DIALOGUE PIDGIN TEXT ADAPTATION VIA CONTRASTIVE FINE-TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

The surging demand for multilingual dialogue systems often requires a costly labeling process for each language addition. For low resource languages, human annotators are continuously tasked with the adaptation of resource-rich language utterances for each new domain. However, this prohibitive and impractical process can often be a bottleneck for low resource languages that are still without proper translation systems nor parallel corpus.

In particular, it is difficult to obtain task-specific low resource language annotations for the English-derived creoles (e.g. Nigerian and Cameroonian Pidgin). To address this issue, we utilize the pretrained language models i.e. BART which has shown great potential in language generation/understanding – we propose to finetune the BART model to generate utterances in Pidgin by leveraging the proximity of the source and target languages, and utilizing *positive* and *negative* examples in constrastive training objectives. We collected and released the first parallel Pidgin-English conversation corpus in two dialogue domains and showed that this simple and effective technique is suffice to yield impressive results for English-to-Pidgin generation, which are two closely-related languages.

1 INTRODUCTION

Task-oriented dialog systems are becoming prevalent in various daily activities such as ticket booking and restaurant reservations Peng et al. (2020). In a typical task-oriented dialog system, the natural language generation (NLG) module plays a crucial role by converting semantic representations into responses in natural language Langkilde & Knight (1998). As such, NLG plays a significant impact on the users' experience.

Unfortunately, these dialogue systems are mostly built for high resource languages such as English, which makes it less user-friendly in regions where English is not widely-spoken or is spoken differently Khalil (2020); Yusupujiang & Ginzburg (2021). This includes African countries such as Nigeria and Cameroon where the demand for these technologies are growing at a rapid rate Khalil (2020). There is then a need to adapt[1]/translate these existing well-developed dialogue systems in high resource languages into its target language counterparts. However, techniques in machine translation are still under-developed as some language pairs are completely devoid of parallel corpus Adelani et al. (2021); Chang et al. (2020). This situation is further exacerbated by the fact that the targeted NLG requires a domain-specific translation and annotation.

To this end, we present a preliminary study on the domain-specific adaptation of high resource English language model into the English-related Pidgin languages consisting of the Nigerian (Naija) and Cameroonian Pidgin (Yaounde). In particular, we explore on a low resource scenario where the goal is to translate the English utterance into conversation (dialogue) Pidgin sentences. This scenario consists of a few parallel data and a larger quantity of monolingual text in both languages. To facilitate the research, we collected the first parallel English-Pidgin dialogue corpus and release it along with the public Nigerian (Naija) Ogueji & Ahia (2019) and Cameroonian Pidgin spoken corpus Green et al. (2016). We further propose a simple technique that allows to finetune a pretrained English language model into generating Pidgin that can be used to collect dialogue text in dialogue

[1] We use the word "adapt" here since the translation process can slightly change the content for the regional end users.

systems. To do so, we employ the technique of contrastive learning where the off-the-shelf English language model adjusts its useful prior knowledge into producing Pidgin, which is a closely-related language Ayafor (2008); Faraclas (2008). This process is enhanced by having the model to discern between positive/negative examples as defined by (1) the in/out-domain English utterances, and (2) the English and Pidgin texts. Overall, we made the following contributions:

1. We release the first parallel data consisting of ∽ 200 Nigerian/Cameroonian Pidgin and English pairs in multiple dialogue domains including the restaurant and drone simulation.
2. We showed the efficacy of the proposed constrastive finetuning technique as being both simple and effective in creating natural Pidgin text with high-fidelity.

2 Related Work

Contrastive learning has been widely used in various tasks – language modeling (Huang et al., 2018), unsupervised word alignment (Liu & Sun, 2015), caption generation (Mao et al., 2016; Vedantam et al., 2017), and machine translation (Yang et al., 2019). Representations are learned with contrasting positive pairs and negative pairs: Chopra et al. (2005); Weinberger & Saul (2009); Schroff et al. (2015) leverage a triplet loss to separate positive examples from negative examples in metric learning. Chen et al. (2020) shows that contrastive learning can boost the performance of self/semi-supervised learning in computer vision tasks.

In natural language processing, contrastive learning has been widely used. (Mikolov et al., 2013) predicts neighbouring words from context with noise-contrastive estimation (Gutmann & Hyvärinen, 2012) while Logeswaran & Lee (2018) samples two contiguous sentences for positive pairs and the sentences from other document as negative pairs. Our work is in the same vein where positive/negative examples are provided to enhance cross-lingual sentence representations.

3 Relations of Pidgin to English

Variations of Nigerian Pidgin are spoken across West and Central Africa in countries such as Benin, Ghana, and Cameroon Faraclas (2013) as a result of contacts between Africans and English-speaking sailors and traders Gilman (1980)[2]. Altogether, these languages evolved into a closely related group of languages whose vocabulary is predominantly English, spoken by Africans in West Africa and by their descendants in the Western Hemisphere Gilman (1980).

