Native-like Expression Identification by Contrasting Native and Proficient Second Language Speakers

Oleksandr Harust* **Yugo Murawaki** **Sadao Kurohashi**
Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo-ku, Kyoto, 606-8501, Japan
oharust@gmail.com {murawaki, kuro}@i.kyoto-u.ac.jp

Abstract

We propose a novel task of native-like expression identification by contrasting texts written by native speakers and those by proficient second language speakers. This task is highly challenging mainly because 1) the combinatorial nature of expressions prevents us from choosing candidate expressions a priori and 2) the distributions of the two types of texts overlap considerably. Our solution to the first problem is to combine a powerful neural network-based classifier of sentence-level nativeness with an explainability method that measures an approximate contribution of a given expression to the classifier's prediction. To address the second problem, we introduce a special label `neutral` and reformulate the classification task as complementary-label learning. Our crowdsourcing-based evaluation and in-depth analysis suggest that our method successfully uncovers linguistically interesting usages distinctive of native speech.

1 Introduction

We propose a novel task of native-like expression identification (NLEI) by contrasting texts written by native and proficient second language (L2) speakers. Our primary motivation lies in the observation that native English speakers often fail to be understood even by proficient L2 speakers (Hazel, 2016). Take the following sentence for example:

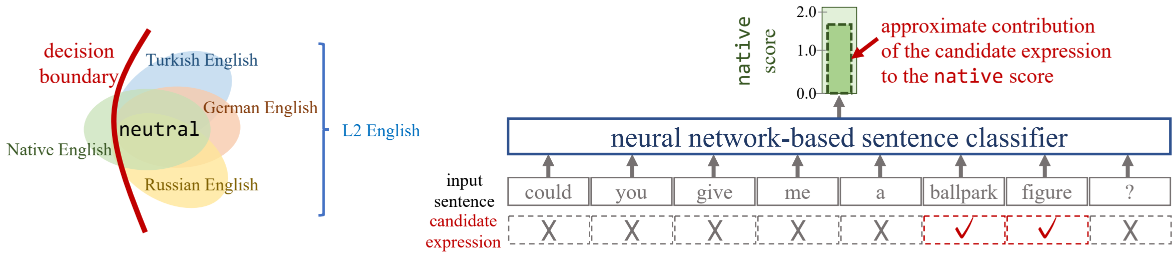
Could you give me a ballpark figure?

Ballpark figure is a fairly common American English idiom meaning "an approximate figure or quantity". Despite being a simple combination of two basic words, this expression is enigmatic to many L2 speakers, leading to communication breakdowns. We will refer to such expressions as native-like expressions. We believe that native speakers, no less than L2 speakers, would do well to adapt their speech in an international setting so as to maximize mutual comprehension, and that one effective strategy is to avoid native-like expressions of this sort. However, the hegemony of English in international communities has led to many monolingual English speakers lacking the notion of what it is like to speak a second language, making it particularly difficult for them to identify native-like expressions on their own. For this reason, a system that automatically detects native-like expressions would help native speakers improve their international outlook.

NLEI itself has other potential applications. It could prove useful as a tool for advanced language learners to find new, fluent expressions to acquire in any given text. It can also be used as a method of linguistic inquiry for examining the differences between first- and second-language acquisition. If we manage to draw cross-lingually valid generalizations from English data, they would have a significant impact. In many other languages, there has been an increasing recognition of the importance of effective communication with linguistic minorities, especially in emergency situations (Uekusa, 2019), but texts written by L2 speakers are not large enough to enable data-driven approaches.

[*] Oleksandr Harust now works at sinops Inc., Osaka, Japan.

Proceedings of the 28th International Conference on Computational Linguistics, pages 5843–5854
Barcelona, Spain (Online), December 8-13, 2020



(a) (b)

Figure 1: Overview of our task. (a) A schematic illustration of texts written by native and L2 English speakers. We assume that the two types of distributions are different but nevertheless overlap considerably. We work on identifying a subregion characteristic of native speakers. (b) The proposed method. We train a neural network-based sentence classifier that gives a `native` score to a given input. After that, we feed a candidate expression, in addition to the input sentence, to the classifier to approximately divide the `native` score into the contribution made by the candidate expression and that made by the rest of the sentence.

NLEI poses two key technical challenges. First, native-like expressions are often context-sensitive, and more importantly, can consist of several words. In traditional word-based approaches, simple frequency-based statistical tests such as a chi-squared test (Baayen, 2008) can be used to contrast text written by native and proficient L2 speakers. Such statistical tests are not applicable to our task because the combinatorial nature of native-like expressions prevents us from choosing candidate expressions a priori. Second, the distributions of the two types of texts cannot totally be separated because, as illustrated in Figure 1(a), they do overlap considerably. This hampers discriminative approaches to the task that are known to be more powerful than simple statistical tests.

To address the first challenge, we combine a neural network with an explainability method. We build a sentence-level classifier using a BiLSTM with subword embeddings as input, with the hope that it is expressive enough to implicitly exploit native-like expressions for classification. After narrowing down the targets to native-like sentences, we use a method named contextual decomposition (Murdoch et al., 2018; Singh et al., 2019) to measure an approximate contribution of a given expression to the classifier's prediction (Figure 1(b)). We repeat this for multiple candidate expressions in a given sentence to choose the most appropriate one.

