Identifying negative language transfer in learner errors using POS information

Leticia Farias Wanderley
EdTeKLA Research Group
Department of Computing Science
University of Alberta
fariaswa@ualberta.ca

Carrie Demmans Epp
EdTeKLA Research Group
Department of Computing Science
University of Alberta
cdemmansepp@ualberta.ca

Abstract

A common mistake made by language learners is the misguided usage of first language rules when communicating in another language. In this paper, n-gram and recurrent neural network language models are used to represent language structures and detect when Chinese native speakers incorrectly transfer rules from their first language (i.e., Chinese) into their English writing. These models make it possible to inform corrective error feedback with error causes, such as negative language transfer. We report the results of our negative language detection experiments with n-gram and recurrent neural network models that were trained using part-of-speech tags. The best performing model achieves an F1-score of 0.51 when tasked with recognizing negative language transfer in English learner data.

1 Introduction

Advances in grammatical error correction (GEC) research allow writing editors to provide real-time corrective feedback to language learners. GEC systems are trained on parallel erroneous and corrected learner data to detect and suggest corrections for user errors. State-of-the-art GEC systems apply neural machine translation to learn how to correct grammatical errors in English learner data (Bryant et al., 2019). These systems also use metadata, such as the learners' native languages and their proficiency levels in the target language, to support the GEC task (Nadejde and Tetreault, 2019).

The direct corrective feedback derived from GEC systems' predictions helps learners improve their writing and increase their language understanding. However, direct corrective feedback is not the only type of information that can aid language learning (Bacquet, 2019). Learners also benefit from metalinguistic feedback. This feedback type can support learners' reflections on their language usage and, consequently, increase their grammatical awareness (Karim and Nassaji, 2020).

One metalinguistic phenomenon that often causes confusion and errors in learner writing is the occurrence of language transfer. The language transfer phenomenon is characterized by learners reusing rules from their first languages (L1s) when communicating in a second one[1] (L2) (Lado, 1957). When the L1 and L2 grammars diverge, learners who apply L1 rules make mistakes. This particular version of language transfer is called negative language transfer, and it is a significant source of errors in second language learner speaking and writing.

This paper explores the application of language models to represent language structures and detect negative language transfer in learner essays written by Chinese native speakers. The proposed methods aim to identify incorrect learner utterances that have structural patterns that are more similar to the learners' L1 (Chinese) than to their L2 (English). In this paper, we demonstrate that a recurrent neural network trained to identify a part-of-speech (POS) tag sequence's source language outperformed POS n-gram models in negative language transfer detection. The RNN model achieved an F1-score of 0.51 on the negative language transfer detection task analysing POS tag sequences extracted from learner errors. The output of this language model can be used to inform metalinguistic feedback, tying the incorrect utterances to differences between the L1 and English rules. By enabling negative language transfer detection in learner writing, language learners will have access to interpretable information about their errors' potential causes along with direct corrective feedback for those errors.

[1] By second language, we mean any additional language beyond the learner's mother tongue.

Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications, pages 64–74

2 Related work

When writing in another language, writers are prone to leave behind certain traces of their native languages. This fact is the foundation for the native language identification task. This task makes use of computational models to determine the L1 of a text's author (Tetreault et al., 2013; Malmasi et al., 2017). Native language identification models aim to uncover usage patterns that distinguish native speakers of a certain language when writing in English. Rabinovich et al. (2018) explored the usage of cognates in non-native English writing to cluster the authors' text according to their native languages' families. Cognates are words from distinct languages that are similar in meaning and form. As demonstrated by Rabinovich et al. (2018), non-native English writers show a preference for using English words that have cognates in their L1. Furthermore, Flanagan et al. (2015) have shown that it is possible to employ the authors' error patterns to discriminate between L1s, as English learners' writing errors are highly correlated to their native languages.

Learner writing errors are useful in yet another computational task. The aforementioned grammatical error correction task aims to detect and correct errors in learner data (Ng et al., 2013, 2014; Bryant et al., 2019). State-of-the-art GEC systems model the task as a neural machine translation procedure in which the erroneous learner data is the source language, and its corrected version is the target language (Bryant et al., 2019). Throughout the years, researchers have explored the impact of using L1-specific learner data to train GEC systems: see Rozovskaya and Roth (2011), Chollampatt et al. (2016), and Nadejde and Tetreault (2019) for examples. These researchers have found that whether GEC models are exclusively trained with L1-specific data or whether this data is only used to fine-tune the model, GEC benefits from information about learners' L1s. In addition to that, Nadejde and Tetreault (2019) explored fine-tuning GEC systems to learners' English proficiency levels and L1s. They found that information about both learner aspects improves GEC performance.

