Towards an open-domain chatbot for language practice

Gladys Tyen[1], Mark Brenchley[2], Andrew Caines[1], Paula Buttery[1]
[1] ALTA Institute & Computer Laboratory, University of Cambridge, United Kingdom
[2] Cambridge University Press & Assessment, University of Cambridge, United Kingdom
{gladys.tyen,andrew.caines,paula.buttery}@cl.cam.ac.uk
mark.brenchley@cambridge.org

Abstract

State-of-the-art chatbots for English are now able to hold conversations on virtually any topic (e.g. Adiwardana et al., 2020; Roller et al., 2021). However, existing dialogue systems in the language learning domain still use handcrafted rules and pattern matching, and are much more limited in scope. In this paper, we make an initial foray into adapting open-domain dialogue generation for second language learning. We propose and implement decoding strategies that can adjust the difficulty level of the chatbot according to the learner's needs, without requiring further training of the chatbot. These strategies are then evaluated using judgements from human examiners trained in language education. Our results show that re-ranking candidate outputs is a particularly effective strategy, and performance can be further improved by adding sub-token penalties and filtering.

1 Introduction

Studies in second language acquisition have shown that interaction is an important aspect of language learning (e.g. Loewen and Sato, 2018; Plonsky and Oswald, 2014; Mackey, 2013; Long, 1996). However, interaction typically involves one-on-one sessions with a teacher, which can be costly or may simply be unavailable to some learners. In addition, learners may experience language anxiety during interaction, which can be detrimental to learning (Horwitz, 2001).

Artificial dialogue systems provide an alternative way to learn through interaction. Learners can chat with the system in their target language at their own convenience, without needing a teacher.

Existing systems typically rely on handcrafted rules, and require learners to practise within a specified context (e.g. shopping at a supermarket) (cf. Bibauw et al., 2019). They are therefore quite limited in scope and require much manual work to anticipate possible responses.

In our work, we leverage existing chatbot technology that can generate responses in virtually any topic (Roller et al., 2021). As a first step towards integrating this technology into language education, we experiment with ways to **adjust the difficulty** of chatbot messages to a specified level (e.g. one that matches the learner's proficiency level).

Our contributions are as follows:

1. We propose two types of decoding-based strategies for adjusting the difficulty of generated text – **vocabulary restriction** and **re-ranking** – as well as ways to augment them.
2. In total, we implemented 5 different variants of these strategies, and we release the code and demo at `https://github.com/WHGTyen/ControllableComplexityChatbot/`.
3. For our evaluation process, we generated self-chats from the chatbot and determined their difficulty level and quality. We release the annotated data alongside the code and demo.

2 Related work

We provide an overview of four related topics: 1) dialogue systems; 2) decoding strategies for adding various desired attributes to text; 3) text simplification methods for transforming *existing* text (instead of *generating* text at a specified difficulty level); and 4) methods for predicting linguistic complexity.

2.1 Dialogue systems

Dialogue systems can be classified into goal-oriented systems, which are designed for a specific task, or non-goal-oriented systems, designed for general "chit-chat" (Chen et al., 2017). In this paper, we focus on open-domain text systems for chit-chat, which allow learners to practise chatting in any topic they choose.

Early open-domain systems relied on pattern matching and rules (e.g. Weizenbaum, 1966; Car-

Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022), pages 234 - 249
July 15, 2022 ©2022 Association for Computational Linguistics

penter), but were somewhat limited in their conversational ability. More recent neural dialogue systems can produce a wider range of responses using generative models, retrieval-based models, or some combination of both (e.g. Papangelis et al., 2021; Adiwardana et al., 2020; Roller et al., 2021). Generative models produce entirely new sentences using beam search or similar decoding algorithms, while retrieval-based models select appropriate responses from an existing corpus.

Dialogue systems are also present in the Computer Assisted Language Learning (CALL) literature. Bibauw et al. (2019) present an investigation of dialogue-based CALL, where, notably, only 22 out of 96 systems allow completely free dialogue. Of those, most rely on handcrafted rules, and none make use of neural methods. To our knowledge, our work is the first attempt to use a neural generative chatbot for language learning purposes.

2.2 Decoding strategies

Neural models for text generation tasks typically maximise the likelihood of the generated output using beam search (e.g. Li et al., 2016; Rush et al., 2015; Sutskever et al., 2014). However, the most likely output may not be the most desirable one – e.g. in this paper, we would like to produce outputs of a particular difficulty level. One way to achieve desirable results is to further fine-tune the language model (e.g. Welleck et al., 2019; Roller et al., 2021), but this requires having (or being able to easily generate) data containing such desired traits.

Instead, it is possible to change the decoding strategy to produce desired outputs without further training of the language model. For example, to increase semantic and linguistic diversity, researchers have proposed changing the softmax temperature (Ackley et al., 1985; Caccia et al., 2020), or to use Stochastic Beam Search (Kool et al., 2019), top-k sampling (Fan et al., 2018), Nucleus Sampling (Holtzman et al., 2020), or conditional Poisson stochastic beam search (Meister et al., 2021). Decoding strategies have been also employed to control other linguistic attributes, such as output length (Kikuchi et al., 2016), style (Ghazvininejad et al., 2017), repetition, specificity, response-relatedness and question-asking (See et al., 2019). To our knowledge, our methods are the first to adjust the difficulty level during decoding.

2.3 Text simplification

Text simplification (TS) is the task of transforming complex text into simpler, more readable text that conveys the same meaning. Previous approaches are typically only designed to simplify where possible (e.g. Nisioi et al., 2017; Zhang and Lapata, 2017). More recently, methods have been proposed for controllable TS, where text can be simplified to a desired level of difficulty, such as Scarton and Specia (2018) and Nishihara et al. (2019), though both methods require training of a sequence-to-sequence model from scratch. Maddela et al. (2021) use a hybrid approach where the degree of simplification operations can be controlled, though not explicitly to a specified difficulty level.

Existing TS methods apply *transformative* operations to an existing piece of text. The main drawback is that some complex words may be impossible to simplify as there is no simpler alternative that conveys the same meaning (Shardlow, 2014). Our paper takes a different approach entirely, and instead adjusts difficulty of text when it is *generated*.

There is also research on specific operations within the TS pipeline. In particular, we discuss complex word identification (CWI) in subsection 2.4 below. Other operations such as sentence splitting (Narayan et al., 2017; Aharoni and Goldberg, 2018) and paraphrase generation (Gupta et al., 2018; Fu et al., 2019) are also transformative operations, where the outcome needs to convey the same meaning as the original input. Our generation methods do not have the same constraint.