We address the inseparability problem by introducing a special label `neutral` to indicate the overlapping region. The resulting three-way classifier can be trained under the framework of complementary-label learning (Ishida et al., 2017; Yu et al., 2018). By letting non-distinctive sentences be absorbed into `neutral`, the classifier is able to choose distinctively native-like sentences from sentences written by native speakers.

Due to the exploratory nature of the task, it is hard to evaluate the proposed method. Nevertheless, we performed crowdsourcing for quantitative evaluation, which was followed by in-depth manual analysis of the detected expressions. We found that the scores given by the proposed method weakly correlated with aggregated ratings provided by L2 crowdworkers. Remarkably, the proposed method often identified expressions that consisted of words so basic that the traditional word-based models would have deemed them easy for L2 speakers.

2 Related Work

2.1 Word-based Models for Second Language Learners

A growing body of research adopts NLP techniques to assist second language learners. While grammatical error correction (Dale and Kilgarriff, 2011; Ng et al., 2013; Ng et al., 2014) has arguably been the most actively studied, a number of researchers have also worked on identifying words that are difficult for learners.

The existing approaches to modeling words can be grouped into type-based and token-based ones.

The goal of type-based approaches is to estimate learners' vocabulary proficiency. To predict whether a learner knows a given word, logistic models based on item response theory are often used (Ehara et al., 2012; Ehara et al., 2013). Token-based approaches are formalized as complex word identification (CWI) (Paetzold and Specia, 2016; Yimam et al., 2018). CWI is a task that aims to identify words in texts that might present a challenge for target readers, who are often but not always second language learners.

Our task is closer to CWI in that both tasks are designed to handle context sensitivity: depending on surrounding context, a given expression can convey different meanings and hence can be easy or difficult. As the name suggests, however, the focus of CWI is on words, although phrases are not entirely absent from the data. Our departure from word-based models is motivated by skepticism about the idea that simpler words result in better comprehension. Another major difference is that while CWI is usually framed as a supervised learning task where words in texts are annotated regarding complexity, we assume that in our task, only writer attributes are available as indirect supervision signals. Finally, our overarching goal is different from those of learner-oriented studies in that we aim to change the behavior of native speakers, rather than L2 speakers, in the settings of international communication.

2.2 Controlled Languages and Text Simplification

There have been a number of attempts to create lexically and grammatically restricted subsets of English, grouped under the umbrella term "controlled languages" (CL). Although a large portion of CLs are domain-specific, there also exist general-purpose CLs, most notably Basic English (Ogden, 1930) — one of the oldest and most influential controlled languages. The vocabulary of Basic English is stripped down from regular English to 850 word forms, with verbs being especially restricted to just 18.

The largest collection of texts that is claimed to use Basic English is the Simple English Wikipedia (SEW). SEW contributors purport to adhere to the principle of using "simpler" vocabulary and avoiding idioms,[1] but this is not strict, and the case-by-case judgments are largely left to writers' and editors' discretion. The surprising unsimplicity of SEW has been pointed out by Xu et al. (2015). Another study concludes that in practice the vocabulary richness of SEW is the same as in regular English Wikipedia (EW), and that the decrease in complexity is mostly due to usage of shorter sentences, while syntax itself is not drastically simplified (Yasseri et al., 2012). Finally, most of the editors seem to be native speakers of English, and the SEW is apparently failing to reach its target audience of L2 speakers, students, and developmentally disabled people.

SEW, aligned with EW, is often used for the task of text simplification (Alva-Manchego et al., 2020). A popular subtask of text simplification is lexical simplification, where difficult words are substituted with simple words. In fact, CWI is often treated as a prerequisite of lexical simplification. Although we limit ourselves to detection, paraphrasing is an interesting direction to explore. It should be noted again that our focus on native speakers and longer expressions makes our task unique and distinctive.

2.3 Native Language Identification

From a technological point of view, the proposed method has a close connection to native language identification (NLI) (Koppel et al., 2005; Tetreault et al., 2013; Malmasi et al., 2017; Goldin et al., 2018). In its simplest form, NLI is formalized as a binary classification task where the goal is to determine whether the writer is a native or L2 speaker. Goldin et al. (2018) worked on an English corpus in which L2 speakers were highly advanced, almost at the level of native speakers (Rabinovich et al., 2018). As we see in Section 5.1, we use this dataset in our task.

Although we also train a classifier, an important difference from NLI is that classification is not our goal but an intermediate task. While Goldin et al. (2018) used as the input a text chunk large enough to reveal the writer's identity, we classify single sentences in order to narrow down potential occurrences of native-like expressions. To this end, we are eager to eliminate the impact of extralinguistic patterns as much as possible while Goldin et al. (2018) used them to improve the classification performance. For

[1] https://simple:wikipedia:org/wiki/Wikipedia:HowtowriteSimpleEnglishpages

example, the place name *New Orleans* is suggestive of American identity. In our task, however, it is to be masked because it is not a native-like expression.