English non-native speakers transfer patterns and rules from their L1s into English writing. This phenomenon has been explored by native language identification and grammatical error correction systems. Moreover, the phenomenon has been demonstrated in experimental investigations of the contrastive analysis hypothesis (Lado, 1957), as reported in papers by Wong and Dras (2009) and Berzak et al. (2015). Berzak et al. (2015) found a correlation between structural error distribution in English as a second language learners' writing and typological differences between the learners' L1s and English. Their work used the L1's structural properties to predict challenging grammatical areas in English as a second language writing.

One of the challenges for language learners is that they are often unaware that they reuse structures from their L1s (Wanderley and Demmans Epp, 2020). One way to alleviate this difficulty is to provide error feedback that contains information about possible error sources (Bacquet, 2019). Linguists and education researchers have long debated what is the best way to provide feedback in a second language setting. Some argue that no feedback should be given, while others posit that learners should have access to a different range of error feedback types, such as direct, indirect, comprehensive, and focused (Bitchener et al., 2005). More recent research has found that providing metalinguistic feedback can increase language learners' writing accuracy (Bacquet, 2019; Karim and Nassaji, 2020). Unlike direct corrective feedback, which simply highlights and corrects learner errors, metalinguistic feedback can help learners understand their errors by calling attention to the errors and describing possible causes (Lyster and Ranta, 1997). To understand whether students value metalinguistic feedback about negative language transfer, Watts (2019) examined Japanese students' familiarity with the phenomenon. These learners highlighted how being aware of language transfer effects helped improve their English writing.

While natural language processing (NLP) tasks and language learners benefit from information about negative language transfer, we are not aware of previous research that automatically detects this phenomenon. Recently, Monaikul and Di Eugenio (2020) applied traditional and neural natural language processing methods to detect preposition omission errors with the aim of informing negative language transfer feedback. Preposition errors are one of the most common types of negative language transfer errors in second language learning. In their paper, the authors proposed applying the preposition error detection output to generate metalinguistic contrastive feedback for language learners and teachers (Monaikul and Di Eugenio, 2020).

Dataset	Number of sentences
Global Voices	138 582
WMT19	11 960
Combined	150 542

Table 1: Training datasets sizes

Error type	Count
Structural negative language transfer	1457
Structural not negative language transfer	914
Total	2371

Table 2: Structural negative language transfer error counts from the test dataset

We envision a similar application of our results. However, our focus is on detecting additional negative language transfer errors. We want to detect all those that come from learners inappropriately applying structural patterns from their L1s when writing in English.

3 Data

3.1 Training data

In this paper, we explore the application of shallow syntactic language models in negative language transfer detection. We use the part-of-speech sequences within a language to create what we call a shallow syntactic language model, thus representing language structures. Instead of representing the probability of word sequences in a language, these models compute the likelihood of POS tag sequences. As we wanted to model L1 and English structures with these models, we needed textual data in both languages. To ensure that the shallow syntactic language models in the L1 and in English were equivalent in number of training samples and context, parallel data in those languages were used for training.

Two parallel datasets were used to train the Chinese and English models. These datasets are available online and are aligned at the sentence level (Tiedemann, 2012). The first data source is the Global Voices dataset (Prokopidis et al., 2016). It contains multilingual parallel media stories from the Global Voices website[2]. The second dataset is the news test set from the ACL 2019 Fourth Conference on Machine Translation (WMT19)[3] (Barrault et al., 2019). Table 1 presents the number of parallel sentences in each dataset. These datasets were chosen because they contain text that observes simple yet grammatical language structures, patterns that language learners are familiar with and encouraged to use.

[2] https://globalvoices.org/

[3] http://statmt.org/wmt19/translation-task.html

3.2 Test data

Our test data consists of error annotated English learner data. This data was extracted from the First Certificate in English (FCE) dataset (Yannakoudakis et al., 2011). The FCE exam is an upper-intermediate English language certification. The 1244 essays in the FCE dataset were written by English as a second language learners who took the exam between 2000 and 2001. These essays were manually error annotated. During the annotation process, the FCE annotators applied the error coding system described in Nicholls (2003) to highlight writing errors and correct them.

