Towards a Method to Classify Language Style for Enhancing Conversational Systems

Paulo Cavalin*, Victor H. Alves Ribeiro*, Marisa Vasconcelos*, Claudio Pinhanez*, Julio Nogima*, Henrique Ferreira†
*IBM Research
São Paulo, Brazil
pcavalin@br.ibm.com
†Universidade Federal do ABC - UFABC

***Abstract*—Chatbots have received significant attention in the last years. These systems have improved operational efficiency by reducing the cost of customer service and more and more are used in customer service channels. More recently, with the addition of speech capabilities added to those systems, bot language modifications have to be done in order to improve user experience. In this paper, we analyze the language used in those systems using datasets collected from real chatbots. For that, we first propose models that are able to identify the language style (writing, speech, and computer-mediated) using as linguist features syntactic (Biber's dimensions) and sentence embeddings as well as state-of-art datasets. Our results show that our models were able to distinguish among the three classes with an accuracy of up to 66%. Finally, we evaluate real chatbots systems, either speech- and text-based ones, using our proposed models. We found that these real chatbots are generally using a computer-mediated style, but there is indication that their developers tend to adapt the language style to be more speech-like when text-to-speech is used.**

***Index Terms*—Text simplification, Multiple classifiers, Dynamic selection**

I. Introduction

Dialogue systems are becoming increasingly common over recent years, owing to a considerable advance in machine learning techniques which perform intent classification, answer generation, text to speech (TTS), and speech to text (STT) conversion. Such technologies have attracted the attention of industry that have tied them together to build up voice-based chatbots which can perform a great deal of customer care interactions, seamlessly as if the service is provided by a real agent.

The main objective of this paper is to present a study on the language style used in the development of such dialogue systems. We constrain the term "language style" here to determine whether the utterances generated by such systems are more similar to those commonly used by people in writing, chatting, or spoken contexts. To determine whether the appropriate language style is employed in such systems, we built and use machine learning classifiers. The main goal is to have a better understanding of the language style which has been employed to create such systems; and of how the language style can affect the efficiency of voice-based dialogue systems.

To our knowledge, there has been no previous study on how different styles of language may affect interaction with speech-based conversational systems. However, there is extensive previous work on how people perceive and are affected by other aspects of computer-generated speech and text, especially in the body of literature created by Clifford Nass and his colleagues [1], [2]. For instance, users are more susceptible to make buying decisions based on how they perceive a computer generated voice according to gender, ethnicity, and extroversion [3]–[6], even when they are clearly told that they are interacting with a computer. People seem to not be able to avoid ascribing human traits to machine-generated voices, and to notice, sometimes unconsciously, even small defects and inconsistencies in them.

By analogy, it seems to be necessary to use the correct style of language, written or spoken, according to how the final user will interact with the conversational system. In practice, however the development of chatbots is usually done by domain experts and not by linguistic experts or writing professionals. Even if clear rules existed about writing speech vs. written text, it would be difficult for domain experts to follow those rules since they may just not have the language perception and the crafting abilities to write text in different styles.

Another issue is that, to make chatbot development more user-friendly, STT and TTS modules are usually general-purpose plugins which are put on top of text-based chatbots. Particularly for the TTS systems, the way that chatbot responses are written may negatively affect the generated speech, or at least result in awkward unnatural spoken responses, possibly harming the user experience as pointed out before for other voice features. It is thus desirable that the responses of voice-based systems should be closer to speech style (*speech-like*) as possible. Having a tool which assesses how much *speech-like* is a given text and suggests text modifications to turn the text more suitable for a given TTS system can help chatbot developers who, as noticed before, more often than not have limited writing abilities.

Fortunately, many studies in Linguistics [7]–[9] have looked into the issue of defining the characteristics indicating the

differences between *speech* and *writing* modes of language, and more recently, *computer-mediated communication (CMC)* styles, such as the ones used in text-based chats. In this work, we present an investigation of the feasibility of developing a classifier to predict the language style of a text based on those studies. Our study aimed at validating seminal work in the field [10] and compare it with more recent deep learning approaches. Additionally, we applied the developed methods on real-world data to better understand their language styles.

For achieving our goals, we have first performed a data analysis on publicly-available datasets which correspond to each of those three linguistic contexts, and conducted a deep investigation of such sets by considering the linguistic features and dimensions proposed by Biber86 along with clustering approaches. With those sets we trained and evaluated different classifiers, using both the linguistic features used by Biber86 and more modern feature extraction methods, such as sentence embeddings provided by BERT [11]. Finally, we applied the trained classifiers on real-world chatbots data, from which we took advantage of TTS tags (such as SSML) and emojis to define a level of ground-truthness for quantitative data analysis.

The main observed outcome is that the majority of the real-world chatbot texts tend to be classified more frequently as either speech or CMC, and consequently closer to the desired linguistic style for such TTS systems. Another observation is that there is a great deal of classification confusion between CMC and the speech, which can either urge for more specific features to differentiate samples from these styles, or indicate that these types of language styles are more similar than previously thought.