2.4 Linguistic complexity

Lexical complexity Previous work on lexical complexity typically involves predicting the complexity of words within a context. There have been multiple shared tasks related to lexical complexity: the 2016 CWI shared task (Paetzold and Specia, 2016b) to identify complex words; the 2018 CWI shared task (Yimam et al., 2018) also to identify complex words, and to predict the probability that a word is complex; and the 2021 Lexical Complexity Prediction shared task (Shardlow et al., 2021) to predict the difficulty level of words, as determined by Likert-scale annotations. Submissions to the first two shared tasks were mostly dominated by feature-based approaches (e.g. Gooding and Kochmar, 2018; Kajiwara and Komachi, 2018). The 2021 shared task was won by Pan et al.

(2021) using an ensemble of pre-trained Transformers (Vaswani et al., 2017), but submissions with feature-based models also ranked highly (Mosquera, 2021; Rotaru, 2021).

Readability assessment Beyond the word level, research on linguistic complexity is typically done on long-form texts. Traditionally, researchers have derived formulae such as the Flesch-Kincaid score (Kincaid et al., 1975) and the Coleman-Liau index (Coleman and Liau, 1975) to estimate readability, but many such formulae do not account for semantic content or linguistic structure, or are outperformed by data-driven methods (Si and Callan, 2001). Later machine learning methods for readability assessment may rely on feature extraction (e.g. Meng et al., 2020; Deutsch et al., 2020; also see Martinc et al. (2021) for an analysis of different approaches).

3 Implementation

We propose 5 different decoding strategies to adjust the chatbot difficulty to one of 6 CEFR levels[1].

For our implementations, we used Facebook's Blender 2.7B (version 1[2], generator model) (Roller et al., 2021) as the basis, though our methods can be used or adapted to other language models that use beam search or sampling-based generation.

For comparability, all strategies use top-k sampling[3] (Fan et al., 2018) using $k = 40$ with a beam size of 20. We did not use the additional safety mechanisms (as described by Roller et al. (2021)) to ensure fair comparison of results.

Some of our methods use regression, either during generation or beforehand. For all regression tasks described below, we use a continuous scale from 0 to 5 to denote CEFR values, even though they are typically differentiated by qualitative properties (Council of Europe, 2020). This is because:

a) We have limited training data, and a scalar value provides more information to a regression model than a classification label; and

b) Due to the subjectivity of difficulty levels, there are often situations where examiners refer to values *between* CEFR levels, such as a "high B2/low C1". Using a continuous scale allows us to represent such in-between values.

3.1 Method 1: Vocabulary restriction with EVP

As a baseline strategy, we implemented a simple vocabulary filter based on a list of words manually labelled by CEFR[4]. The English Vocabulary Profile[5] (EVP) (Capel, 2015) maps 6,750 words and phrases to one or multiple CEFR levels according to their particular sense and usage. If the lowest[6] CEFR level of a word/phrase is higher than the target CEFR level, we prevent that word/phrase from being generated by setting the probability to 0[7,8]. For example, the word *absolutely* is labelled as B1, C1, or C2 depending on its usage. If the target CEFR level is B1 or above, the word is allowed and will retain its original probabilities; if the target CEFR level is A2 or below, the word will always have a probability of 0, and can never be generated.

As the EVP does not contain proper nouns, we also added a list of the most common first and last names (Social Security Administration; United States Census Bureau), and U.S. state names[9]. All

[1]Throughout this paper, we draw on the Common European Reference Framework (CEFR) (Council of Europe, 2020) to denote proficiency levels. An international standard for describing language ability, the CEFR organises ability into 6 levels, beginning with A1, continuing to A2, B1, B2, C1, and ending with C2, representing mastery of a second language.

[2]Version 2 had not been released at the time of our experiments.

[3]We use top-k sampling here because it was found to be equivalent to the default settings (in Roller et al., 2021) of beam size 10, with beam blocking of 3-grams and a minimum length of 20. Beam search, however, is deterministic, so top-k sampling allows us to generate multiple self-chats using the same settings.

[4]We chose to use a manually curated list to minimise errors, and because such vocabulary lists are often available as a language learning resource for widely-spoken languages. Alternatively, it is possible to produce similar word lists either in an unsupervised or semi-supervised manner (e.g. Jenny Ortiz-Zambrano, 2020).

[5]`https://www.englishprofile.org/wordlists/evp`

[6]We ignore the higher CEFR labels and collapse words with multiple meanings or usages into a single entry, because it is often impossible to determine the correct meaning during generation, when the rest of the sentence is missing.

[7]After modifications, the sum of all "probabilities" would no longer be 1, though we continue to refer to these as probabilities for the sake of exposition.

[8]Blender uses Byte-Level BPE tokenisation (Roller et al., 2021; Sennrich et al., 2016), so a word is not fully formed until the subsequent sub-token (beginning with a whitespace or punctuation denoting a word boundary) is chosen. To ensure that the 0 probabilities are assigned to the correct word, we also assign them to subsequent sub-tokens that begin with a word boundary.

[9]Since Blender is pre-trained on data from Reddit, where a large part of the user base comes from the U.S., we found that many of the dialogues contained names of U.S. states. We also noticed that when vocabulary is restricted and no proper names are allowed, the generated text sometimes contained approximations of locations in the U.S., such as "I live in wash in n" for *I live in Washington*. For this reason, we decided

added entries assume a CEFR level of A1, and so are always allowed regardless of the target CEFR level.

3.2 Method 2: Vocabulary restriction with extended EVP

Unfortunately, manually curated word lists typically have limited coverage. For example, the EVP only contains 6,750 words and phrases. For our 2nd method, we extended the list by training a regressor to predict the CEFR level of words outside of the EVP. We opted for a feature-based approach that is based purely on the surface word form rather than different word senses (as above), due to the lack of 1) training data and 2) available context while decoding[10]. Adapting the winning system (Gooding and Kochmar, 2018) of the 2018 CWI shared task, we selected a subset of features that are non-context-dependent[11]. The list of features used can be found in the appendix.

As in the original paper, we used the Random Forest implementation from *scikit-learn*[12] , but for regression instead of binary classification. CEFR levels in the training data were converted into integers from 0 to 5 (inclusive), and predicted values were rounded to the nearest CEFR level.

To evaluate this word complexity prediction model, we randomly selected one-fifth of the original EVP data as the test set, taking care to ensure that words from the same root cannot be found in both the training and test set. Results are shown in Table 1.

After evaluation, the model was re-trained on the whole dataset, then used to predict the CEFR levels of an additional 10,000 of the most common[13] English words that are not in the EVP. The prediction of CEFR levels is done beforehand to minimise computational costs at runtime.

to include in our vocabulary a list of U.S. states, along with popular first and last names from the U.S.

[10] As above, the CEFR level of a word is used to determine the probability of a sub-token at a given time step during decoding, where the rest of the sentence is still missing.