The FCE dataset is suitable to our task because each essay is annotated with the native language of its writer (Yannakoudakis et al., 2011). In addition to the original FCE data, the highlighted errors were annotated with information regarding their relation to negative language transfer. Each error in the dataset is accompanied by a flag indicating whether its grammatical structure resembles the learner's L1. The negative language transfer annotation process was performed by a native speaker of English and Mandarin Chinese who teaches Chinese as a foreign language. The annotation results can be accessed through our research laboratory's website[4]. In total, there are 3092 errors in essays written by Chinese native speakers. Out of these, 1776 (57.4%) are negative language transfer related (Wanderley et al., 2021). The negative language transfer annotation allows the evaluation of our proposed detection methods. It is our gold-standard.

Since we aim to detect structural negative language transfer, we filtered the errors on the test set to only contain structural errors. We borrow some of the definition of structural errors from Berzak et al. (2015), filtering out spelling and word replacement errors from the test dataset. The resulting dataset contains 2371 structural error instances. Among those, 61.45% are related to negative language transfer and 38.55% are not (see Table 2).

[4] https://github.com/EdTeKLA/LanguageTransfer

Language	Min	Q1	Median	Q3	Max
English	1	12	19	29	258
Chinese	1	6	10	16	230

Table 3: Length statistics of part-of-speech tagged training sequences

	Training split	Evaluation split
Chinese	120 433	30 109
English	120 433	30 109

Table 4: Number of sequences belonging to the training and evaluation splits that were used in hyperparameter tuning

3.3 Data preprocessing

As previously mentioned, we used Chinese and English parallel data to build shallow syntactic language models. These models were trained with POS tag sequences extracted from the parallel training sentences. The data was POS tagged using the Python library spaCy[5].

SpaCy is an NLP library that has pre-trained Chinese and English POS taggers. SpaCy's pre-trained taggers allowed the sentences from the training data to be POS tagged according to their languages. They were tagged either by a POS tagger trained on Chinese text or by one trained on English text. Table 3 summarizes the training sequences' lengths. While the training sentences were tagged by one of two POS taggers (i.e., English or Chinese), the FCE dataset sentences were only tagged with spaCy's English POS tagger.

The Universal Dependencies[6] POS tagset was used. This treebank defines syntactic annotations that are common among a variety of languages, including English and Chinese (Nivre et al., 2016). It fits our experiments as it allows us to model language structures using a shared POS tagset, and consequently, consistently compare a single POS tag sequence across languages.

4 Methods

4.1 N-gram baseline

N-gram language models are a statistical NLP approach to language modelling that represents a language as its token sequence distribution. These models output the likelihood of a sequence belonging to the language they represent (Jurafsky and Martin, 2009). N-gram shallow syntactic language models represent a language's structure as the distribution of POS tag sequences in the language. These models function in a straightforward way. During training they derive a probability distribution over all possible POS tag sequences in the training data. Then, when provided a test sequence, they compute its likelihood based on the training probability distribution.

N-gram shallow syntactic language models were used to assign probabilities to POS tag sequences extracted from learner errors. As one n-gram model is only able to represent one language. We trained one model on POS tag sequences extracted from the English sentences and another model on POS tag sequences extracted from the Chinese text. After the models were trained, each erroneous sequence in the FCE dataset was evaluated by both n-gram models separately, and the probability values output by the n-gram models were compared. If the Chinese model's output for a POS tag sequence was higher than the English model's output, the error which generated that POS tag sequence was classified as negative language transfer.

The n-gram shallow syntactic language models used in our experiments were trained using the Python interface for KenLM[7]. KenLM is a language model package that applies several performance improvement techniques to its n-gram modelling implementation. For example, KenLM applies modified Kneser-Ney smoothing to the n-gram distribution (Heafield et al., 2013). Moreover, this language model implementation improves querying time by using trie data structures and linear probing to store and retrieve n-gram probabilities (Heafield, 2011).

The best parameter setting for the n-gram models was selected through a systematic search over the parameter space. In this process, 20% of each monolingual training dataset was kept as an evaluation set. The remaining 80% was used to train n-gram shallow syntactic language models with varying n-gram lengths (2 to 6). See Table 4 for the number of training and evaluation sequences used in this tuning process. The n-gram models that achieved the best accuracy (96.94%) in identifying the language of POS tag sequences extracted from evaluation sentences had n = 5. That is, they analysed one sequence of five POS tags at a time.

[5] https://spacy.io/

[6] https://universaldependencies.org/

[7] https://kheafield.com/code/kenlm/

4.2 Recurrent neural network approach

Recurrent neural networks (RNNs) are artificial neural networks known to effectively model sequential data, such as natural language (Dyer et al., 2016). Mikolov et al. (2010) have shown that RNN architectures outperform other statistical methods in language modelling tasks. In RNN architectures, each unit's output is based not only on its respective input but also on the output of preceding units. Hence, these networks are able to take previous context into account when processing current information. This recurrent aspect of RNN architectures is particularly relevant to language modelling tasks, in which tokens are often related to remote neighbours.

