hard to compare. This is why they analysed the usage of prepositions in these narratives by focussing only on the verbs take and sit. Jarvis & Odlin (2000) explained that Finnish ESL learners prefer to use different prepositions than Swedish ESL learners. Furthermore, the Finnish learners omit prepositions where it is not allowed to omit, while the Swedish learners do not.

Summarized, the study by Jarvis & Odlin (2000) shows that the use of prepositions of Finnish and Swedish ESL learners follows a different distribution. This could be implemented into the maximum entropy classifier. Implementing the L1 transfer into a classifier would raise the possibility to focus on a contexts in which a speaker of a certain L1 would make the same mistake, where a speaker of a different L1 would not. More concretely, if a Finnish learner would omit prepositions where it is not allowed, where a Swedish learner would (almost) never do this, this could help the classifier analyse these contexts more thoroughly. In this way the classifier could detect more mistakes, thus increasing the level of recall.

[2] Special thanks to Walther Glödstaf for helping with these examples.

4. Proposal

Even though the conclusions from Jarvis & Odlin (2000) and Sulkala (1996) suggest that speakers of an agglutinative language behave similarly in respect to prepositional errors in ESL, there is need for more solid evidence before implementing this abstract concept of language transfer into the classifier. For that purpose, I propose to replicate the study done by Jarvis & Odlin (2000), but this time using speakers of Hungarian, which is an agglutinative language like, and also closely related to, Finnish. In order to have a comparable experimental group that lives in very similar conditions, like in Jarvis & Odlin (2000), I propose to also test Hungarian participants that have German as their L1, since German is a minority language in Hungary. German, like Swedish and English, uses prepositions and should thus behave differently than Hungarian in this test. Jarvis & Odlin (2000) only focussed on spatial-reference prepositions. Likewise, this paper will also focus on this type of prepositions. Including other types of prepositions would go beyond the scope of this proposal.

As mentioned before, Jarvis & Odlin (2000) predominantly looked at only two different verbs in both Swedish and Finnish. Ideally, more verbs would be analysed. However this would go beyond the scope of this proposal, since the analysis is a time consuming process. Thus, the replication study will consider the same verbs. Like in Jarvis & Odlin (2000) the participants will be in their third, fifth, or seventh year of learning English, both for the Hungarian speakers and for the German speakers. The participants will be shown the same 8-minute silent film (*Modern Times*) and will be asked to write a narrative immediately after, just like in Jarvis & Odlin (2000). Furthermore, a group of adult speakers of English will be compiled to function as a control group. The study is replicated as precisely as possible in order to make the results better comparable. This is necessary, since language transfer is a phenomenon that shows a high level of variation. In order to isolate language transfer, other possible influences (like age of the participants and the task at hand) should be accounted for by exactly replicating Jarvis & Odlin (2000).

5. Implications

Not only can the results of the proposed study strengthen the claims Jarvis & Odlin (2000) made, the collected data will also be very useful for comparison. The maximum entropy classifier could be used on the newly gathered data and on the data from Jarvis & Odlin (2000). The results could be helpful in generalizing the results for agglutinative languages.

The experiment I proposed can have multiple outcomes, of which I will discuss three. Firstly, if the results are alike for all four L1s (Finnish, Swedish, Hungarian, and German), or not similar at all, then language transfer might not be traceable (yet) in this data, or at least not with this method. Secondly, it could be possible that the languages used in Jarvis & Odlin (2000) show similar patterns, and the two languages from the replication study do the same. This would imply that the socio-topological environment (Finland versus Hungary) affects the use of prepositions in ESL. However, this seems highly unlikely, since Jarvis & Odlin (2000) already showed that Swedish and Finnish learners use prepositions differently and I expect the same results for German and Hungarian learners. Lastly, if prepositional errors are indeed influenced by language transfer, the results of the experiment would show that the agglutinative languages pattern together and the non-agglutinative languages pattern together. This last outcome would make language transfer a valuable attribute that can improve the maximum entropy classifier.

If indeed I find the third possible outcome, mentioned above, then language transfer should be implemented into the maximum entropy classifier, at least to make a difference between agglutinative languages and non-agglutinative languages. This should help increase the recall, as was argued before, because the classifier would have additional conditions or rules that can be applied to analyse a certain context and evaluate it. Importantly, it should be stressed that Jarvis & Odlin (2000), as well as this replication study, only focus on spatial-reference prepositions. Including different prepositions would help improve the classifier even further.