Importantly, we *postulate that the relatedness of the Pidgin language to English can be utilized* to formulate an effective contrastive learning setup such that an English language model can be readily adapted into its Pidgin variant. This assumption is not unfounded as English is the lexifier language of Pidgin language Gilman (1980), thereby consisting of a large amount of loaned English vocabulary. This motivates our use of an English language model as a strong prior to generate Pidgin text.

4 Methodology

In order to adapt the English language model to generate Pidgin dialogue texts, we propose a contrastive learning framework to expose the model to various valid or incorrect output sequences for a given input sentence. Specifically, we train the model in a two-stage process where for each stage its contrastive loss goes below the threshold 0.1:

(1) **Stage 1** is the *domain-targeting phase* where the language model is finetuned to generate in-domain dialogue utterances. As such, the primary source of positive samples are the English dialogue utterances. Further, we include a general-domain, non-conversation English texts. Non-conversation English texts are closer in the embedding space to the dialogue English texts, so they serve as a meaningful source of negative examples that help to pull the embedding projections to separate in/out domain texts.

[2] It is also possible that the English of the sailors was itself pidginized before contact with the Africans, as a result of the multilingual nature of the ships' crews.

Stage 1:
$+$ **(Dialogue English)**: There is a pub Blue Spice in the riverside area.
$-$ **(English)**: Mauritius was voted Vice President.
Stage 2:
$+$ **(Pidgin)**: Na for di main general hospital for di regional capital, Mekelle dem dey treat sick pipo.
$+$ **(Dialogue Pidgin)**: One pub Blue Soice dey for riversde area.
$-$ **(English)**: We treat our youth differently because they are just coming up.

Figure 1: Positive ($+$) and negative ($-$) examples.

(2) On the other hand, **stage 2** is the *language-converting phase* that primarily aims to ensure that the model learns to distinguish Pidgin from English sentences. In this phase, all monolingual Pidgin texts are defined to be the positive samples. In order to induce the language model to generate the corresponding Pidgin texts in stage 2, one source of negative samples are thus the randomly sampled English sentences that exclude the dialogue English texts. We obtained these monolingual English sentences from a general domain and force the model to predict them to be negative. Lastly, when available for the setting, we consider the few-shot Pidgin dialogue utterances obtained from the dialogue system can be used as positive samples. We display some examples in Figure 1.

Both stages follow the contrastive learning framework (Chen et al., 2020), where we train the model to learn the representations of the ground truth sentence by contrasting the positive pairs with the negative pairs. We project the source and target text sequences onto the latent embedding space. Then we maximize the similarity between the pair of source and target sequence; while minimizing the similarity between the negative pairs as follows:

$$\mathcal{L}_{cont}(\theta) = \sum_{i=1}^{N} \log \frac{\exp(\text{sim}(z_x^{(i)}, z_y^{(i)})/\tau)}{\sum_{z_y^{(j)} \in S} \exp(\text{sim}(z_x^{(i)}, z_y^{(j)})/\tau)} \tag{1}$$

$\mathbf{z}^{(i)} = [\mathbf{h}_1^{(i)} \cdots \mathbf{h}_T^{(i)}] \in \mathbb{R}^{d \times T}$ is a concatenation of the decoder hidden states $\mathbf{h}_t^{(i)}$ of the target sentence $y^{(i)}$ across the sequence of length T for sequence S. Sim$(\cdot, \cdot)$ is a cosine similarity function and τ is the temperature factor set to 0.5.

Sequence-to-sequence Finetuning. To fine-tune the pretrained BART Lewis et al. (2020) models for generation, we assume a dataset where each example (x, y) is an (English, Pidgin) pair. We train the student model using the standard cross entropy loss:

$$\mathcal{L}_{\text{seq}} = -\sum_{t=1}^{T} \log p(y_{t+1}|y_{1:t}, x) \tag{2}$$

where T in the target sequence length p is the model's predicted probability for the correct word. As such, the total objective for each batch update is given as $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{seq}} + \mathcal{L}_{\text{cont}}$.

5 Pidgin Dialogue Corpus

To create the Pidgin dialogue corpus, we consider two publicly available English dialogue datasets (i.e. E2E Novikova et al. (2017) of restaurant domain and DroneParrot[3])) and translate the English utterances into Pidgin. We show the statistics for each datasets in Table 1 and observe that the Pidgin language generally has less characters (e.g. "kontri" as opposed to "country") – owing to the evolution of its make-shift morphology where words are spelled in a reduced way based on its spoken, colloquial form Gramley & Pätzold (2004).

[3] Dialogue corpus for drone-human communication.