One advantage of RNNs over n-gram models in detecting negative language transfer is that a single network is able to learn how to differentiate between languages. As opposed to the n-gram baseline that needed one model for English and one for Chinese, a single RNN model can be trained to identify a POS tag sequence's originating language based on its structure. During training, the RNN model learned how to predict language from POS tag sequences extracted from English and Chinese text. Then, for each erroneous POS tag sequence in the FCE dataset, the trained RNN would predict English or Chinese, according to which language structure the POS tag sequence most resembled. If the network returned "Chinese", the error which generated that POS tag sequence was flagged as negative language transfer.

The RNN implementation from the Python library PyTorch[8] was used in our experiments. The model was trained for ten epochs with Adam optimization (Kingma and Ba, 2015). Each of the 17 POS tags in the Universal Dependencies tagset was converted into a one-hot-encoding vector of length 17, i.e., each distinct tag in the tagset was represented by one position in the vector. Using these vector representations, the POS tag sequences in the training and test datasets were converted to POS tag vector arrays, preserving the original POS tag sequences' orders.

Similarly to the n-gram baseline, the RNN hyperparameters were selected from a set of options in the parameter space. The evaluation of the hyperparameter settings was made on 20% of the training data (the evaluation split), while the network was trained on the remaining 80% of the training data. Table 4 provides the number of POS tag sequences included in each split. In total, four parameters were tuned - the number of hidden units (8, 16, 32, 64, 128, 256, or 512 units); the learning rate (0.01, 0.001, 0.0001, 0.00001, or 0.000001); the batch size (1, 2, 4, 8, 16, or 32 samples per mini batch); and the loss function (negative log likelihood[9] or binary cross entropy with logits[10]). The RNN models analysed in the hyperparameter tuning procedure were trained for 10 epochs. The best performing model, which achieved 95.16% accuracy on the evaluation set, had 16 hidden units, learning rate = 0.0001, mini batch size = 1, and negative log likelihood as its loss function.

Note that the best evaluation accuracy achieved in the RNN hyperparameter tuning was not as high as the one achieved in the n-gram models' tuning procedure. These results are a reflection of the type of data being modelled in this task. The POS tag sequences extracted from sentences in Chinese and English are from high-quality (i.e., grammatical and manually translated) data that is characterized by a high degree of regularity which is well-represented by n-gram models. Each n-gram model trained on one of the parallel datasets can accurately classify POS tag sequences as belonging to a language or not. The n-gram model for one language cannot, however, infer whether the POS tag sequences belong to the language represented by the other model. In this aspect, RNN-based models are more fitted to the task. They are able to specifically differentiate between Chinese and English structures and decide to which language a POS tag sequence belongs.

4.3 Test sequences

Among other aspects, structural errors in the FCE dataset vary with regards to length and dependency. For example, a wrong word order error consists of several words incorrectly arranged. The number of incorrect words in wrong word order errors is always greater than one, while an unnecessary adverb error, for instance, consists of one single incorrect word. As for the dependency aspect, the incorrect words from an error usually are in conflict with another word in the sentence. The positions of the word involved differ according to error type. For

[8] https://pytorch.org/

[9] https://pytorch.org/docs/stable/generated/torch.nn.NLLLoss.html

[10] https://pytorch.org/docs/stable/generated/torch.nn.BCEWithLogitsLoss.html

Incorrect sentence	**Error subtype**	**Padded error span**	**Error + unigram span**	**Error + bigram span**
Moreover, the stars and artists were *only from* six countries.	Wrong word order	were *only from* six AUX ADV ADP NUM	*only from* six ADV ADP NUM	*only from* six countries ADV ADP NUM NOUN
They flew out of the window and hit *directly* the headmaster's head.	Unnecessary adverb	hit *directly* the VERB ADV DET	*directly* the ADV DET	*directly* the headmaster ADV DET NOUN
This are only my immature views.	Pronoun agreement	*This* are DET AUX	*This* are DET AUX	*This* are only DET AUX ADV
This *remind* me of what I experienced.	Verb agreement	This *remind* me DET VERB PRON	*remind* me VERB PRON	*remind* me of VERB PRON ADP

Table 5: Learner errors and test windows examples

example, a pronoun agreement error is often caused by disagreement with the subsequent word. However, a verb agreement error usually comes from an incongruence with its preceding word. Table 5 presents learner error examples along with their distinct lengths and dependencies. These distinctive aspects prompted the design of experiments that evaluated different spans of erroneous POS tag sequences.