3 Native-like Expression Identification

Given texts written by native and proficient L2 speakers, the task of NLEI is to find any possible native-like expressions in them. We assume that each sentence is tied to a writer-attribute label, $\texttt{native}_{\text{writer}}$ or $\texttt{L2}_{\text{writer}}$. It is important to note that we assume no annotation of native-like expressions themselves.

Due to the exploratory nature of the task, it is difficult to give a precise definition to native-like expressions. Any word or sequence of words is considered native-like if it is more commonly used by native speakers compared to L2 speakers.[2] An additional condition is that native-like expressions must not include domain-specific words or named entities. Since topics that native speakers write about are often different from L2 speakers, domain-specific expressions such as *baseball*, and *electoral wipeout*, as well as country names and so on, could help reliably distinguish the two groups. However, since they do not constitute distinctive usages that we are aiming to identify, we do not treat them as native-like expressions.

There are no clear-cut criteria for determining the boundary of a native-like expression. From the sentence,

> *What on Earth are you yammering on about?*

some might want to extract *yammering on* while others may prefer *yammering on about*. We decide to select at most one native-like expression per sentence to sidestep the overlap problem. We also adopt a relaxed matching criterion for evaluation.

4 Proposed Method

We build a neural network-based classifier that receives a sentence as the input and returns a three-dimensional vector representing the labels `native`, `neutral`, and `L2` (Section 4.1). It is trained under the framework of complementary-label learning (Section 4.2). The classifier is then combined with an explainability method named contextual decomposition to locate a native-like expression in a given native-like sentence (Section 4.3).

4.1 BiLSTM Classifier of Sentence-level Nativeness

For a given sentence with tokens $x_{1:N} = (x_1, \cdots, x_N)$, the classifier outputs a three-dimensional score vector $\mathrm{y} = (y_{\texttt{native}}, y_{\texttt{neutral}}, y_{\texttt{L2}}) \in \mathbb{R}^3$. We build the classifier with a BiLSTM (Graves and Schmidhuber, 2005). After transforming the input into an embedding sequence $e_{1:N} = (e_1, \cdots, e_N)$, we feed the vectors into the BiLSTM to obtain context-aware representations:

$$h_i = \mathrm{BiLSTM}_i(e_{1:N}).$$

We then apply the average pooling and two linear transformations with a hyperbolic tangent activation function to obtain the output vector:

$$\mathrm{y} = \mathrm{Linear}(\tanh(\mathrm{Linear}(\mathrm{AvgPooling}(h_{1:N})))).$$

For classification, we select $\mathrm{argmax}_i\, y_i \in \mathrm{y}$, but we are more interested in the vector y itself.

[2] Note here that even though we assume that a lack of L2 *production* would often indicate potential L2 *comprehension* problems, we do not claim the two to be equivalent: an L2 speaker does not necessarily fail to understand a given native-like expression they are unlikely to produce themselves: it might well be contained in their passive vocabulary, or perhaps inferred from the context. Identifying passive vocabulary from a static corpus is generally very hard, and NLEI is a reasonable first step toward detecting potential obstacles to international communication from the abundant corpora of L1 and L2 productions.

4.2 Complementary-Label Learning

Even though we try to build a three-way classifier, we only have access to two writer-attribute labels, $\texttt{native}_{\text{writer}}$ and $\texttt{L2}_{\text{writer}}$. The trick we employ here is called complementary-label learning (Ishida et al., 2017; Yu et al., 2018). A complementary label specifies a class to which the input does *not* belong. In out task, the writer-attribute label $\texttt{native}_{\text{writer}}$ is recast as a complementary sentence label `not-L2`, meaning that the input may belong to either `native` or `neutral` but certainly not to `L2`. Similarly, the writer-attribute label $\texttt{L2}_{\text{writer}}$ is mapped to the complementary sentence label `not-native` (either `neutral` or `L2`). Our setting is unusual for a complementary-label learning task in that we never observe `not-neutral`, but it does work in practice.[3]

To define the loss, we normalize y and project the result into the space of complementary labels, `not-native`, `not-neutral`, and `not-L2` in this order:

$$\mathrm{g} = Q \cdot \mathrm{Softmax}(\mathrm{y}),$$

where Q is a transition matrix,

$$Q = \begin{pmatrix} 0 & 0.5 & 1-d \\ d & 0 & d \\ 1-d & 0.5 & 0 \end{pmatrix}.$$

d is a small discount factor which we found stabilized training. For simplicity, we assume that sentences labeled `not-native` and `not-L2` are equal in size although it is not difficult to handle data imbalance. We compute the cross-entropy loss of g with respect to the complementary label and perform backpropagation to update the parameters.

To gain an insight, suppose that the classifier vacillates between `native` and `neutral` during training. Q moves the (normalized) `native` score directly to `not-L2` (i.e., $\texttt{native}_{\text{writer}}$) while the `neutral` score is evenly distributed to `not-native` ($\texttt{L2}_{\text{writer}}$) and `not-L2` ($\texttt{native}_{\text{writer}}$). Since the reference label is either `not-native` or `not-L2`, the `neutral` score leads to a stable, moderate loss. That makes `native` (and `L2`) a relatively high-risk/high-return bet, yielding either a large loss or a small loss. Thus, the classifier ends up giving a relatively large score to `neutral` when the input is not distinctive.