[11] Other features used in the original paper were context-dependent, and so were unsuitable for our use case.

[12] https://scikit-learn.org/

[13] Word frequency is estimated from the Exquisite Corpus, which combines frequencies from Wikipedia, subtitle corpora, news corpora, Google Books, and other resources. https://github.com/LuminosoInsight/exquisite-corpus

Spearman's ρ	Pearson's r	MAE
0.694	0.712	0.826

Table 1: Spearman's and Pearson's correlation and mean absolute error (MAE) of predicted CEFR levels of words in the EVP. Both correlation statistics are significant ($p \leq 0.001$). MAE of 1 corresponds to a difference of 1 CEFR level.

3.3 Method 3: Re-ranking

One main drawback of vocabulary restriction is that text difficulty is not necessarily determined by vocabulary choice alone. We want to generate outputs that are of the appropriate difficulty level in terms of structure, content, as well as choice of words.

For our 3rd method, we propose a re-ranking method that considers multiple candidate messages, before selecting the most appropriate one. As described in section 3, our models use beam size = 20, which generates 20 candidate messages for every message sent to the user.

We first trained a regressor to predict the CEFR level of sentences. When the chatbot is in use, the regressor will predict the CEFR level of all candidate messages, allowing us to compute a score that combines the original ranking and the predicted CEFR. This score will then be used to re-rank the candidates, and the top candidate message will be sent to the user.

For the regressor, we used a RoBERTa model pre-trained on a dynamic masked language modelling (Liu et al., 2019), which is then distilled (Sanh et al., 2019), as implemented in Huggingface Transformers (Wolf et al., 2020). We fine-tuned this model to predict text difficulty on the Cambridge Exams (CE) dataset (Xia et al., 2016), which contains English texts from Cambridge Exams aimed at learners of different CEFR levels. However, instead of training our model on entire texts, we used spaCy (Montani et al., 2021) to detect sentence boundaries, and trained the model to predict the CEFR level from individual sentences, as they are more similar in length to the messages generated by Blender. Since the prediction of candidate messages must occur during live interactive use, the distilled version of the model was chosen to minimise computational overhead.

As with the previous word complexity prediction model, we randomly selected one-fifth of the CE sentences as a test set for the sentence complexity prediction model, taking care to ensure that sentences from the same text cannot be found in both the training set and the test set. After evaluation, we re-trained the model using all available data, to be used to generate text. Initial evaluation results from the test set are shown in Table 2.

Spearman's ρ	Pearson's r	MAE
0.701	0.734	0.634

Table 2: Spearman's and Pearson's correlation and mean absolute error (MAE) of predicted CEFR levels of sentences from the Cambridge Exams dataset. Both correlation statistics are significant ($p \leq 0.001$). MAE of 1 corresponds to a difference of 1 CEFR level.

Our proposed re-ranking procedure accounts for:

(1) $P(C_i)$ the original probability of each **C**andidate C_i according to the chatbot language model

(2) L_{C_i} the difficulty **L**evel of each **C**andidate

(3) L_t the **t**arget difficulty **L**evel

We compute the new rank R of each candidate C_i by taking the average of its original rank (based on $P(C_i)$) and its difficulty rank (based on distance away from the target difficulty level).

$$R = \frac{r(P(C_i)) + w \cdot r(|L_t - L_{C_i}|)}{2} \quad (1)$$

r denotes a function that returns the rank of the candidate out of all candidates. That is: $r(P(C_i))$ is the ranking of probabilities from the model (where higher probability = higher rank), and $r(|L_t - L_{C_i}|)$ is the ranking of distance to target difficulty (where smaller distance = higher rank). r essentially normalises the probability values and difficulty levels before the final rank is computed. w is an adjustable weight that controls the importance of distance to the target difficulty.

To select a value for w, we manually annotated ideal rankings for 10 sets of 20 candidate outputs, and found that the original rank and the difficulty rank contribute equally to the final rankings. Therefore, for methods 3, 4 and 5, we use $w = 1$.

3.4 Method 4: Re-ranking with sub-token penalties

With method 3, we sometimes found that all 20 generated candidates would be of a similar difficulty level, which may be some distance away form the learner's CEFR level. For example, we might have 20 candidate responses at C1 level, while the learner is a B1 speaker.

In order to increase the probability that a candidate message is at the target CEFR level, we implemented an additional penalty system, which penalises sub-tokens that are too difficult. Sub-tokens that are too easy are not penalised, as many are words serving grammatical functions or common words, which are also frequently used in difficult texts. Note that, as with vocabulary restriction, penalties must be assigned to sub-tokens rather than words, because words are not fully formed until a following sub-token with a word boundary is chosen.

The penalty for a given sub-token varies depending on how difficult it is. To determine the CEFR level of a sub-token, we tokenised the texts in the CE dataset to identify which sub-tokens appeared at which CEFR level. The lowest CEFR level is then chosen for each sub-token. For example, a sub-token that appears in a B1 exam but not in A1 or A2 exams will be classified as a B1 sub-token.

The penalty values scale according to the difference between the target CEFR level and sub-token's CEFR level. For example, an A2 chatbot will assign a smaller penalty (or a larger weight) to a B1 sub-token (e.g. *_absolutely, _opportun, initely*) than a B2 sub-token (e.g. *_humanity, _adop, iable*)[14]. The penalty values are taken from a Gaussian distribution, where μ is a CEFR level difference of 0, and σ is 2 CEFR levels[15].

The new probability of a given sub-token is therefore calculated as follows:

$$p' = \begin{cases} p \cdot \varphi(L_s - L_t) & \text{if } L_s > L_t \\ p & \text{otherwise} \end{cases} \quad (2)$$

where p and p' denote the original and new probability respectively, L_s is the CEFR level of the sub-token, and L_t is the target CEFR level. φ represents the Gaussian distribution described above.

[14] where _ represents a whitespace character.

[15] We settled on this value for σ for relatively lenient penalties, because:

a) The Cambridge Exams dataset only contains 331 texts (averaging at 531 words each), so a low frequency token of e.g. B1 level may only appear at B2 level or above. Having more lenient penalties can account for such potential discrepancies.

b) If the resulting candidate is too difficult, it is likely to be filtered out in the re-ranking process.

However, this value can be adjusted based on the language model or applicational needs.

3.5 Method 5: Re-ranking with sub-token penalties and filtering

With method 4, we noticed that occasional non-sense words are generated. This was typically due to how penalties are assigned to sub-tokens rather than words: for example, on one occasion, *backpacking* was generated as *backpicking*.

To combat this, we added a vocabulary filter[16] to look for words that are out-of-vocabulary, ignoring capitalised words and punctuation. If a candidate message contains such a word, it is removed from the pool of candidates.