The negative language transfer detection models received POS tag sequences extracted from learner errors as test input. To attempt to capture error dependencies in these sequences, we defined and tested three different POS tag sequence spans. Regardless of the span selected, all of the error's POS tags were included in the test sequences. The spans only defined how much of the error's surrounding context was included in the test POS tag sequences. They defined whether the POS tags preceding and following the errors would be part of the sequence evaluated by the shallow syntactic language models.

The first input span considered was the padded error sequence. These test sequences contained the POS tag that precedes the erroneous tags, the tags extracted from the error, and the subsequent POS tag, which is the one that comes immediately after the erroneous tags. The preceding and subsequent tags were only included if there were words before and after the error, e.g., when the error occurred on the first word of a sentence, the preceding tag was not included in the input test sequence as it did not exist. The second and third spans that were tested took into account the POS tags extracted from the error followed by POS tags from words that followed the error. One of them considers the tag from the word that immediately follows the error, while the other considers the tags from the error's two subsequent words. They are referred to as "error + unigram" and "error + bigram" spans. Table 5 illustrates the POS tag spans used in the experiments.

There is one type of error that required an adjustment to the span definitions. Omission errors, such as missing noun errors, happen when the learner has not included a word that was necessary. If represented by the error + unigram span following its general definition, these errors would consist of an isolated tag, but this representation could not convey much of the error's context. For this reason, omission errors were represented by slightly modified versions of the error + unigram and error + bigram spans. When representing missing word errors, the error + unigram span consisted of the POS tags extracted from the two words that followed the error. Similarly, the error + bigram span consisted of the POS tags extracted from the three words after the error when an omission error occurred. This adjustment was envisioned to maintain the amount of information represented by each error span consistent throughout the experiments.

Compared to the POS tag sequences used for training the n-gram and RNN models, the test POS tag sequences extracted from learner errors are short. Table 6 presents a summary of the test sequences' lengths. The median values for these sequence lengths indicate that most learner errors involved a single incorrect word.

Span	Min	Q1	Median	Q3	Max
Padded error	1	3	3	3	14
Error + unigram	1	2	2	3	13
Error + bigram	1	3	3	4	14

Table 6: Part-of-speech tagged test sequence lengths

		P	R	F1
N-gram				
	Padded error	0.68	0.32	0.43
	Error + unigram	0.64	0.34	0.45
	Error + bigram	0.66	0.27	0.38
RNN				
	Padded error	0.69	0.34	0.46
	Error + unigram	0.67	**0.41**	**0.51**
	Error + bigram	**0.70**	0.35	0.46

Table 7: Negative language transfer detection results

The decision to analyse manually annotated learner errors implies a setting that may be hard to reproduce, as in reality learner error information is not always available or correctly identified. The POS tag sequences analysed by the shallow syntactic language models were extracted from sentences in which the exact location and type of the errors was known. Although grammatical error correction research has achieved impressive results, it is not a solved task. In applying negative language transfer detection techniques to GEC system outputs it would be necessary to keep in mind and design around possible misclassifications.

5 Results

Table 7 presents the precision, recall, and F1-score results for the structural negative language transfer detection task. The error spans being equal, the RNN approach results outperformed the n-gram baseline in all metrics. Both approaches yielded the highest recall and F1-scores when analysing errors represented by the error + unigram span. However, the precision scores achieved in these experimental settings were the lowest ones within each approach.

The n-gram baseline paired with the padded error representation was more precise in classifying negative language transfer than the RNN when it analysed errors represented with the error + unigram span. However, the recall scores yielded by these two experimental combinations show that the RNN approach was able to retrieve more negative language transfer errors than the n-gram baseline. Although both n-gram and RNN approaches achieved analogous precision results, especially when analysing padded errors, the n-gram baseline was consistently outperformed on recall. These results indicate that comparable precision came with a lower recall cost for the baseline.

The high precision compared to low recall rates achieved by the n-gram baseline can be attributed to the fact that these models compute their output values from independent probability distributions derived from the training data. This may indicate that these models excel in classifying negative language transfer when there is a clear distinction in usage of the error structure between the languages. For example, if a learner applied a language structure that is frequent in Chinese but extremely rare in English, the Chinese n-gram model would assign a much higher probability than the English n-gram model, and the error would be correctly classified as negative language transfer. However, these very prominent usage differences are scarce in our data. As the FCE test-takers are learning English, they may not fully apply Chinese rules but instead combine rules from the two languages. This interlingual process makes the error structures ambiguous to our independent n-gram models.