4.3 Contextual Decomposition for Finding Native-like Expressions

Once the classifier chooses sentences with high `native` scores, we want to locate native-like expressions in them. Let $S \subseteq \{1, \cdots, N\}$ be a subset of the input that we consider as a candidate expression. We use contextual decomposition (CD) (Murdoch et al., 2018; Singh et al., 2019) to calculate the approximate contribution of S to the `native` score. We repeat this for multiple candidate expressions in a given sentence to choose an appropriate one.

The key idea of CD is that if a decomposition operation is defined for every neural network layer, we can propagate the decomposed input to a decomposed output by simply tracing the forward computation. For a vector v going inside the network, let $\beta(v)$ and $\gamma(v)$ be the contributions of S and $\overline{S}$, respectively ($v = \beta(v) + \gamma(v)$). The decomposition is pretty straightforward for the embedding layer: $\beta(e_i) = e_i$ and $\gamma(e_i) = \mathbf{0}$ if $i \in S$; otherwise $\beta(e_i) = \mathbf{0}$ and $\gamma(e_i) = e_i$.

For a linear layer with a weight matrix W and a bias b, the input $v^{L-1} = \beta(v^{L-1}) + \gamma(v^{L-1})$ is transformed into the output $v^L = \beta(v^L) + \gamma(v^L)$ as follows.

$$\beta(v^L) = W\beta(v^{L-1}) + \frac{|W\beta(v^{L-1})|}{|W\beta(v^{L-1})| + |W\gamma(v^{L-1})|} \cdot b$$

$$\gamma(v^L) = W\gamma(v^{L-1}) + \frac{|W\gamma(v^{L-1})|}{|W\beta(v^{L-1})| + |W\gamma(v^{L-1})|} \cdot b$$

[3] Complementary-label learning was originally formalized by Ishida et al. (2017), but we adopt a variant method proposed by Yu et al. (2018), who removed the assumption that complementary labels were uniformly chosen during data construction.

The first terms are, again, straightforward: the weight matrix is multiplied individually by $\beta(v^{L-1})$ and $\gamma(v^{L-1})$. The remaining problem is how to partition the bias term. Singh et al. (2019) found that partitioning the bias in proportion to the absolute values of the first terms empirically worked well.

The hyperbolic tangent is non-linear, but the following formulae provide a good linearization of the relationship between the input $v^{L-1} = \beta(v^{L-1}) + \gamma(v^{L-1})$ and the output $v^L = \beta(v^L) + \gamma(v^L)$.

$$\begin{aligned}\beta(v^L) =& \frac{1}{2}\left(\tanh(\beta(v^{L-1})) + \tanh(\beta(v^{L-1}) + \gamma(v^{L-1})) - \tanh(\gamma(v^{L-1}))\right)\\ \gamma(v^L) =& \frac{1}{2}\left(\tanh(\gamma(v^{L-1})) + \tanh(\beta(v^{L-1}) + \gamma(v^{L-1})) - \tanh(\beta(v^{L-1}))\right)\end{aligned}$$

Murdoch et al. (2018) elaborate a decomposition operation for the LSTM. It is easy to extend the operation to the BiLSTM because the concatenation of the forward and backward hidden vectors is linear. The remaining layer, average pooling, is also linear. In this way, the output y is decomposed into $\beta(\mathrm{y})$ and $\gamma(\mathrm{y})$. We use $s_{\mathrm{CD}} = \beta_{\texttt{native}}(\mathrm{y})$ as the CD score of the candidate expression.

5 Experiments

5.1 Data and Preprocessing

We used the L2-Reddit corpus (Rabinovich et al., 2018), a collection of native and L2 English sentences extracted from Reddit, a community-driven discussion website. On some portions of Reddit, a user can indicate his/her country of origin with a metadata tag; it can be used to assume the user's native language.[4] All submissions by such users were extracted and split into sentences, resulting in a corpus of over 250 million sentences produced by 45,000 users from 50 countries. The label $\texttt{native}_{\text{writer}}$ was assigned to sentences produced by users from Australia, Canada, Ireland, New Zealand, the United Kingdom, and the U.S. while the remaining sentences were given the label $\texttt{L2}_{\text{writer}}$.

A notable feature of the L2-Reddit corpus is that the proficiency level of L2 English utterances tends to be very high: spelling and grammar mistakes are very uncommon, and the syntactic complexity and use of colloquialisms makes L2 nearly indistinguishable from native-produced utterances. This gives us reason to hope that the expressions that get large native scores from our method will provide an insight into the distinctive features of native English, since in most other respects the utterances are very similar.

To reduce the impact of noisy data on the classifier, we removed certain kinds of sentences from the dataset: *a*) very long sentences (more than 80 tokens), *b*) very short sentences (less than 6 tokens), *c*) repeating sentences (found more than 5 times in the corpus), *d*) sentences containing too much punctuation (more than 15% of total tokens in the sentence), and *e*) sentences containing too many named entities (see below; more than 25% of total tokens in the sentence).