II. Related work

Studies in Linguistics which analyze differences between written and spoken language styles are by no means a new field of study, with the first attempts dating back to the 1980s [12]–[15]. Some of those studies [14], [15] intended to find characteristics to distinguish speech and writing. Those differences focused on how language is produced and received. In general, since writing is slower than speech, it allows written language to be more planned, less fragmented, and more syntactically integrated than speech. For instance, [7] suggested that written text has a greater diversity, more difficult words, more nouns, and more adjectives than spoken language.

The research presented in [8], [10] has been very influential in the area. [8] performed a quantitative investigation of spoken-vs-written differences by analyzing large collections of spoken and written data. Applying *Multidimensional (MD)* analysis to that data, Biber could identify a set of linguistic syntactic features which represent underlying communicative functions served in English. The co-occurrence of a group of features is called a *dimension* and different co-occurrence patterns represent different dimensions.

More recently, using the same six dimensions identified by [8], and inspired by the increasing production of *Computer-Mediated Communication (CMC)* (e.g., the kind of language used in Internet and mobile text-based chat), [9] investigated the linguistic characteristics of two genres of CMC: synchronous and asynchronous chats. Using the same methodology as in [8], the MD analysis was applied to four different datasets: a conversational writing corpus composed by Internet relay chat and split-window ICQ, a spoken corpus based on a subset of the Santa Barbara Corpus [16], and the original spoken and writing corpus from [10]. Jonsson observed that certain CMC genres, when modeled using Biber's dimensions, presented features which resembled spoken conversations or showed a certain degree of orality.

For this paper, we will take advantage of the findings of Biber's and Jonsson's works to build our proposed model. Notice that the linguistic features used on those works might be more indicative of subtle language style than more modern approaches, such as sentence embeddings, which are mainly built upon the idea of content-encoding.

III. Data Validation

The first step of our exploratory study was to evaluate Biber's linguistic features for building up a classifier for language style. It consisted of performing data definition and validation using gathered public-available datasets and applying data analytics procedures with the purpose of better understanding such datasets.

A. Data Definition

As defined by jonsson15, the language style of a text can be classified into three main classes: writing, CMC, and speech. One goal of this work is to be able to define datasets which represent such classes but, unfortunately, the datasets of Jonsson's research are not publicly available. For this reason, we looked for publicly-available datasets which could be viable alternatives for those.

For written style, one type of data which we believe to directly represent such language style is *Wikipedia* articles. Generally, such articles provide descriptions for a given subject, represented by texts providing some level of writing features such as discussed before. Thus, for this work to be as reproducible as possible, we make use of two publicly-available datasets created with Wikipedia articles: *Question-Answer (QA)* [17] and *WikiLarge (Wiki)* [18]. While the former contains factoid question/answer pairs, the latter comprises parallel pairs of texts used for text simplification tasks. These two sets consist of a total of 624,749 samples.

For spoken style, we consider four English-language datasets which contain transcriptions of human-to-human conversations: *Switchboard (SwitchB)* [19], *Santa Barbara Corpus (SBC)* [16], *Taskmaster2 (TM2)* [20], and the user side of *Coached Conversational Preference Elicitation (CCPE-User)* [21]. SwitchB consists of telephone conversations of about 70 different topics. SBC is mainly composed of face-to-face conversation but the dataset also contains other data claimed as "other ways that people use language in their everyday lives", such as telephone conversations, card games,

TABLE I: Summary of the different datasets used in this study.

Style	Dataset	#Full	#Filtered (%)
Writing	QA	29,243	23,457 (80.2%)
Writing	WikiLarge (Wiki)	595,506	427,777 (71.8%)
CMC	CCPE-Agent (CCPE-A)	5,594	3,926 (70.2%)
CMC	IRC	219,825	100,676 (45.8%)
Speech	Taskmaster2 (TM2)	13,953	5,909 (42.3%)
Speech	CCPE-User (CCPE-U)	6,376	3,811 (59.8%)
Speech	Santa Barbara Corpus (SBC)	29,242	15,929 (54.5%)
Speech	Switchboard (SwitchB)	199,762	103,645 (51.9%)

on-the-job talk, and so on. TM2 is composed of person-to-person conversations in a Wizard-of-Oz-based spoken dialogue system where the users interact with a real person believing it was an automated system. The CCPE dataset is somewhat similar to TM2 but only the user side of the dialogue is composed of real speech. Summing up all the utterances from the four datasets, there is a total of 249,333 samples available.

Finally, for the CMC category, we aimed at gathering data representing texts created by users in some chat-based environment. With that in mind, we collected two datasets: the *IRC disentanglement dataset* [22], and the agent side of *CCPE (CCPE-Agent)* [21]. The IRC dataset comprises group-based dialogues on Internet Relay Chat rooms, and the CCPE-Agent contains texts which were converted to speech with text-to-speech (TTS) tools. Therefore, in total, the CMC datasets provide a total of 225,419 samples.

A detailed list of the datasets is presented in Table I. It is worth mentioning that, for each dataset, we perform just basic pre-processing steps, such as tokenization and removal of unwanted tokens (e.g. usernames), which may negatively affect the computation of linguistic features. The last column of that table is about a filtering procedure explained later in the paper.