4 Evaluation

For each of our 5 methods, we generated 300 self-chat dialogues using Blender, where the chatbot talks to itself (Li et al., 2016). Each self-chat was generated using the settings for a specific CEFR level: for example, method 1 at B1 level would only generate vocabulary at B1 level or below; method 3 at C1 level would re-rank outputs based on how close it is to C1 difficulty.

Then, to determine whether these methods are truly able to generate messages at the intended level, we recruited English language examiners to judge the *true* difficulty level of each self-chat.

We chose self-chats rather than human-model chats (i.e. chats between a human and the language model) for three reasons: firstly, because we did not want the examiner's judgement of the chatbot output to be biased by the proficiency level of the user; secondly, because it is cheaper and less time consuming to generate self-chats; and finally, because second language users may struggle to communicate with the chatbot. Additionally, previous work comparing self-chats to human-model chats found that they produced similar results (Li et al., 2019).

Each self-chat consists of 18 messages, all prompted by an initial message, "Hello!". An example of a generated self-chat can be found in the appendix. The 300 dialogues for each method are split evenly into 6 sets of 50, each set targeting a different CEFR level. An additional 100 dialogues were generated without any modifications for comparison, resulting in an overall total of 1600 dialogues (see Table 3).

We recruited 10 English language examiners from Cambridge University Press & Assessment.

	1	2	3	4	5	B
A1	50	50	50	50	50	0
A2	50	50	50	50	50	0
B1	50	50	50	50	50	0
B2	50	50	50	50	50	0
C1	50	50	50	50	50	0
C2	50	50	50	50	50	0
N/A	0	0	0	0	0	100
Total	300	300	300	300	300	100

Table 3: Number of self-chats generated for each method / CEFR combination. **B** refers to self-chats generated using the original **B**lender configurations with no modifications, which cannot be targeted at a given CEFR level.

All 10 examiners were provided with a set of genre-specific descriptors adapted from the CEFR[17]. In addition, to assess the general quality of the produced text, each message in the dialogue was labelled according to whether it was sensible and whether it was specific (following Adiwardana et al., 2020), as well as whether it was grammatical. Examiners were given additional guidance on edge cases to support their application of these labels.

Each dialogue was annotated by at least 3 different examiners. For the final results, disagreements between examiners are resolved by taking the average of all annotations. The inter-annotator agreement for our CEFR annotations is 0.79, measured with weighted Fleiss' κ (Fleiss, 1971), and assuming equal distance between CEFR levels. For the grammatical, sensible, and specific labels, we used Gwet's AC_1 (Gwet, 2014)[18]. The agreement scores are 0.62 for grammaticality labels, 0.23 for sensibleness labels, and 0.67 for specific labels.

Agreement in sensibleness is noticeably lower than the others: feedback from annotators suggested that sensibleness of a particular message is often unclear when the previous context already contained messages that were not sensible. Experimental results from Adiwardana et al. (2020) suggest that agreement scores may be higher if annotators are only asked to label single responses within a pre-written, sensible context. However, they also note that "final results are always aggregated labels", so the overall proportion of sensible

[16] We use a list of words (containing only letters) from `https://github.com/dwyl/english-words`.

[17] Descriptors were adapted from 3 CEFR scales: Overall Reading Comprehension, Overall Oral Interaction, and Conversation. The descriptors we used in our experiments can be found in the appendix.

[18] rather than Krippendorff's α, because our data is very skewed (containing 87.0% grammatical, 75.7% sensible, and 91.6% specific responses), and AC_1 accounts for marginal probabilities.

Method	Spearman's ρ	Pearson's r	MAE	%gramm.	%sensible	%specific
Original	N/A	N/A	N/A	**90.2%**	**81.9%**	**94.1%**
Method 1	0.229	0.243	1.410	89.0%	76.5%	91.4%
Method 2	0.196	0.194	1.461	89.5%	77.3%	91.7%
Method 3	0.719†	0.707†	1.120	87.4%	77.0%	93.0%
Method 4	0.680†	0.681†	1.174	87.2%	76.1%	91.9%
Method 5	**0.755†**	**0.731†**	**1.090**	87.3%	76.4%	92.1%

Table 4: Table showing, for each method: Spearman's and Pearson's correlation between target CEFR and true CEFR; mean absolute error (MAE) of target CEFR compared to true CEFR, where MAE of 1 corresponds to a difference of 1 CEFR level; and percentage of grammatical, sensible, and specific responses. † indicates a significant correlation ($p \leq 0.001$). For all 5 methods, proportions of grammatical, sensible, and specific messages are found to be statistically equivalent (Wellek, 2010) to the original ($\epsilon = 0.001$, $p \leq 0.001$).

labels is still indicative of chatbot quality, despite relatively low agreement scores.

5 Results and discussion

Our results are in Table 4. To our knowledge, this is the first attempt at adjusting text difficulty during open-ended text generation – therefore, we were unable to find comparable results to be included here. Where possible, we include results for the original (unmodified) generation method, which cannot be targeted at any specific CEFR level.

For each of the methods, we compared the target CEFR to the CEFR determined by examiners – henceforth referred to as the *true* CEFR. Spearman's ρ and Pearson's r show the correlation between the two, and MAE is the mean absolute error, where an MAE of 1 refers to the difference of 1 CEFR level. %gramm., %sensible, and %specific refers to the percentage of grammatical, sensible, and specific responses out of all responses (excluding the original "Hello!" prompt).

From the correlation and MAE scores, we can see that the re-ranking methods work best, with method 5 – reranking with sub-token penalties and filtering – achieving the strongest correlation and lowest MAE between the target and true CEFR. Both vocabulary-based methods performed poorly, achieving almost no correlation and high MAE scores. This is somewhat surprising, as one might expect vocabulary to be a key factor in determining text difficulty. We suspect that this is because many of the easier words in the EVP also have more difficult word senses, but our method only considered the lowest CEFR level. Additionally, we looked at a set of 10 randomly sampled dialogues and counted 66 multi-word expressions (MWEs) in total, averaging at 0.36 MWEs per message. MWEs might often be more difficult than their constituent words individually: for example, the idiom *a cut above the rest* consists of words that are individually simple, but the phrase itself is relatively complex. Unfortunately, our vocabulary restriction methods are not able to account for this.