To further minimize differences between the native and L2 sentences in the corpus, we masked any named entities — most often proper nouns such as *Bay Area*, *Canadians*, *Angela Merkel*, etc. This would have the effect of reducing the classifier's ability to rely on country-specific topical content and other "easy" clues to as to whether a sentence is native-produced or not. The named entities were detected using several algorithms from the Stanford CoreNLP package (Manning et al., 2014).

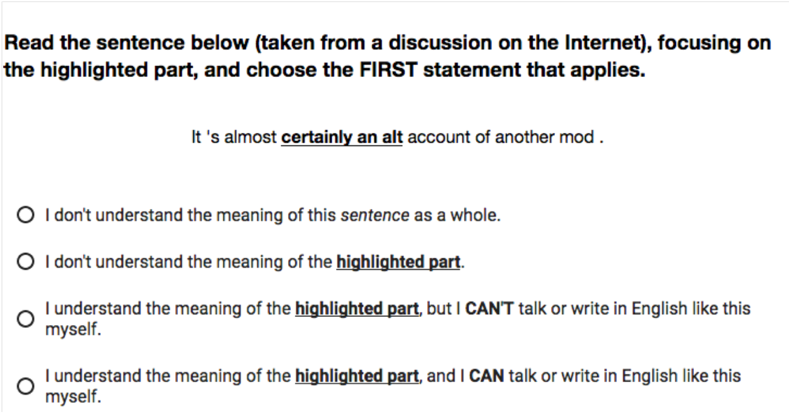
As a result of preprocessing, the original corpus of some 240M sentences was reduced to 146M sentences (or about 73M each of native and L2 sentences). The training, validation, and test subsets amounted, respectively, to 95%, 1%, and 4% of the final dataset.

5.2 Details of the Proposed Method

We tokenized sentences into WordPiece (Wu et al., 2016) subword tokens with a 30,000 token vocabulary, and used the pre-trained 768-dimensional embeddings provided by Devlin et al. (2019), available through the transformers[5] package. We did not update the embeddings during training because,

[4] Even though there may be a mismatch for some users (a user resides in the U.S. but is not a native speaker), Rabinovich et al. (2018) provide proof that in most cases, user-specified country is a sufficient proxy for native language.

[5] https://github.com/huggingface/transformers



Read the sentence below (taken from a discussion on the Internet), focusing on the highlighted part, and choose the FIRST statement that applies.

It 's almost **certainly an alt** account of another mod .

○ I don't understand the meaning of this *sentence* as a whole.

○ I don't understand the meaning of the **highlighted part**.

○ I understand the meaning of the **highlighted part**, but I **CAN'T** talk or write in English like this myself.

○ I understand the meaning of the **highlighted part**, and I **CAN** talk or write in English like this myself.

Figure 2: The task screen for an L2 crowdworker.

somewhat surprisingly, this setting yielded more intuitively plausible CD scores than a fine-tuned model and a cold-start model. We used a BiLSTM with 200-dimensional hidden vectors, the Adam optimizer (Kingma and Ba, 2015) with the learning rate of 0.001, the batch size of 1024 sentences, and the discount parameter d of 0.001. We selected the model state after 3 epochs of training because it got the best validation score.

The classifier labeled the sentences in the test data as follows: 8.2% for native, 81.7% for neutral, and 10.2% for L2. This result was consistent with our observation that the distributions of sentences written by native and L2 speakers had a huge overlap. Removing sentences labeled neutral, we focused on classification performance for the remaining sentences. The output label native was judged correct if the given sentence was tied to native$_{\text{writer}}$. For the native label, we obtained the precision of 79.5%, the recall of 81.1%, and the F1 score of 80.3%. In preliminary experiments, we confirmed that a baseline binary classifier of the writer-attribute labels topped out just around 60% in terms of accuracy. The large jump in performance indicates that the proposed three-way classifier successfully identified distinctively native-like sentences by letting non-distinctive sentences be absorbed into neutral.

Next, we identified native-like expression from sentences in the training data. We began by selecting native$_{\text{writer}}$ sentences classified as native. There were multiple ways to enumerate candidate expressions, including systems for constituency parsing and multi-word expression identification. For simplicity, however, we opted for extracting all up-to-5-grams (unigrams, bigrams, ... 5-grams) within the sentence. We selected sentences for which the largest CD score was no less than 0.7. In many cases, more than one candidate expression in a sentence surpassed this threshold. To avoid the overlap problem, and for reasons of evaluation convenience, we extracted only the first candidate. With the threshold of 0.7 for the CD scores, about 71% of the native-like sentences remained.

5.3 Crowdsourcing-based Evaluation

To quantitatively evaluate the results, we asked L2 crowdworkers to rate the detected expressions. Before the actual evaluation, we conducted additional rule-based filtering and some screening by native speakers of English, as explained in detail in Appendix A.

We hired crowdworkers on the Amazon Mechanical Turk platform. Since this platform did not provide an option to filter workers by native language, we resorted to using country of residence as a proxy. We only accepted answers from workers *not* residing in Australia, Canada, Ireland, New Zealand, the United Kingdom, or the U.S., the same six countries chosen for the L2-Reddit corpus.