B. Datasets Analysis

With the previously-mentioned datasets, we conducted some data analyses to validate whether the datasets of choice are actual representations of the three language styles, and also to prepare the data for removing noise for training machine learning classifiers.

This analysis considers the 47 linguistic features used in [10]. Those features take into account metrics such as the average token ratio, the occurrence of first-person pronouns and of present-tense verbs, just to name a few, which can provide evidence of a particular language style. For instance, the presence of first-person pronouns can be a strong indication of speech, while third-person pronouns can point towards a writing style.

For computing Biber's linguistic features we make use of the *Biberpy Python* package,[1] which unfortunately contains implementations for only 42 of the 67 features. Luckily, those features are the most common ones. Notice that previous works such as Biber86 and jonsson15, as far as we understood, heavily depended on manual annotation of such linguistic features. Thus, even though not all the linguist features are implemented, we believe that a sufficient set of them is likely to be enough for conducting an automated analysis. With that, we have been able to associate to each text the occurrence of each of each of the 42 linguistic features. And, similarly to the work in [9], those occurrences have been normalized by the occurrence at every 1000 words.

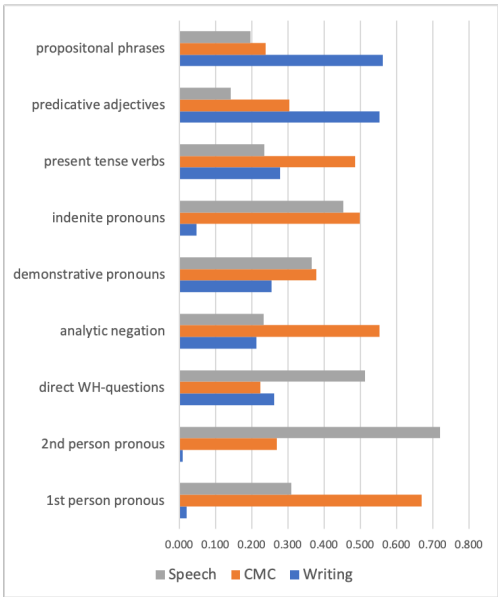
The first validation consisted of comparing the normalized frequency of nine different features with the values reported in [9, p. 150, Table 4.6],[2] which is depicted in Figure 1. Notice that we perform this validation using datasets different than those used in [9], and with computer-determined features instead of manual annotators. As observed, the frequencies are not identical but they preserve some similarities, such as the highest occurrence of propositional verbs in writing, and the prevalence of indefinite pronouns in CMC and speech. The largest differences are related to the CMC style, which can be expected since our data contains more truly conversational CMC contexts.

Following, we shifted our focus to validate, quantitatively, whether the linguistic features would provide enough evidence to differentiate the datasets and, as a consequence, allow us to train a machine learning classifier. Our first attempt consisted of making use of Biber's six dimensions, which consists of summing up a given set of features for each dimension [9, p. 53, Table 2.2]. Unfortunately, that first trial was unsuccessful, provided that we have not been able to observe patterns in the computed dimension values, considering mean, standard deviation, range, minimum and maximum values.

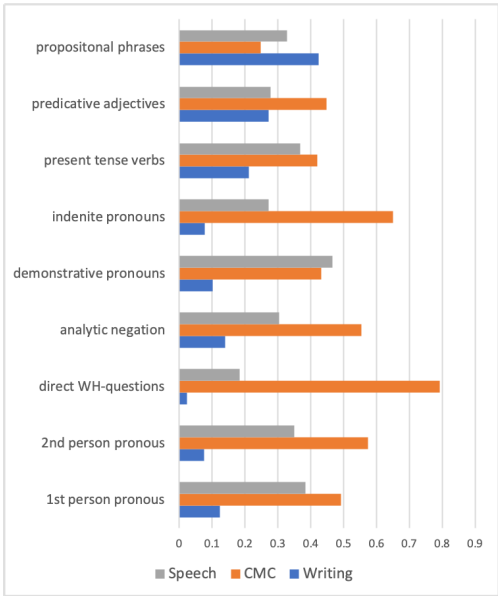
Next, we moved to a clustering-based strategy, with the following premise: if the data is clustered, we can evaluate the frequency of occurrence of samples in specific clusters as an indication of the separability of the datasets. Thus, we applied the k-means algorithm, with K arbitrarily set to 20, on all of the linguistic feature vectors extracted from all samples. We then counted the occurrence of the samples in each cluster. The result, normalized in the 0-1 range, is presented in Figure 2. We observe that there is one cluster (number 19) which concentrates a large number of examples for all datasets. We observe also that datasets from the same class, such as QA and WikiLarge, tend to have a larger occurrence of samples in the same clusters, such as clusters 1 and 11. And the same can be observed with the spoken datasets in clusters 0 and 12, for instance. The only datasets that are more difficult to find a pattern from are the CMC ones, which sometimes tend to co-occur more with writing datasets and other times with speech. But that is somewhat expected, given that CMC tends to float between written and spoken styles, as reported by jonsson15.Notice that he CCPE-Agent dataset is highly present in clusters 6 and 16, suggesting that it may contain text of a different style than its CMC counterpart.