5.1 CEFR distribution

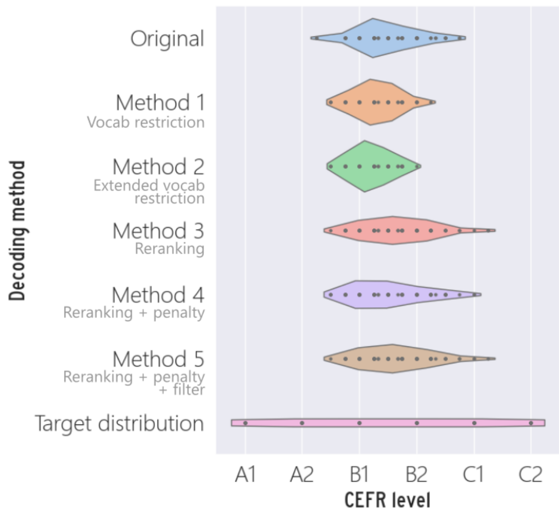

Figure 1: The first 6 violin plots shows distribution of CEFR levels for each method, along with the original version, all scaled for comparison. The last violin plot shows the target distribution for our 5 methods. It is spread evenly across all CEFR levels, as we had generated the same number of self-chats for each level (see Table 3).

Figure 1 contains violin plots showing the distribution of CEFR level for each method, including the original version with no modifications. The last violin plot is the ideal distribution for our 5 variants, which is evenly spread throughout all 6 levels. Out of the 5 variants, the 3 re-ranking variants have the most even distribution.

One surprising finding from this study is that the CEFR level of dialogues that were generated *without* modifications to difficulty were mostly in the B1 to B2 range, rather than C1 and C2, which would be closer to a "native-like" difficulty, i.e. intended for users of native proficiency. Since the restriction-based methods served to reduce the difficulty level rather than increase it, the result is that none of the generated dialogues were labelled as C1 or C2, so the CEFR distribution clustered around the B1 level. On the other hand, the reranking based methods performed better because, when the target CEFR is C1 or C2, the reranking procedure would select texts that are more difficult than the most likely output. However, we suspect that the imbalance of CEFR levels in dialogues generated by the original model affects all 5 variants, and may adversely affect the MAE scores.

Another observation from our data is that none of our dialogues were labelled as A1, which is the lowest CEFR level. We suspect that this is because dialogic communication is inherently difficult for beginners, and there are simply too few topics and words that are suitable for all A1 learners. For example, the CEFR scale for written interaction simply states that A1 learners "can ask for or pass on personal details" (Council of Europe, 2020). However, we leave it to future work to explore other ways of generating dialogue data at A1 level.

5.2 Message quality

While our grammaticality, sensibleness, and specificity scores were found to be statistically equivalent ($\epsilon = 0.001$, $p \leq 0.001$), it may not be surprising to see a slight degradation of quality in Blender's messages when using our decoding methods. Our methods are designed to reject the most likely output if its difficulty level is not appropriate, and to select the next best output that falls within our constraints.

The focus of this paper is on the decoding methods rather than the original language model, which may be improved on or replaced by a different generative model. However, we acknowledge that the quality of messages may detract from the learning experience, particularly ones that are not grammatical or not sensible.

According to the inter-annotator agreement scores in section 4, there was relatively little agreement on what was considered sensible. In future work, it would be important to refine the criteria to better evaluate the quality of messages. Additionally, it may be possible to implement style classifiers or contradiction detection tools to mitigate this issue.

Sampling 100 messages from ones which were considered ungrammatical by at least one examiner, we identified three types of 'ungrammaticality':

- Around half (51) involved colloquialisms (e.g. "LOL", comma splicing, and other capitalisation or punctuation errors) that are ungrammatical in written English, but are more accepted in online messaging.
- More than a quarter (29) contained awkward phrasing depending on the context and/or was marked as not sensible. This becomes a grey area where it is difficult to determine whether the intended meaning was not sensible, or if the surface linguistic form was incorrect.
- Only a fifth (20) were clearly ungrammatical (e.g. "on your free time") or involved a spelling mistake (e.g. "clausterphobia").

Since Blender was pre-trained on large amounts of data from Reddit (Roller et al., 2021), it is unsurprising to see internet colloquialisms in the generated messages. While this may not be desirable for formal written work, learners are likely to come across similar forms of language in online or computer-mediated interaction. Alternatively, it may be possible to use grammatical error detection tools or style classifiers to filter out these messages. We leave to future work to investigate ways of filtering undesirable messages.

6 Conclusion and future work

This paper presents an initial foray into using open-domain chatbots for language practice. We propose and compare 5 variants of decoding strategies to adjust the difficulty level of generated text. We then evaluate these methods on dialogue generation, and find that our re-ranking strategies significantly outperform vocabulary-based methods. Our best variant achieved a Spearman's ρ of 0.755 and Pearson's r of 0.731.

Our current work only looks at self-chat difficulty from a teacher/examiner's perspective, which may not transfer well to interactive difficulty. It is also important to ensure that language learners would benefit from this endeavour. For our future work, we will directly engage with learners to investigate the utility and impact of chatbots on language learning.

However, there are also areas where the chatbot needs to be significantly improved upon. For example, to cater for A1 learners, we need to be able to generate messages at A1 difficulty. This paper only looked at text complexity in terms of vocabulary, but it may also be possible to adjust the complexity by paraphrasing or altering the sentence structure.

We also need to ensure that generated messages are grammatical, sensible, specific, and appropriate. There is ongoing research on grammatical error detection (cf. Wang et al., 2021), toxic language detection (e.g. Dinan et al., 2019), and improving dialogue consistency (e.g. Li et al., 2020), which can be used to improve the chatbot.

Additionally, a language learning chatbot can be further augmented with other technologies to enhance the user experience, such as grammatical error correction tools, dictionary lookup, or personalisation mechanisms. However, it is not always clear what tools or mechanisms would best facilitate language learning: for example, immediate grammar correction could distract and cause communication breakdown (Lyster et al., 2013). We leave this investigation to future research.

7 Ethical concerns

By building a chatbot for language learning, we hope to make interactive, open-domain language practice more accessible to all learners. However, there are ethical risks that must be considered before providing learners with such a tool. In particular, we highlight three areas in which open-domain chatbots may have harmful effects, especially for younger learners.

1. **Toxic language**
   Open-domain chatbots are typically based on pre-trained large language models. These models, especially ones trained on internet data, are known to produce outputs that are toxic (e.g. Gehman et al., 2020) or that contain harmful biases (e.g. Nadeem et al., 2021; Sheng et al., 2019). There is existing and ongoing research on ways to mitigate these outputs (e.g. Xu et al., 2020; Dinan et al., 2019; Faal et al., 2022; Dinan et al., 2020), though Gonen and Goldberg (2019) argue that debiasing methods are insufficient and do not remove bias entirely. It remains an important ethical concern, especially for younger learners. For our experiments, we only recruit adult participants, who are warned about such messages beforehand.