The results in Table 7 suggest that the RNN's capability of distinguishing between languages might make it more suitable for the negative language transfer detection task. The RNN model that used the error + unigram test span achieved the highest recall and F1-score results in correctly classifying errors as negative language transfer. Moreover, the RNN model with the error + bigram span yielded the highest precision results across all experiments.

6 Discussion

The error coding used by the FCE annotators is useful to our error analysis as it allows the learner errors to be grouped by error subtype (Nicholls, 2003). Analysing which error subtypes cause the shallow syntactic language models to misclassify negative language transfer errors provides insights as to how our methods could be revised to better support the task.

Our models performed poorly when detecting negative language transfer in replacement errors. Replacement errors commonly occur when the learner uses the correct part-of-speech, but incorrect word in a sentence. For example, the word

"many" in the sentence "how many you charge?" was flagged as a quantifier replacement error in the FCE dataset. Both n-gram and RNN shallow syntactic language models performed poorly in identifying whether replacement errors were related to negative language transfer. This may have occurred due to the POS tag representation of these errors not conveying a lot of information about the error's source. The erroneous and correct POS tag sequences follow the same pattern, a pattern that is acceptable in English.

The POS tag representation of errors also hindered negative language transfer detection for errors involving verb tenses, forms, and agreement. The Universal Dependencies tagset annotates verbs with one of two tags, "VERB" or "AUX". None of these tags contains information about verb features, such as tense, form, or person. These features would be useful in detecting verb usage patterns that are more common in Chinese than English and could be related to negative language transfer. These error subtype results indicate that the models would benefit from a more detailed POS tagset. A tagset that encapsulates more word features in their tags, such as the Penn Treebank tagset (Marcus et al., 1993). The challenge in using a more detailed tagset is that it needs to be common between English and the learners' L1 (in our case, Chinese). The Universal Dependencies tagset is the only tagset we are aware of that fits this commonality requirement.

Chinese speakers who are learning English as a second language usually have trouble with punctuation marks, as these symbols are employed differently in Chinese (Liu, 2011). For example, in Chinese, commas are used to mark the end of a complete thought (Xue and Yang, 2011). In our experiments, the POS tag span selection was fundamental to the detection of negative language transfer in missing punctuation errors. Both the n-gram and RNN approaches achieved the highest F1-scores when detecting negative language transfer in missing punctuation errors when using the padded error span. When evaluating the padded error span for missing punctuation errors, the RNN model achieved an F1-score of 0.48. The error + unigram and error + bigram spans yielded F1-scores of 0.43 and 0.41, respectively. A few possible explanations for this disparity in scores are that the errors' previous contexts helped the models correlate the errors to the Chinese language, or the lack of punctuation in between the spans' POS tags is divergent from English grammar rules.

Another common mistake for Chinese speakers who are learning English is determiner omission. As determiners do not exist in the Chinese language, learners tend to neglect their application in English (Robertson, 2000). Contrary to our initial assumption, missing determiner errors were better classified by the RNN model when using the error + unigram span. The previous context from the padded error span did not help the classification. One hypothesis for this behaviour is that it is common for determiners, as opposed to other parts-of-speech that usually follow determiners, to appear at the beginning of sentences in English. As there are no determiner equivalents in Chinese, whenever the RNN analysed a POS tag sequence that should begin with a determiner in English, but would not in Chinese, the sequence was classified as negative language transfer. In these cases, the padded error span contained another POS tag preceding the position where the determiner should have been. The RNN model might have been confused by this preceding POS tag, as there are mid-sentence situations in English in which determiners are not necessary.

The hypothesis that the RNN expected determiners to be at the beginning of English sequences as opposed to Chinese ones is corroborated by the error + bigram span also yielding a higher F1-score than the padded error span in this setting. The error + bigram span is analogous to the error + unigram span as it begins with the POS tag extracted from the word that should be preceded by a determiner.