[1]https://github.com/ssharoff/biberpy

[2]For the CMC classes, we present an average of the three distinct CMC classes.



(a) This work



(b) [9, p. 150, Table 4.6]

Fig. 1: Frequencies of linguistic features in our 8 datasets, compared to Jonssons' results with its own datasets, using computer-determined features instead of manual annotators.

Figure 3 illustrates the enhanced ability of the proposed k-means based evaluation method to find similarities and dissimilarities among the language styles. As can be observed in Figure 3b, in general the proposed method results in lights colors, i.e. less similarity, when considering sets from different language styles. One remarkable observation is that, even though the IRC dataset in principal belongs to the CMC class, the results indicate that the language on the set is similar to speech. CCPE-A on the other hand, stands more apart from writing and speech.

IV. Model Development and Evaluation

In this section, we describe the development of the machine learning classifier models and the evaluation of the datasets described in the previous section. The methods described herein are evaluated with two distinct training sets: a) *Full dataset*, with all samples; and b) *Filtered dataset*, where samples from cluster 2, as depicted in Figure 2, were discarded because of their high frequency in almost all datasets. The number of filtered samples in each dataset is depicted in the last column of Table I, followed by the percentages they correspond to in the original datasets[3].

Our evaluation is based on a leave-one-dataset-out approach, where each dataset is used as the test set and the remainder as the training set. The final evaluation metrics are computed on the confusion matrix which aggregates all individual confusion matrices. For validating state-of-the-art (SOTA) metrics for that matter, and also to contextualize it into current's SOTA deep learning methods, we implemented two distinct methods.

The first proposed method uses Biber's linguistic features[4] to train a classifier. After experimenting with a few different base methods, such as Logistic Regression and Support Vector Machines, the best results were found with Multi-Layer Perceptron (MLP) neural network, with one 200-neurons hidden layer, trained with the Adam optimizer with the learning rate set to 0.001. We refer to this method as Biber-MLP for the sake of simplicity.

Table II presents the results of Biber-MLP models. In the full dataset, the models achieved an accuracy of about 46.40%, and in the filtered dataset, the accuracy increased to 62.92% (an increase of 16%). Thus, it is clear that filtering samples from confusing clusters improves the performance of the classifier, and especially provides a boost in detecting speech style. However, as observed in the data analysis, the CMC class was the hardest one to detect, and the best accuracy was only up to 2.28%. That class presents a high degree of confusion with the speech class, which demonstrates that CMC language style has many features similar to the spoken language.

We have then performed the same experiments with BERT-based neural networks, where Biber's linguistic features were replaced by BERT sentence embeddings [11]. Such embeddings consist of a 768-dimensional vector, computed with a pre-trained encoder based on a Transformer neural network. As mentioned, the idea is to compare the performance of the Biber's linguistic features against more recent advances made by deep learning in the NLP field. The main results are presented in Table III. As observed, this classifier achieved an accuracy of about 66.03% on the full dataset, showing a considerably better result than Biber-MLP on this set. However, in the filtered dataset, BERT models performance decrease to up to 27.36% of accuracy.

[3]We are aware that the dataset are unbalanced and futher investigation should be done in that respect, but this is out of the scope of this work

[4]We use only the 31 features implemented in the Biberpy tool.

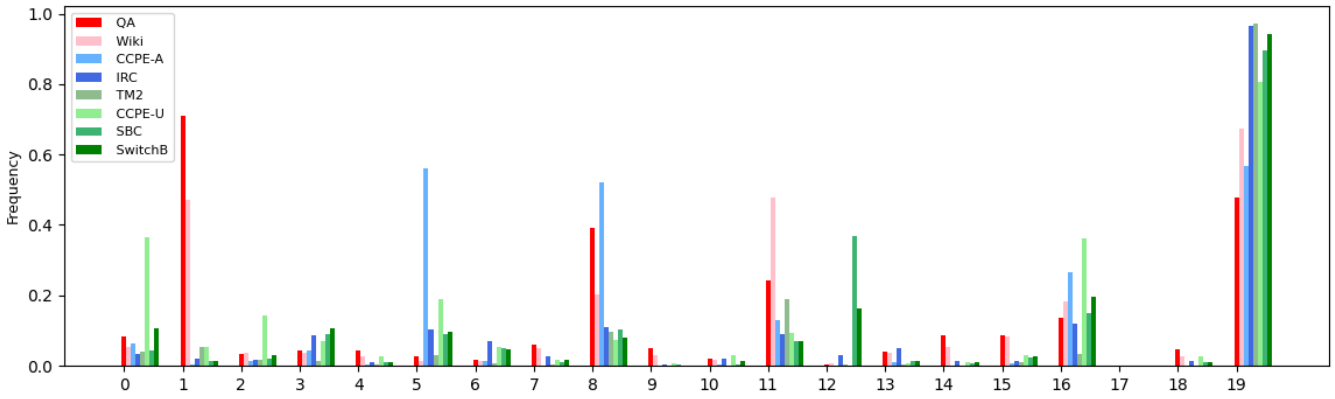


Fig. 2: Normalized occurrence frequency of texts from the 8 datasets in each cluster automatically generated by a k-means algorithm, K=20.