2. **Inaccurate information**
   Large language models are also known to hallucinate knowledge during text generation (Roller et al., 2021; Maynez et al., 2020). While there is ongoing work to reduce this (Zhao et al., 2020; Komeili et al., 2020, e.g.), users should also be made aware that the information generated by a chatbot may not be accurate.

3. **Human likeness**
   Users should know that they are interacting with a machine rather than a human. Weizenbaum (1966) remarks, "In human conversation a speaker will make certain (perhaps generous) assumptions about his conversational partner." This is also known as the ELIZA effect (Hofstadter, 1995), which affects a user's perception of and emotional response to a chatbot.

In our experiments, evaluation was done through self-chats, and annotators did not interact with the chatbot directly. All annotators involved are adults and were asked to identify nonsensical or inaccurate statements (sensibility), and to flag any inappropriate language. In total, 6 of the 1600 (0.4%) self-chat dialogues we generated contained inappropriate language, or touched on inappropriate topics. We will make use of these in future work to address inappropriate chatbot turns.

Acknowledgements

We thank Helen Allen, Sian Gooding, David Strohmaier, and Andrew Rice for their advice and support. We also thank our anonymous reviewers for their valuable comments and suggestions. This paper reports on research supported by Cambridge University Press & Assessment. This work was performed using resources provided by the Cambridge Service for Data Driven Discovery (CSD3) operated by the University of Cambridge Research Computing Service (www.csd3.cam.ac.uk), provided by Dell EMC and Intel using Tier-2 funding from the Engineering and Physical Sciences Research Council (capital grant EP/P020259/1), and DiRAC funding from the Science and Technology Facilities Council (www.dirac.ac.uk).

References

David H. Ackley, Geoffrey E. Hinton, and Terrence J. Sejnowski. 1985. A learning algorithm for boltzmann machines. *Cognitive Science*, 9(1):147–169.

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot.

Roee Aharoni and Yoav Goldberg. 2018. Split and rephrase: Better evaluation and stronger baselines. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 719–724, Melbourne, Australia. Association for Computational Linguistics.

Serge Bibauw, Thomas François, and Piet Desmet. 2019. Discussing with a computer to practice a foreign language: research synthesis and conceptual framework of dialogue-based call. *Computer Assisted Language Learning*, 32(8):827–877.

Massimo Caccia, Lucas Caccia, William Fedus, Hugo Larochelle, Joelle Pineau, and Laurent Charlin. 2020. Language gans falling short. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Annette Capel. 2015. The english vocabulary profile. In Julia Harrison and Fiona Barker, editors, *English Profile in Practice*, chapter 2, pages 9–27. UCLES/Cambridge University Press.

Rollo Carpenter. jabberwacky - live chat bot - ai artificial intelligence chatbot - jabber wacky - talking robot - chatbots - chatterbot - chatterbots - jabberwocky - take a turing test - loebner prize - chatterbox challenge - entertainment robots, robotics, marketing, games, digital pets - jabberwhacky. Available at `http://www.jabberwacky.com/` (2022/05/17).

Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *ACM SIGKDD Explorations Newsletter*, 19(2):25–35.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283–284.

William Coster and David Kauchak. 2011. Simple English Wikipedia: A new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 665–669, Portland, Oregon, USA. Association for Computational Linguistics.

Council of Europe. 2020. *Common European Framework of Reference for Languages: Learning, Teaching Assessment*, 3rd edition. StrasBourg.

Tovly Deutsch, Masoud Jasbi, and Stuart Shieber. 2020. Linguistic features for readability assessment. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–17, Seattle, WA, USA → Online. Association for Computational Linguistics.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188, Online. Association for Computational Linguistics.

Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 2019. Build it break it fix it for dialogue safety: Robustness from adversarial human attack. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4537–4546, Hong Kong, China. Association for Computational Linguistics.

Farshid Faal, Jia Yuan Yu, and Ketra Schmitt. 2022. Reward modeling for mitigating toxicity in transformer-based language models.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Yao Fu, Yansong Feng, and John P Cunningham. 2019. Paraphrase generation with latent bag of words. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

Marjan Ghazvininejad, Xing Shi, Jay Priyadarshi, and Kevin Knight. 2017. Hafez: an interactive poetry generation system. In *Proceedings of ACL 2017, System Demonstrations*, pages 43–48, Vancouver, Canada. Association for Computational Linguistics.

Yoav Goldberg and Jon Orwant. 2013. A dataset of syntactic-ngrams over time from a very large corpus of English books. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages

241–247, Atlanta, Georgia, USA. Association for Computational Linguistics.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. 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 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

Sian Gooding and Ekaterina Kochmar. 2018. CAMB at CWI shared task 2018: Complex word identification with ensemble-based voting. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 184–194, New Orleans, Louisiana. Association for Computational Linguistics.

Ankush Gupta, Arvind Agarwal, Prawaan Singh, and Piyush Rai. 2018. A deep generative framework for paraphrase generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Kilem L Gwet. 2014. *Handbook of Inter-Rater Reliability*, 4th edition. Advanced Analytics, LLC, Gaithersburg, MD.

Douglas R Hofstadter. 1995. *The ineradicable Eliza effect and its dangers*, chapter Preface 4.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Elaine Horwitz. 2001. Language anxiety and achievement. *Annual Review of Applied Linguistics*, 21:112–126.

Arturo Montejo-Ráez Jenny Ortiz-Zambrano. 2020. Overview of alexs 2020: First workshop on lexical analysis at sepln. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)*, volume 2664, pages 1–6.

Tomoyuki Kajiwara and Mamoru Komachi. 2018. Complex word identification based on frequency in a learner corpus. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 195–199, New Orleans, Louisiana. Association for Computational Linguistics.

Yuta Kikuchi, Graham Neubig, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. 2016. Controlling output length in neural encoder-decoders. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1328–1338, Austin, Texas. Association for Computational Linguistics.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation Of New Readability Formulas (Automated Readability Index, Fog Count And Flesch Reading Ease Formula) For Navy Enlisted Personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2020. Internet-augmented dialogue generation.

Wouter Kool, Herke Van Hoof, and Max Welling. 2019. Stochastic beams and where to find them: The gumbel-top-k trick for sampling sequences without replacement. In *International Conference on Machine Learning*, pages 3499–3508.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Austin, Texas. Association for Computational Linguistics.

Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. 2020. Don't say that! making inconsistent dialogue unlikely with unlikelihood training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4715–4728, Online. Association for Computational Linguistics.

Margaret Li, Jason Weston, and Stephen Roller. 2019. Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons. In *33rd Conference on Neural Information Processing Systems: Conversational AI Workshop*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Shawn Loewen and Masatoshi Sato. 2018. Interaction and instructed second language acquisition. *Language Teaching*, 51(3):285–329.