The task screen is shown in Figure 2. For each sentence, five crowdworkers were asked to read the sentence with the relevant expression highlighted and to answer a multiple-choice question, roughly ranging from least to most familiar: hard-sent, hard, known, and used as shorthand.

It turned out that the L2 crowdworkers were, or at least pretended to be, highly proficient in English. The distribution of the answers was skewed toward used: 3.8% for hard-sent, 8.9% for hard, 10.0% for known, and 78.3% for used. Ignoring answers with hard-sent, we aggregated multiple answers for each sentence into a single scalar value:

$$s_{\text{L2}} = \frac{0 \times n_{\texttt{hard}} + 1 \times n_{\texttt{known}} + 2 \times n_{\texttt{used}}}{n_{\texttt{hard}} + n_{\texttt{known}} + n_{\texttt{used}}},$$

where n_* is the corresponding answer counts. The lower s_{L2} is, the less familiar the expression is to L2 speakers.

For comparison, we employed L2 vocabulary knowledge data collected by Ehara et al. (2013). Built on top of the Standard Vocabulary List (SVL12000), a list of fundamental words for English learners, this dataset was constructed by asking 16 learners of English to rate their degree of knowledge of each word (lower is less familiar). For each word, we took an average of the 16 scores. To apply the word-based rating to expressions, we chose the word with the lowest score (i.e., least familiar) and took that to be the score for the entire expression. We refer to this score as s_{word}.

Note that we performed lemmatization for word lookup. We found this process error prone, as there were many false negatives. For a fair comparison, we dropped expressions for which we failed to find any words in the list, leading to 7% reduction of the annotated dataset.

s_{CD} exhibited a weak negative correlation with s_{L2}, with Pearson's $r = -0.26$. For comparison, s_{word} had a stronger correlation of 0.37, also in the range of weak correlation. s_{word}'s relatively strong performance is understandable given that, albeit word-based, it reflects direct human supervision while the proposed method only exploits sentence-level signals. It is interesting that s_{CD} correlated very weakly with s_{word} ($r = -0.09$), indicating that the two methods explored very different phenomena. We investigate this point in the next section.

5.4 In-depth Analysis

We manually evaluated a random sample of 100 sentences, from those for which an expression with the CD score at or above 0.7 was found. According to our subjective analysis, 47 expressions were good (i.e., contained linguistically interesting usages), and the remaining 53 expressions were classified into three groups: 36 for domain-specific expressions, 5 for named entity detection failure, and 12 for others.[6]

Looking first at the not-good expressions, the domain-specific expressions constituted the largest group. These expressions tended to consist of (or contained) words and collocations related to sports, local politics, legal matters, academic affairs, and other subjects that the native speakers in the data were more likely to talk about:

> *No, its pretty damn conclusive, that gun control and gun ownership have no effect on a nations violence rate.*

The debate over *gun control and gun ownership* is a uniquely American topic, but the phrase itself is not linguistically interesting because its meaning is transparent to L2 speakers. Unlike named entities, common nouns can hardly be masked at the preprocessing step. A possible solution to this problem is to perform domain-adversarial training (Ganin et al., 2016) in a topic-aware manner. Additional metadata provided by Reddit may help.

There were also occasional named entities that were not detected by the NER pipeline, often due to lack of proper capitalization in the original sentence. Most of these are of little interest to us:

> *Oh no, the packers are infighting?* (team name)

The last group included the expressions that did not fit neatly into the other two categories. Some of these were hard to understand or interpret qualitatively (and may thus be treated as noise), while others simply did not seem interesting enough to be considered good.

As for good expressions, we found only 10 out of 47 expressions containing difficult-looking words:

[6] Again, we want to stress that both the sentences and the underlined expressions in the examples given were detected entirely automatically: from among all possible up-to-5-grams within a given $\texttt{native}_{\text{writer}}$ sentence that was classified as native by the classifier in Sections 4.1 to 4.2, the underlined expression is the one with the largest CD score — calculated as described in Section 4.3.

They started with a wildly lopsided pitching match-up and kept it respectable.

Seems to be privy to things that normal bloggers are not.

This is a bit surprising but consistent with the very weak correlation between s_{CD} and s_{word}.

Indeed, many expressions detected by the proposed method were combinations of basic words:

(1) *Good luck with it, but I think the location is going to be a hard sell.*

(2) *Do you find that the objectionable ones get better as the glasses and bottles sit out a bit?*

(3) *The ATM at my old work would occasionally toss out bills stuck together.*

It is no wonder that s_{word} wrongly indicated high familiarity for the underlined expressions. Even if not incomprehensible, these expressions appear to require some effort from L2 speakers to be understood. In the case of sentence (1) above, the expression contains a less-common idiom *to be a hard sell*, which in this case means "something that it is difficult to persuade people to buy or accept".[7] In this case, context will need to be taken into account for correct comprehension.

Sentence (2) is harder still: the phrasal verb *sit out* seems to be used creatively to mean that the glasses and bottles, ostensibly containing some sort of beverage, need to stay undisturbed for a while. This idiosyncratic usage, compounded with *a bit* denoting a relatively short span of time, can prove exceptionally hard for an L2 speaker to "parse out", especially if part of a spoken utterance.