TABLE II: Confusion matrix for the Biber-MLP models, which achieved 46.40% of overall accuracy on the full dataset, and 62.95% on filtered. Class-specific accuracy is presented in the Acc. columns.

	Full dataset				**Filtered dataset**			
Classes	Writing	CMC	Speech	Acc (%)	Writing	CMC	Speech	Acc (%)
Writing	419,250	156,043	49,456	67.11	354,030	64,802	24,396	79.88
CMC	75,029	3,682	146,708	1.63	27,421	2,260	69,403	2.28
Speech	26,189	135,901	87,244	34.99	10,392	46,114	55,855	49.71

TABLE III: Confusion matrix of the BERT models, which achieved 66.03% of overall accuracy on the full dataset, and 27.36% on filtered. Class-specific accuracy is presented in the Acc. columns.

	Full dataset				**Filtered dataset**			
Classes	Writing	CMC	Speech	Acc (%)	Writing	CMC	Speech	Acc (%)
Writing	556947	40243	27559	89.15	57112	192000	375637	9.14
CMC	125653	5609	94157	2.49	904	849	223666	0.38
Speech	12599	72851	162707	65.57	454	5186	242517	97.73

In Figure **??** we present fine-grained plots of the results with both models, where each dataset is considered as a different style. The plots represent the confusion matrices for each style, considering the normalized confusion matrices. Considering Biber-MLP models in Figure 4, we see that the datasets belonging to both writing and speech styles tend to present a clear intra-style confusion, which indicates that such datasets, although extracted from different sources, present similar content. One exception is SwitchB, which, despite presenting a high degree of confusion with SBC, presents more confusion with the CCPE-A dataset, which belongs to the CMC class. And that behavior is even more present in the results produced with TM2 as the test set. Thus, although we can find some datasets that strongly represent some styles, such as QA for writing and SBC for speech, some datasets may belong to a gray area where it is hard to find distinction among them. Similar behavior can be observed with BERT, as depicted in Figure 5. Differently from Biber-MLP models results, we notice a higher number of examples classified as IRC using BERT models. Especially with SwitchB, CCPE-A and IRC behave quite differently for the two methods. But still, the presented confusion is similar, since it is between a dataset from the speech class, i.e., SBC, and a CMC one, i.e. either CCPE-A or IRC. As a consequence, these results are not very conclusive in eliciting evidence that one method has significantly superior performance to the other.

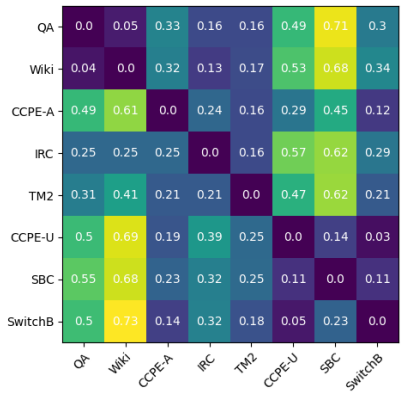
V. Chatbot Data Evaluation

To contextualize this work in a real-world application, we applied the previously-built models (Biber-MLP and BERT), onto a set of datasets from real-world chatbots. The purpose of this evaluation is not only to validate our models on real scenarios but also to evaluate the language that was used on the data of such systems. For achieving this goal, we present in this section a mix of quantitative and qualitative analysis.

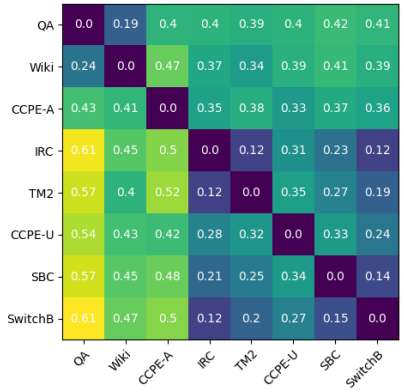
The dataset used for this analysis is composed of the utterances which can be produced by 4,822 different chatbots [5]. Those chatbots were developed on a commercial platform [6] and we had any influence in their development. Notice that they can be related to several different domains, making them very suitable for this analysis which aims to better characterize

[5]The chatbots data were made available, by their authors, to the parent company of the authors for product improvement.

[6]For anonymity reasons, we will not disclose the name of the platform but plan to do it in a camera-ready version.



(a) Biber's dimensions



(b) Proposed approach

Fig. 3: Heatmaps illustrating the similarities among the language styles, where darker colors represent a higher degree of similarity, comparing the original dimensions proposed by [8] and the proposed clustering-based method.