Michael H Long. 1996. The role of the linguistic environment in second language acquisition. In William C. Ritchie and T.K. Bhatia, editors, *Handbook of Second Language Acquisition*, chapter 13, pages 413–454. Academic Press.

Roy Lyster, Kazuya Saito, and Masatoshi Sato. 2013. Oral corrective feedback in second language classrooms. *Language Teaching*, 46(1):1–40.

Alison Mackey. 2013. *Conversational Interaction in Second Language Acquisition*. Oxford Applied Linguistics. Oxford University Press.

Mounica Maddela, Fernando Alva-Manchego, and Wei Xu. 2021. Controllable text simplification with explicit paraphrasing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3536–3553, Online. Association for Computational Linguistics.

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2021. Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, 47(1):141–179.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Clara Meister, Afra Amini, Tim Vieira, and Ryan Cotterell. 2021. Conditional Poisson stochastic beams. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 664–681, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Changping Meng, Muhao Chen, Jie Mao, and Jennifer Neville. 2020. Readnet: A hierarchical transformer framework for web article readability analysis. In *Advances in Information Retrieval*, pages 33–49, Cham. Springer International Publishing.

Ines Montani, Matthew Honnibal, Matthew Honnibal, Sofie Van Landeghem, Adriane Boyd, Henning Peters, Paul O'Leary McCann, Maxim Samsonov, Jim Geovedi, Jim O'Regan, György Orosz, Duygu Altinok, Søren Lind Kristiansen, , Roman, Explosion Bot, Leander Fiedler, Grégory Howard, Wannaphong Phatthiyaphaibun, Yohei Tamura, Sam Bozek, , Murat, Mark Amery, Björn Böing, Pradeep Kumar Tippa, Leif Uwe Vogelsang, Bram Vanroy, Ramanan Balakrishnan, Vadim Mazaev, and GregDubbin. 2021. explosion/spacy: v3.2.0: Registered scoring functions, doc input, floret vectors and more.

Alejandro Mosquera. 2021. Alejandro mosquera at SemEval-2021 task 1: Exploring sentence and word features for lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 554–559, Online. Association for Computational Linguistics.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Shashi Narayan, Claire Gardent, Shay B. Cohen, and Anastasia Shimorina. 2017. Split and rephrase. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 606–616, Copenhagen, Denmark. Association for Computational Linguistics.

Daiki Nishihara, Tomoyuki Kajiwara, and Yuki Arase. 2019. Controllable text simplification with lexical constraint loss. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 260–266, Florence, Italy. Association for Computational Linguistics.

Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P. Dinu. 2017. Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 85–91, Vancouver, Canada. Association for Computational Linguistics.

Charles Kay Ogden. 1930. *Basic English: A general introduction with rules and grammar*. Paul Treber & Co., Ltd., London.

Gustavo Paetzold and Lucia Specia. 2016a. Collecting and exploring everyday language for predicting psycholinguistic properties of words. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1669–1679, Osaka, Japan. The COLING 2016 Organizing Committee.

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

Chunguang Pan, Bingyan Song, Shengguang Wang, and Zhipeng Luo. 2021. DeepBlueAI at SemEval-2021 task 1: Lexical complexity prediction with a deep ensemble approach. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 578–584, Online. Association for Computational Linguistics.

Alexandros Papangelis, Paweł Budzianowski, Bing Liu, Elnaz Nouri, Abhinav Rastogi, and Yun-Nung Chen, editors. 2021. *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*. Association for Computational Linguistics, Online.

Luke Plonsky and Frederick L. Oswald. 2014. How big is "big"? interpreting effect sizes in l2 research. *Language Learning*, 64(4):878–912.

Princeton University. 2010. About WordNet. `https://wordnet.princeton.edu/`.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of* *the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Armand Rotaru. 2021. ANDI at SemEval-2021 task 1: Predicting complexity in context using distributional models, behavioural norms, and lexical resources. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 655–660, Online. Association for Computational Linguistics.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Carolina Scarton and Lucia Specia. 2018. Learning simplifications for specific target audiences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 712–718, Melbourne, Australia. Association for Computational Linguistics.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. 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 1702–1723, Minneapolis, Minnesota. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Matthew Shardlow. 2014. Out in the open: Finding and categorising errors in the lexical simplification pipeline. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1583–1590, Reykjavik, Iceland. European Language Resources Association (ELRA).

Matthew Shardlow, Richard Evans, Gustavo Henrique Paetzold, and Marcos Zampieri. 2021. SemEval-2021 task 1: Lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1–16, Online. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Luo Si and Jamie Callan. 2001. A statistical model for scientific readability. In *Proceedings of the 10th International Conference on Information and Knowledge Management*, pages 574–576. ACM Press.

Social Security Administration. Top names over the last 100 years.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

United States Census Bureau. Frequently occurring surnames from the 2010 census.

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.

Yu Wang, Yuelin Wang, Kai Dang, Jie Liu, and Zhuo Liu. 2021. A comprehensive survey of grammatical error correction. *ACM Transactions on Intelligent Systems and Technology*, 12(5):1–51.

Joseph Weizenbaum. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9(1):36–45.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2019. Neural text generation with unlikelihood training.

Stefan Wellek. 2010. *Testing Statistical Hypotheses of Equivalence and Noninferiority*, 2nd edition, chapter 9.2. Chapman & Hall/CRC, Boca Raton, FL.

Michael Wilson. 1988. MRC psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers*, 20(1):6–10.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text readability assessment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22, San Diego, CA. Association for Computational Linguistics.

Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2020. Recipes for safety in open-domain chatbots.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo 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, New Orleans, Louisiana. Association for Computational Linguistics.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 584–594, Copenhagen, Denmark. Association for Computational Linguistics.

Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020. Knowledge-grounded dialogue generation with pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3377–3390, Online. Association for Computational Linguistics.

A Implementational details

A.1 Blender settings

We used the Blender 2.7B generative model released through ParlAI[19] (model file: `zoo:blender/blender_3B/model`) as the basis for our chatbot models. Table 5 lists the hyperparameters used for all chatbot models in our experiment.

During this project, we noticed that the generated dialogues sometimes contained sequences such as "+/u/dogetipbot", since Blender was pre-trained on large amounts of Reddit data (Roller et al., 2021). As this is beyond the scope of our project, and to prevent this from affecting our results, we decided to filter out sequences containing "u/" and "r/" for all dialogues, so that results are still comparable. This filtering step occurs just before a message is selected from the pool of candidates: if a candidate contains either "u/" or "r/", it is removed from the pool, and the next best candidate is selected and sent to the user.