Further developing the maximum entropy classifier has several benefits, both for the teacher and the learner in ESL settings. One way of making the classifier especially helpful for the learner is suggested by De Felice & Pulman (2009): the classifier could mark the error and then rank several possible alternatives. However, they note that this is mostly useful for the more advanced students. Beyond helping the learners, the classifier could also be directly applied in everyday situations. For instance, it could be a helpful tool for non-native speakers that need to write an application letter. It could even be added to the grammar correcting software currently used in word processing programs.

Lastly, if the classifier is further improved, it can also be helpful to analyse prepositional errors in ESL corpora. As mentioned above in this proposal, different languages might pattern alike, according to similar language background. Not only could the classifier be used to identify other languages that show similar patterns as Finnish and Hungarian, or Swedish and German, it could also be used to identify new patterns. For instance, there are languages, such as Polish, that sometimes use prepositions and sometimes agglutinize. However, further research has to be done in order to permit an analysis of these languages. ■

Received April 2019; accepted April 2019.

References

Angelis, G. D. (2007). *Third or additional language acquisition*. Multilingual Matters.

Bitchener, J., Young, S., & Cameron, D. (2005). The effect of different types of corrective feedback on ESL student writing. *Journal of Second Language Writing*, *14*(3), 191–205.

Chodorow, M., Tetreault, J., & Han, N. R. (2007). Detection of grammatical errors involving prepositions. *Proceedings of the Fourth ACL-SIGSEM Workshop on Prepositions*, 25–30.

Dalgish, G. (1985). Computer-assisted ESL research and courseware development. *Com-*

puters and Composition, *2*(4), 45-62.

De Bot, K., Lowie, W., & Verspoor, M. (2007). A dynamic systems theory approach to second language acquisition. *Bilingualism: Language and Cognition*, *10*(01), 7–20.

De Felice, R., & Pulman, S. (2008). A classifier-based approach to preposition and determiner error correction in L2 English. In *Proceedings of the 22nd International Conference on Computational Linguistics - COLING '08*, *1*, 169–176.

De Felice, R., & Pulman, S. (2009). Automatic detection of preposition errors in learner writing. *CALICO Journal*, *26*(3), 512–528.

Dominguez, L., Arche, M. J., & Myles, F. (2011). Testing the predictions of the feature-assembly hypothesis: evidence from the L2 acquisition of Spanish aspect morphology. In *BUCLD 35: Proceedings of the 35th Annual Boston University Conference on Language Development, 1*, 183-196.

Islam, A., & Inkpen, D. (2010). An unsupervised approach to preposition error correction. In *Proceedings of the 6th International Conference on Natural Language Processing and Knowledge Engineering,* 1–4.

Islam, A., & Inkpen, D. (2011). Correcting different types of errors in texts. In C. Butz & P. Lingras (Eds.), *Advances in Artificial Intelligence* (pp. 192–203). Springer Berlin Heidelberg.

Jarvis, S., & Odlin, T. (2000). Morphological type, spatial reference, and language transfer. *Studies in Second Language Acquisition*, *22*(4), 535–556.

Lardiere, D. (2008). Feature assembly in second language acquisition. *The role of formal features in second language acquisition*, 106-140.

Nicholls, D. 2003. The Cambridge learner corpus: Error coding and analysis for lexicography and ELT. In *Proceedings of the Corpus Linguistics 2003 conference,* pages 572–581.

Odlin, T. (2013). Crosslinguistic influence in second language acquisition. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics*, (pp. 436-486). Oxford, UK: Blackwell Publishing Ltd.

Sulkala, H. (1996). Finnish as a second language for speakers of related languages. In M. Martin & P. Muikko-Werner (Eds.), *Finnish and Estonian: New target languages* (pp. 143–158). Jyväskylä, Finland: Center for Applied Language Studies, Universityof Jyväskylä.

Tomasello, M. (2001). First steps toward a usage-based theory of language acquisition. *Cognitive Linguistics*, *11*(1–2), 61-82.