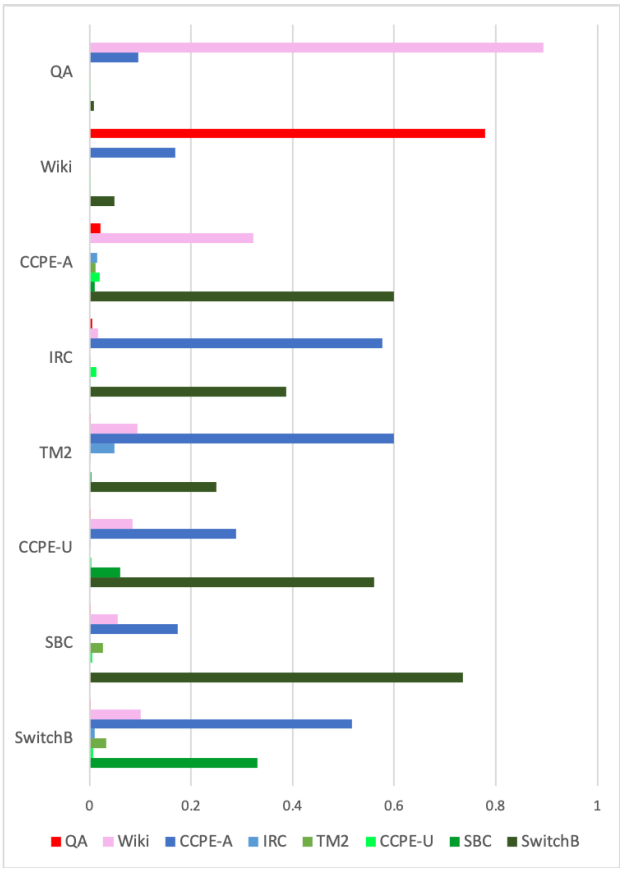


Fig. 4: Plot of the confusion matrix for the Biber-MLP

the language styles employed by chatbot developers, who are generally domain experts but have limited writing abilities.

That said, for the quantitative analysis, we took advantage of some specific features which may indicate whether the chatbot was specifically designed for speech use (i.e., to be coupled with a TTS service) or CMC. With such features, we could define something similar to a ground-truth since they clearly expose the main intention of the developer regarding the type of conversational interface on which the chatbot would be used: either spoken or text-based only.

Two features drove our selection of chatbots. The first feature comprises tags which are specifically used to setup TTS services, which we refer to as TTS tags, such as SSML. The other one is the use of emojis, which can strongly characterize a CMC chatbot. Since there are chatbots in the set that which present both tagged and non tagged texts, we considered all texts from the same chatbot as belonging to the same class, i.e., either speech or CMC, but specify the samples that do not contain those features with the *Other* suffix.

That procedure resulted in four datasets for this evaluation: *Speech-TTS*, with 746 texts containing TTS tags; *Speech-Other*, composed of 1,922 texts not containing TTS tags but that belong to the same chatbots which contain *Speech-TTS* samples; *CMC-Emoji*, comprising 461 text samples containing emojis; and *CMC-Other*, with 16,007 text examples not containing emojis but that belong to the same chatbots that contain *CMC-Emoji* samples. These datasets sum up a total of 19,136 text samples, extracted from 134 chatbots, where 46 of which contained TTS tags and 88 contained emojis.

The main results, represented by normalized confusion matrices with the classes, are presented in Figure 6. Overall we observe that all four sets are more frequently classified as CMC style, using both Biber-MLP and BERT models. It is interesting to notice that in the set supposed to be more speech-like (Speech-TTS), the number of samples classified as speech is about twice as high as those classified as writing. In our opinion, that shows that the TTS tags can indeed point out to texts with a linguistic style closer to speech, which is desirable for good TTS performance. Considering the CMC-Emoji set, the results with the BERT model indicate that this classifier has superior classification performance compared to Biber-MLP, given that about 80% of the samples were classified as CMC, while using Biber-MLP model this number is only up 50%. We note also that for both Speech-Other and CMC-Other, there is a higher occurrence of samples classified as a written style than spoken style. That might be an indicator that the developer