[19] `https://parl.ai/`

Hyperparameter	Value
Beam size	20
Top-k	40
Temperature	1.0
Beam delay	30
Beam length penalty	0.65
Beam block n-gram	3
Beam context block n-gram	3
Number of encoder layers	2
Number of decoder layers	2
Embedding size	2560
Number of attention heads	32
Hidden layer dropout	0.1
Attention dropout	0.1
Activation function	GELU

Table 5: Hyperparameters and corresponding values used for our chatbot models.

A.2 Word difficulty prediction model

For our word difficulty prediction model used for method 2 (restriction with extended EVP), we used the `RandomForestRegressor` from the *scikit-learn* library[20]. We used the following features:

- Word length
- Number of syllables
- Number of WordNet synsets (Princeton University, 2010)
- Number of WordNet hypernyms
- Number of WordNet hyponyms
- Word frequency in subtitles from Movies and Series for Children in the SubIMBD corpus (Paetzold and Specia, 2016a)
- Word frequency in the SimpleWiki[21] (Coster and Kauchak, 2011)
- Word presence in Ogden's Basic English list (Ogden, 1930)
- Word frequency according to syntactic-ngrams compiled by Goldberg and Orwant (2013)
- Number of phonemes (from the MRC Psycholinguistic Database, Wilson, 1988)
- Kucera-Francis frequency norms (MRC)
- Thorndike-Lorge frequency (MRC)
- Familiarity (MRC)
- Concreteness (MRC)
- Imageability (MRC)
- Age of acquisition (MRC)

[20] `https://scikit-learn.org/`
[21] `https://simple.wikipedia.org/`

Hyperparameter	Value
Number of estimators	5000
Splitting criterion	MSE
Min. number of samples for splitting	2
Min. number of samples at leaf node	1
Min. impurity decrease	None
Sample weighting	None

Table 6: Hyperparameters and corresponding values used for our word difficulty prediction model.

Hyperparameter	Value
Learning rate	9.737×10^{-6}
Batch size	32
Number of epochs	3
Random seed	18
Number of encoder layers	6
Embedding size	768
Number of attention heads	12
Hidden layer dropout	0.1
Attention dropout	0.1
Activation function	GELU

Table 7: Hyperparameters and corresponding values used for our sentence difficulty prediction model.

CEFR levels in the training data were converted into integers from 0 to 5 (inclusive), and predicted values were rounded to the nearest CEFR level accordingly. Hyperparameters for this model are listed in Table 6.

A.3 Sentence difficulty prediction model

For the sentence difficulty prediction model in our reranking-based methods, we used the `distilroberta-base` implementation from Huggingface Transformers (Wolf et al., 2020), and added a regression head to output a value representing difficulty. We tuned the learning rate, batch size, number of epochs, and random seed for this model using Optuna[22]. The final hyperparameters for this model are listed in Table 7.

The training data is taken from the Cambridge Exams dataset (Xia et al., 2016), where the text is split up into sentences using SpaCy's (Montani et al., 2021) `en_core_web_sm` model. As above, CEFR levels in the training data were converted into integers from 0 to 5 (inclusive), and predicted values were rounded to the nearest CEFR level accordingly.

[22] https://optuna.org/

C2	Appropriate for a reader who can understand and interpret critically virtually all messages and dialogues including abstract, structurally complex, or highly colloquial text. Appropriate for a reader who can appreciate subtle distinctions of style and implicit as well as explicit meaning, including idiomatic expressions and colloquialisms.
C1	Appropriate for a reader who can understand in detail complex dialogues, whether or not they relate to their own area of speciality, provided he/she can reread difficult messages. Appropriate for a reader who can understand a wide range of demanding messages, and recognise implicit meaning, including emotional, allusive, and joking usage of language.
B2	Appropriate for a reader who can understand the main ideas of complex dialogues across a range of concrete and abstract topics, and has a broad active reading vocabulary, but may experience some difficulty with low-frequency idioms.
B1	Appropriate for a reader who can understand straightforward factual messages on subjects related to his/her field and interest with a satisfactory level of comprehension. Appropriate for a reader who can understand the main points of clear messages on familiar matters regularly encountered in work, school, leisure, etc.
A2	Appropriate for a reader who can understand short, simple messages on familiar matters of a concrete type which consist of high frequency everyday or job-related language (including simple everyday polite forms of greetings and farewells; introductions; invitations and apologies; giving thanks), or shared international vocabulary items. Appropriate for a reader who can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment).
A1	Appropriate for a reader who can understand very short, simple messages a single phrase at a time, picking up familiar names, words and basic phrases and rereading as required. Appropriate for a reader who can understand familiar everyday expressions and very basic phrases aimed at the satisfaction of simple needs of a concrete type.

Table 8: CEFR descriptors provided to examiners to annotate difficulty levels.

Speaker 1:	Hello!
Speaker 2:	Hello there. Do you listen to music? I'm all about rock. Songs that have a melody are my favorite.
Speaker 1:	Yes, I love the melodies in rock music. I think that is the foundation of all music after all.
Speaker 2:	I agree. Music can have many meanings and have spiritual and religious elements too. How about you?
Speaker 1:	Yes, music can have a strong emotional impact too. It can enhance an experience or provoke a response.
Speaker 2:	Definitely! Music also can be interpreted through different cultural lenses as well. Have you ever thought about that?
Speaker 1:	Yes, that is true, since music can come in a wide variety of genres and structures.
Speaker 2:	Absolutely! Since it is so dynamic and changes from time to time, it has a tendency to change and evolve with the culture as well!
Speaker 1:	Yes, I think it is a good thing that music is dynamic though. Cultures have a way of shaping and defining what music is.
Speaker 2:	You're absolutely right! Humans have been using music for communication and for artistic inspiration for much of history so I wouldn't want it any other way!
Speaker 1:	Music can also influence the way we view our relationships, our ideas, and even our language as well.
Speaker 2:	Yes exactly! Music, art, language, entertainment, and science all play major roles in human society!
Speaker 1:	Sometimes I wonder if there is such thing as the soul, because it seems to come to people through their interpretation of music.
Speaker 2:	Yes there definitely is! The soul is definitely subjective and a big part of the human psyche so it makes sense that people interpret music in different ways!
Speaker 1:	I think we see it in art all the time. Artists often are influenced by their influences, such as their relationships.
Speaker 2:	In the past, many considered the relationship between the musician and audience to be one of the most important aspects of the medium!
Speaker 1:	Yes! I agree with that, as artists are affected by their art in a very personal way.
Speaker 2:	Yes and for a lot of artists the subject matter can be important to the storytelling and emotions of their music as well.
Speaker 1:	Indeed. And they often do not even realize how impactful their work is until after the fact.

Table 9: A cherry-picked example of a C1 self-chat dialogue generated using method 5 (reranking with sub-token penalties and filtering).