To further investigate errors in model classification of negative language transfer, we charted the overall task performance and F1-scores for some specific subtypes of learner transfer (Figure 1). This figure shows that although the error + unigram representation yielded the highest overall F1-score in the n-gram baseline setting, certain predefined sub-types of negative language transfer were better represented by the padded error span. This difference was seen when the learner had not used punctuation or determiners even though they were needed. In the missing punctuation case, the n-gram models with the padded error span outperformed all of the RNN results. This indicates that the padded error span and n-gram models were more capable of representing the distinction between Chinese and English punctuation usage. In

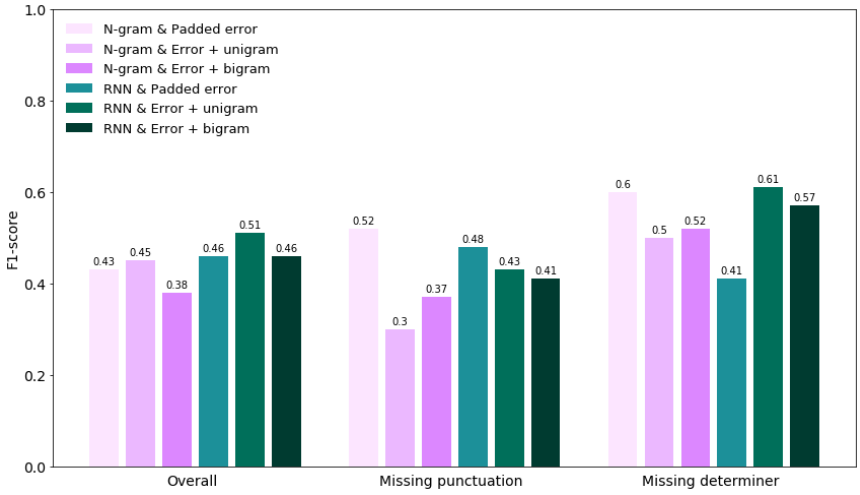

Figure 1: F1-scores for shallow syntactic language models and error spans combinations

this experimental setting, the n-gram models were able to detect when a POS tag sequence was meant to contain a punctuation mark in English but that punctuation mark was not necessary in Chinese.

As for the missing determiner errors, the padded error span also outperformed the other error spans in the n-gram shallow syntactic language models setting. Unlike the RNN, n-gram models do not have such a defined concept of sentence beginning. This distinction may help explain the padded error span's performance in the n-gram setting. The n-gram models' computation is based on how common a POS tag sequence is. In that case, the padded error span POS tags were able to indicate that while a determiner would follow and precede certain tags in English, it would not in Chinese. The n-gram models correctly identified a Chinese structural pattern in missing determiner errors' POS tag sequences because their POS tags were not followed or preceded by determiners.

Overall, the language modelling techniques used in our experiments were able to model the distinction between English and Chinese structures when analysing POS tag sequences extracted from sentences in those languages. When applied to the edits extracted from Chinese native speakers' English writing errors, the models achieved good precision in identifying errors related to negative language transfer. However, they were unable to detect a large portion of the negative language transfer related errors, i.e., low recall scores were observed.

As a future step, it would be interesting to attempt to detect negative language transfer with more powerful language modelling techniques, such as transformer models. Unlike RNN-based language models, transformer-based approaches do not rely solely on recurrent features to represent language structures (Vaswani et al., 2017). The self-attention mechanisms employed by these models could be beneficial in detecting language patterns in learner text that are more similar to the learners' L1s than to English.

7 Conclusion

Our experiments show that shallow syntactic language models can identify some negative language transfer related errors in English learner writing. While improvements can be made regarding the level of detail of the structural representation, our experiments indicate that it is possible to detect aspects of transferred language structures.

Negative language transfer detection methods, such as the ones we described, can support the generation of error feedback that connects learners' mistakes to their first languages. This type of information can help language learners become more aware of rule divergences between English and their L1s, and as a result increase their writing accuracy. In the future, we hope to apply the methods described above to other first languages and learner datasets, such as the Lang-8 Corpus of Learner English (Mizumoto et al., 2011).

Acknowledgements

We acknowledge the financial support of the Natural Sciences and Engineering Research Council of Canada [RGPIN-2018-03834].

References

Gaston Bacquet. 2019. Is corrective feedback in the ESL classroom really effective? A background study. *International Journal of Applied Linguistics and English Literature*, 8:147–154.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Yevgeni Berzak, Roi Reichart, and Boris Katz. 2015. Contrastive analysis with predictive power: Typology driven estimation of grammatical error distributions in ESL. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 94–102, Beijing, China. Association for Computational Linguistics.

John Bitchener, Stuart Young, and Denise Cameron. 2005. The effect of different types of corrective feedback on ESL student writing. *Journal of Second Language Writing*, 14(3):191 – 205.

Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. The BEA-2019 shared task on grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Florence, Italy. Association for Computational Linguistics.

Shamil Chollampatt, Duc Tam Hoang, and Hwee Tou Ng. 2016. Adapting grammatical error correction based on the native language of writers with neural network joint models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1901–1911, Austin, Texas. Association for Computational Linguistics.

Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A. Smith. 2016. Recurrent neural network grammars. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 199–209, San Diego, California. Association for Computational Linguistics.

Brendan Flanagan, Chengjiu Yin, Takahiko Suzuki, and Sachio Hirokawa. 2015. Prediction of learner native language by writing error pattern. In *Learning and Collaboration Technologies*, pages 87–96, Cham. Springer International Publishing.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 690–696, Sofia, Bulgaria. Association for Computational Linguistics.

Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing (2nd Edition)*. Prentice-Hall, Inc., USA.

Khaled Karim and Hossein Nassaji. 2020. The revision and transfer effects of direct and indirect comprehensive corrective feedback on ESL students' writing. *Language Teaching Research*, 24(4):519–539.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Robert Lado. 1957. *Linguistics Across Cultures: Applied Linguistics for Language Teachers*. University of Michigan Press.

Xing Liu. 2011. The not–so–humble "Chinese comma": Improving English CFL students' understanding of multi–clause sentences. In *Proceedings of the 9th New York International Conference on Teaching Chinese*.

Roy Lyster and Leila Ranta. 1997. Corrective feedback and learner uptake: Negotiation of form in communicative classrooms. *Studies in Second Language Acquisition*, 19(1):37–66.

Shervin Malmasi, Keelan Evanini, Aoife Cahill, Joel Tetreault, Robert Pugh, Christopher Hamill, Diane Napolitano, and Yao Qian. 2017. A report on the 2017 native language identification shared task. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 62–75, Copenhagen, Denmark. Association for Computational Linguistics.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernocký, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010*, volume 2, pages 1045–1048.

Tomoya Mizumoto, Mamoru Komachi, Masaaki Nagata, and Yuji Matsumoto. 2011. Mining revision log of language learning SNS for automated Japanese error correction of second language learners. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 147–155, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Natawut Monaikul and Barbara Di Eugenio. 2020. Detecting preposition errors to target interlingual errors in second language writing. In *FLAIRS Conference*, pages 290–293.

Maria Nadejde and Joel Tetreault. 2019. Personalizing grammatical error correction: Adaptation to proficiency level and L1. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 27–33, Hong Kong, China. Association for Computational Linguistics.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The CoNLL-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14, Baltimore, Maryland. Association for Computational Linguistics.

Hwee Tou Ng, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault. 2013. The CoNLL-2013 shared task on grammatical error correction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–12, Sofia, Bulgaria. Association for Computational Linguistics.

Diane Nicholls. 2003. The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT. In *Proceedings of the Corpus Linguistics 2003 conference*, volume 16, pages 572–581.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Prokopis Prokopidis, Vassilis Papavassiliou, and Stelios Piperidis. 2016. Parallel global voices: a collection of multilingual corpora with citizen media stories. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 900–905.

Ella Rabinovich, Yulia Tsvetkov, and Shuly Wintner. 2018. Native language cognate effects on second language lexical choice. *Transactions of the Association for Computational Linguistics*, 6:329–342.

Daniel Robertson. 2000. Variability in the use of the English article system by Chinese learners of English. *Second Language Research*, 16(2):135–172.

Alla Rozovskaya and Dan Roth. 2011. Algorithm selection and model adaptation for ESL correction tasks. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 924–933, Portland, Oregon, USA. Association for Computational Linguistics.

Joel Tetreault, Daniel Blanchard, and Aoife Cahill. 2013. A report on the first native language identification shared task. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 48–57, Atlanta, Georgia. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Leticia Farias Wanderley and Carrie Demmans Epp. 2020. Identifying negative language transfer in writing to increase English as a Second Language learners' metalinguistic awareness. In *Companion Proceedings 10th International Conference on Learning Analytics & Knowledge (LAK20)*.

Leticia Farias Wanderley, Nicole Zhao, and Carrie Demmans Epp. 2021. Negative language transfer in learner English: A new dataset. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. In press.

David Alistair Watts. 2019. RAISING AWARENESS OF LANGUAGE TRANSFER WITH ACADEMIC ENGLISH WRITING STUDENTS AT A JAPANESE UNIVERSITY. *Frontier of Foreign Language Education*, 2:191–200.

Sze-Meng Jojo Wong and Mark Dras. 2009. Contrastive analysis and native language identification. In *Proceedings of the Australasian Language Technology Association Workshop 2009*, pages 53–61, Sydney, Australia.

Nianwen Xue and Yaqin Yang. 2011. Chinese sentence segmentation as comma classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 631–635.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189, Portland, Oregon, USA. Association for Computational Linguistics.