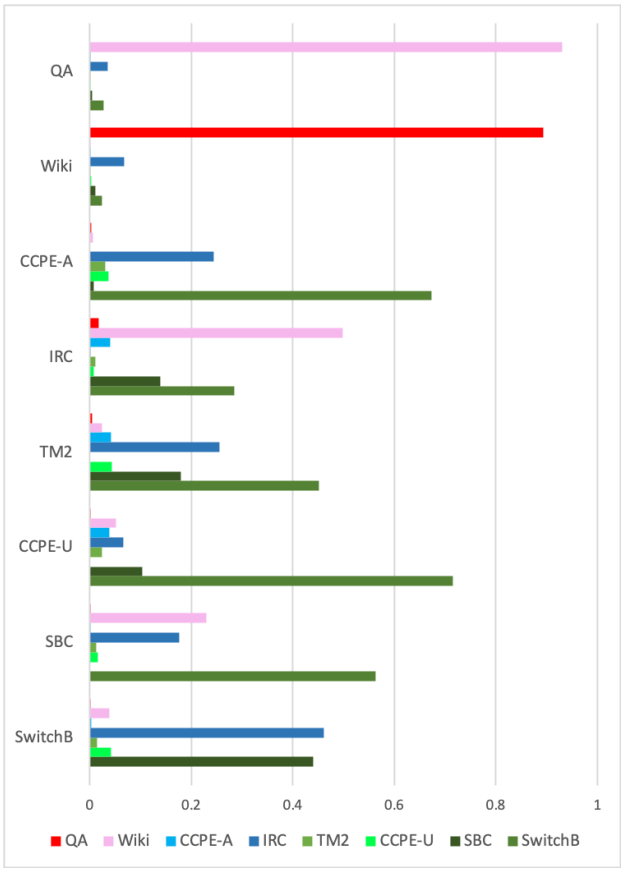
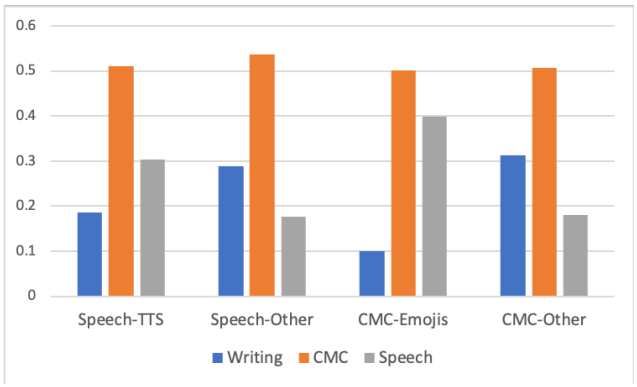
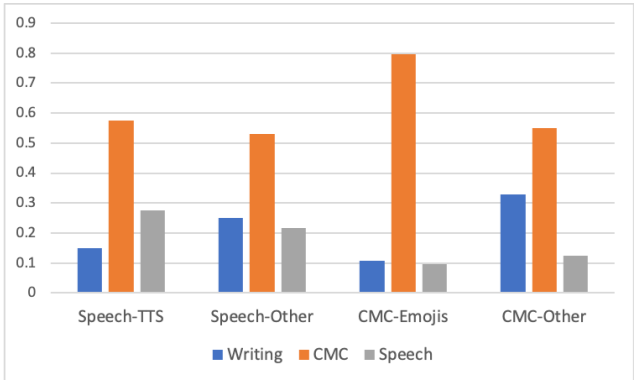


Fig. 5: Plot of the confusion matrix for BERT



(a) Biber-MLP



(b) BERT

Fig. 6: Plot of the detailed confusion matrices for real-world chatbot datasets.

has put more effort to adapt the text to the language that will be used with a specific user interface in mind, such as when TTS tags are used for voice-based interfaces or emojis for conversational ones.

Table IV provides some example sentences from the four chatbot datasets, along with the linguistic style provided by the classifier. Despite the originating dataset, the classification follows some of the typical patterns for the writing, speech, and CMC pointed out by [8]. For instance, in the selected examples, longer texts are usually classified as writing. Texts from writing also do not present singular first-person pronouns along with the text. For the speech and CMC classes, on the other hand, shorter texts with first-person pronouns are presented. However, the distinction between speech and CMC is more difficult, since both present conversational characteristics. One interesting difference among such classes is the acknowledge of human senses in some examples classified as speech, such as references to hearing (third example) and taste (ninth example), marked in bold in the text.

VI. Conclusion

The development of a classifier for language style demonstrated to be challenging, in special considering that the CMC language style seems to have very subtle differences compared to speech. However, we observe that the classifier is generally able to distinguish between writing and speech considerably well. Regarding the comparison of more traditional linguistic metrics as feature set (i.e., Biber's features) vs. deep learning sentence embeddings, we did not observe any significant difference between the methods. Thus, both methods can be reasonable options when developing such kind of classifiers. But the results on the real-world chatbots suggest that the model using BERT sentence embeddings can be better at identifying texts belonging to the CMC class.

Lastly, the evaluation of data from real chatbots indicates that generally, their texts follow a CMC-like style, which would be expected. Nevertheless, in the set with TTS tags, the classifiers were able to find a higher number of speech examples, indicating that chatbots' developers may tailor some of the texts for better results with TTS systems. Consequently, we believe that improving the classification tool can be very helpful to help detecting texts that need further adaptation or the aid of writers with more sophisticated language skills, such as journalists or screenwriters.

References

[1] B. Reeves and C. I. Nass, *The media equation: How people treat computers, television, and new media like real people and places.* Cambridge university press, 1996.

[2] C. I. Nass and S. Brave, *Wired for speech: How voice activates and advances the human-computer relationship.* MIT press Cambridge, MA, 2005.

TABLE IV: Sample sentences from the 4 chatbot datasets and their classified style.