Moreover, it is highly unlikely that L2 speakers can easily imitate the way the basic words are combined. To express a given idea, they would have chosen words with more prototypical meanings in mind and combined them in a more semantically compositional (Crossley and McNamara, 2009) and syntactically straightforward (Wray, 1999) manner. The expression in sentence (3) contains both a creative choice of phrasing – *toss out* instead of, say, *return*, and a reduced relative clause attached to a noun – *bills stuck together*. This latter construction often causes garden-path effects in reading comprehension, and is likely to present difficulties for L2 speakers especially (Juffs, 1998).

In some cases, the detected expressions were not satisfactory with respect to coverage, but nevertheless, the sentences themselves were quite native-like:

Throwing out more than what you want in a negotiation doesn't take smarts.

Its used for casual diving off New Jersey but is currently at quite a depth.

Again, none of the words in these sentences appear difficult in isolation but the way they are used is. To sum up, our method successfully uncovers linguistically interesting usages distinctive of native speech that would slip through unnoticed by traditional word-based models.

6 Conclusion

In this paper, we proposed a novel task of native-like expression identification. We showed that even if no expression-level annotation is available, a powerful neural network combined with an explainability method can detect native-like expressions by contrasting sentences written by native and proficient L2 speakers. Our analysis on expressions detected by the proposed method suggests that even basic words may give some pause to L2 speakers if they are arranged in certain ways.

Although the proposed method in its current form is sufficiently useful for mining native-like expressions, further work is needed to mitigate the effect of domain-specific expressions. An interesting future direction is to create a user interface through which the system gets human feedback. Such direct signals are expected to make the detected expressions more closely match human intuitions (Rieger et al., 2019).

Acknowledgements

We would like to thank Dr. Yo Ehara for his advice on choosing English vocabulary datasets. This work was partly supported by JSPS KAKENHI Grant Number 18K18104.

[7] https://www.ldoceonline.com/dictionary/a-hard-tough-sell

References

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020. Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, 45(1):135–187.

R. H. Baayen. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics using R*. Cambridge University Press, Cambridge.

Scott A Crossley and Danielle S McNamara. 2009. Computational assessment of lexical differences in L1 and L2 writing. *Journal of Second Language Writing*, 18(2):119–135.

Robert Dale and Adam Kilgarriff. 2011. Helping our own: The HOO 2011 pilot shared task. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 242–249, Nancy, France, September. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Yo Ehara, Issei Sato, Hidekazu Oiwa, and Hiroshi Nakagawa. 2012. Mining words in the minds of second language learners: Learner-specific word difficulty. In *Proceedings of COLING 2012*, page 799–814, Mumbai, India, December. The COLING 2012 Organizing Committee.

Yo Ehara, Nobuyuki Shimizu, Takashi Ninomiya, and Hiroshi Nakagawa. 2013. Personalized reading support for second-language web documents. *ACM Transactions of Intelligent Systems and Technology*, 4(2):31:1–31:19, April.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(1):2096–2030.

Gili Goldin, Ella Rabinovich, and Shuly Wintner. 2018. Native language identification with user generated content. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3591–3601.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5-6):602–610.

Spencer Hazel. 2016. Why native English speakers fail to be understood in English – and lose out in global business. `https://theconversation.com/why-native-english-speakers-fail-to-be-understood-in-english-and-lose-out-in-global-business-54436`, 11. Accessed: 2020-06-27.

Takashi Ishida, Gang Niu, Weihua Hu, and Masashi Sugiyama. 2017. Learning from complementary labels. In *Advances in neural information processing systems*, pages 5639–5649.

Alan Juffs. 1998. Main verb versus reduced relative clause ambiguity resolution in L2 sentence processing. *Language Learning*, 48(1):107–147.

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

Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 2005. Automatically determining an anonymous author's native language. In Paul Kantor, Gheorghe Muresan, Fred Roberts, Daniel D. Zeng, Fei-Yue Wang, Hsinchun Chen, and Ralph C. Merkle, editors, *Intelligence and Security Informatics*, pages 209–217, Berlin, Heidelberg. Springer Berlin Heidelberg.

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, September. Association for Computational Linguistics.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

W. James Murdoch, Peter J. Liu, and Bin Yu. 2018. Beyond word importance: Contextual decomposition to extract interactions from LSTMs. In *International Conference on Learning Representations*.

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, August. 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, June. Association for Computational Linguistics.

Charles Kay Ogden. 1930. *Basic English: A general introduction with rules and grammar*. Paul Treber.

Gustavo Paetzold and Lucia Specia. 2016. SemEval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569.

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.

Laura Rieger, Chandan Singh, W James Murdoch, and Bin Yu. 2019. Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. *arXiv*, 1909.13584.

Chandan Singh, W. James Murdoch, and Bin Yu. 2019. Hierarchical interpretations for neural network predictions. In *International Conference on Learning Representations*.

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, June. Association for Computational Linguistics.

Shinya Uekusa. 2019. Disaster linguicism: Linguistic minorities in disasters. *Language in Society*, 48(3):353–375.