Dataset	Class	Text
Speech-TTS	Writing	- If you qualify for an electronic transfer to your bank account, that is the fastest way for us to issue your payment. Otherwise, the payment may take a few business days to reach you.
	CMC	- Should we send the details to the contact number you provided earlier?
	Speech	- I **heard** that you came here not only for the boat trip but also to practice your English, am I right?
Speech-Other	Writing	- Condominium and homeowner association fees are not included in your monthly mortgage payments. You are responsible for paying condominium and the fees directly to your association.
	CMC	- Ana, I'm glad I could assist you with your questions. Always here for you. Have a great day!
	Speech	- Hi there. How can I help you this morning?
CMC-Emoji	Writing	- Our aim is to provide accessible medical specialist healthcare to everybody. (:D) We connect you with the doctor you need without the need to travel.
	CMC	- Ok, I'm here if you need me! Take care (:D)
	Speech	- That would be the chicken burrito (lol) **So good!** Would you like a recipe?
CMC-Other	Writing	- Our store is located in front of Terminal B of the Airport and it will be directly accessible from the Arrival and Departure levels. The store can also be accessed via Terminals C and D via link bridges.
	CMC	- Yay! Great news you want to rebook with our platform and we would love to have you stay again with us. We open for rebooking during term 1 so be sure to keep an eye out around site for the latest information and offers.
	Speech	- Sure, I've cancelled what we were doing. Is there anything else I can help you with today?

[3] C. Nass and K. M. Lee, "Does computer-synthesized speech manifest personality? experimental tests of recognition, similarity-attraction, and consistency-attraction." *Journal of experimental psychology: applied*, vol. 7, no. 3, p. 171, 2001.

[4] C. Nass and S. Najmi, "Race vs. culture in computer-based agents and users: Implications for internationalizing websites," 2003.

[5] H. Giles and K. R. Scherer, *Social markers in speech*. Cambridge,[Eng.]; New York: Cambridge University Press; Paris: Éditions de ..., 1979.

[6] S. T. Fiske and S. E. Taylor, *Social cognition: From brains to culture*. Sage, 2013.

[7] J. A. DeVito, "A linguistic analysis of spoken and written language," *Central States Speech Journal*, vol. 18, no. 2, pp. 81–85, 1967.

[8] D. Biber, *Variation across Speech and Writing*. Cambridge University Press, 1988.

[9] E. Jonsson, *Conversational Writing - A Multidimensional Study of Synchronous and Supersynchronous Computer-Mediated Communication*. Peter Lang International Academic Publishers, 2015.

[10] D. Biber, "Spoken and written textual dimensions in english: Resolving the contradictory findings," *Language*, vol. 62, no. 2, pp. 384–414, 1986.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[12] D. Tannen, C. Li, S. Thompson, P. Tannen, R. Freedle, S. Heath, G. Green, J. Goody, A. Hildyard, D. Olson *et al.*, *Spoken and Written Language: Exploring Orality and Literacy*, ser. Advances in discourse processes. ABLEX Publishing Corporation, 1982. [Online]. Available: https://books.google.com.br/books?id=xJxsAAAAIAAJ

[13] M. Halliday, *Spoken and Written Language*, ser. Language and Learning Series. Deakin University, 1985. [Online]. Available: https://books.google.com.br/books?id=T9RpAAAACAAJ

[14] W. Chafe and D. Tannen, "The relation between written and spoken language," *Annual Reviews Anthropology*, vol. 16, pp. 383–407, 1987.

[15] W. Chafe, J. Danielewicz, B. C. f. t. S. o. W. University of California, C. for the Study of Writing at Berkeley, and C. Mellon, *Properties of Spoken and Written Language*, ser. Technical report (Center for the Study of Writing at Berkeley and Carnegie Mellon). Center for the Study of Writing, 1987.

[16] D. Bois, J. W., W. L. Chafe, C. Meyer, S. A. Thompson, R. Englebretson, and N. Martey, "Santa barbara corpus of spoken american english, parts 1-4," *Philadelphia: Linguistic Data Consortium*, 2005.

[17] N. A. Smith, M. Heilman, and R. Hwa, "Question generation as a competitive undergraduate course project," in *Proceedings of the NSF Workshop on the Question Generation Shared Task and Evaluation Challenge*, 2008.

[18] X. Zhang and M. Lapata, "Sentence simplification with deep reinforcement learning," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2017, pp. 595–605. [Online]. Available: http://aclweb.org/anthology/D17-1063

[19] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing - Volume 1*, ser. ICASSP'92. USA: IEEE Computer Society, 1992, p. 517–520.

[20] G. Research, *Taskmaster-2*, 2020 (accessed June 22, 2020). [Online]. Available: https://research.google/tools/datasets/taskmaster-2/

[21] F. Radlinski, K. Balog, B. Byrne, and K. Krishnamoorthi, "Coached conversational preference elicitation: A case study in understanding movie preferences," in *Proceedings of the Annual SIGdial Meeting on Discourse and Dialogue*, 2019.

[22] J. K. Kummerfeld, S. R. Gouravajhala, J. J. Peper, V. Athreya, C. Gunasekara, J. Ganhotra, S. S. Patel, L. C. Polymenakos, and W. Lasecki, "A large-scale corpus for conversation disentanglement," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 3846–3856. [Online]. Available: https://www.aclweb.org/anthology/P19-1374