Alison Wray. 1999. Formulaic language in learners and native speakers. *Language teaching*, 32(4):213–231.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv*, 1609.08144.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Taha Yasseri, Andras Kornai, and Janos Kertesz. 2012. A practical approach to language complexity: a Wikipedia case study. *PloS one*, 7(11).

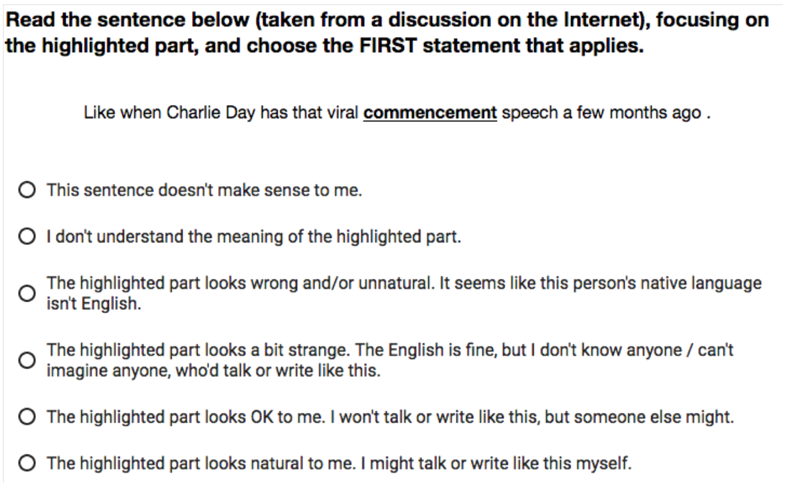
Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo H Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. A report on the complex word identification shared task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 66–78.

Xiyu Yu, Tongliang Liu, Mingming Gong, and Dacheng Tao. 2018. Learning with biased complementary labels. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 68–83.

Appendix A Preliminaries to Crowdsourcing-based Evaluation

In preparation to the crowdsourcing-based evaluation, we filtered out noisy and incomprehensible sentences by conducting additional rule-based filtering and a screening by native speakers of English.

As stated in Section 5.2, we set the threshold of 0.7 for the CD scores and we did apply this to the dataset to be annotated. For sanity check purposes, however, we initially created a more representative sample of the corpus. Because CD scores exhibited a skewed distribution, we did this by combining two sets: 1) We randomly selected 10,000 sentence-expression pairs for which the top CD score was greater



Read the sentence below (taken from a discussion on the Internet), focusing on the highlighted part, and choose the FIRST statement that applies.

Like when Charlie Day has that viral **commencement** speech a few months ago .

○ This sentence doesn't make sense to me.

○ I don't understand the meaning of the highlighted part.

○ The highlighted part looks wrong and/or unnatural. It seems like this person's native language isn't English.

○ The highlighted part looks a bit strange. The English is fine, but I don't know anyone / can't imagine anyone, who'd talk or write like this.

○ The highlighted part looks OK to me. I won't talk or write like this, but someone else might.

○ The highlighted part looks natural to me. I might talk or write like this myself.

Figure 3: The task screen for a native crowdworker.

than 0.5. 2) We created another set of 10,000 sentence-expression pairs for which the top CD score was not necessarily greater than 0.5 (i.e., we selected random expressions with random scores).

To make the data more appropriate for human evaluation, we removed the sentence-expression pairs in the following cases:

- when the sentence contained a profanity;
- when the sentence was longer than 30 tokens;
- when the expression contained a named entity;
- when the expression consisted of a single stopword[8] or a non-alphabetical character;

This left us with 10,877 sentences from the original 20,000. From these remaining sentences, we randomly selected 1,000 sentences.

Next, we asked native crowdworkers to check the sentences. We hired native crowdworkers on Amazon Mechanical Turk. Workers residing in the above-listed six countries were treated as native speakers of English. The task screen for native crowdworkers is shown in Figure 3. The instruction at the top was the same as that given to L2 workers, but this time we provided six options, roughly ranging from least natural to most natural.

As expected, a large majority of the expressions were familiar and natural to the native crowdworkers, as both the readers and the writers were native speakers. The workers chose the most-natural option in four out of five cases (3968 out of 5000 answers total). This indicates that most of the sentences and expressions in the selected subset were very familiar, natural-sounding, and easy to understand for native speakers of English.

Some of the sentences did get bad scores, however. Because we are not interested in sentences that even native speakers cannot make sense of, we removed these from the subsequent evaluation. Specifically, we removed 26 sentences for which at least two workers out of five gave the first three options. We also removed several sentences for which less than five answers were given (due to issues with the crowdsourcing interface). As a result, 958 sentences remained.

As described in Section 5.3, we removed sentences for which none of the words in the target expression matched SVL12000. This reduced the number of sentences to 890. Applying the CD score threshold of 0.7, we chose 287 sentences as the final dataset.

[8] *Stopwords* tend to be the most common words in the language: function words, pronouns, auxiliary verbs, and so on. We used a 127-word list from `https://gist.github.com/sebleier/554280`, apparently from an older version of the NLTK library for the Python language.