Do.	**Method**	**Un-Naija**	**Naija**	**Yaounde**	**N+Y**
E2E	COPY	**16.84**	16.84	0.00	-
	NMT	0.00	28.26	0.00	0.00
	XLM-R	3.57	33.69	2.51	2.65
	BART	2.12	49.86	3.76	3.38
	+Stage-1	7.42	52.42	4.29	3.56
	+Stage-2 (Ours)	8.46	**56.35**	**5.73**	**4.41**
Drone	COPY	0.28	0.28	0.00	-
	NMT	0.00	9.18	0.00	0.00
	XLM-R	1.39	20.71	1.37	1.39
	Ours	**3.28**	**34.62**	**2.63**	**2.55**

Table 2: English to Pidgin translation performance in BLEU-4 with Naija (N) and Yaounde (Y) in both domains (*Do.*).

	Domain	**Split**	**Naija**	**Yaounde**
Para	E2E	Train	40	-
		Dev	30	-
		Test	30	10
	Drone	Train	40	-
		Dev	30	-
		Test	30	-
Mono	General	Pd	$57,549$	$3,108$
		En	$57,549$	

Table 1: Statistics of the dataset. For each part of the dataset, the number of sentences for both monolingual (*Mono*) and parallel (*Para*) data in two dialogue domains.

While the language tend to have a smaller vocabulary set than its lexifier (i.e. English)[4], Pidgin diverges profusely in terms of syntax. Thus, the difference in word-ordering is where the anticipated challenges come from in terms of utterance adaptation.

6 Experiments

Training Details. We fine-tune the large BART Lewis et al. (2020) model for 200 steps using an Adam optimizer Kingma & Ba (2014) with $\beta_1 = 0.9$, $\beta_2 = 0.999$, 0.1 weight decay, 0.1 dropout, 0.1 attention dropout, 0.1 label smoothing, 6% warmup steps and a learning rate of 3e-5 (See Appendix). The final outputs are generated using beam search with a beam size of 3.

Compared Approaches. We compare the proposed technique with contrastive learning setup with various methods including simply using the source as target (**COPY**):

NMT: It is a semi/un-supervised technique that takes sentences from bilingual/monolingual corpora in two different languages and maps them into the same latent space via iterative back-translation Lample et al. (2018). In addition, we include the *supervised* bidirectional training objectives.

XLM-R: This approach finetunes the pretrained multilingual language model in Conneau et al. (2020) on parallel data following the proposed objectives as in **Ours** as we now discussed.

Ours: As our proposed approach, the pretrained language model BART Lewis et al. (2020) is finetuned as in §4 where **BART** simply uses the objective $\mathcal{L}_{\text{seq}}$, and **Stage-1** and **Stage-2** are incrementally added.

Experimental Scenarios. To validate our approach, we finetune **BART** with both Naija and Yaounde utterances. In Table 2, **Un-Naija** means that only the monolingual English and Naija texts

[4] English is the lexifier language of Pidgin where a large set of vocabulary are loaned.

are provided. **Naija** and **Yaounde** means training on parallel/monolingual data and testing on their respective languages; while **N+Y** means training on both Naija/Yaounde data and testing on Yaounde.

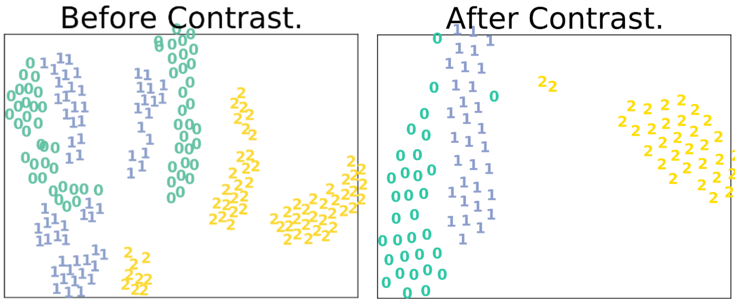
7 Results and Analysis

In Table 2, we display the results for experiments in both unsupervised and few-shot scenarios. We observe that even with an extremely small amount of annotations, the proposed technique (**Ours**) consisting of both training stages outperforms all benchmarks (**NMT** and **XLM-R**) by as much as 28.09 BLEU – which also goes to show that having multilingual representation (**XLM-R**) is not more beneficial. Next, we show the effectiveness of the contrastive learning setup where **BART** without contrastive losses is substantially lower than **BART+Stage-1** by 2.56 BLEU points, and 6.49 points lower than **Ours**. This shows that providing the positive/negative examples can help project the embedding space into more accurate representations for each examples. Further, we observe a consistent lower scores for the Yaounde Pidgin – which corroborates with past findings that indicates its influence from other sources of European languages such as the Portuguese and French Gilman (1980). However, when both *Naija* and *Yaounde* corpus are combined, we observe a drop in performance, which suggests that the morphological differences between the two languages (i.e. Naija and Yaounde) are interfering with the latent representations, causing it to performing sub-par.

Model	**E2E**		**Drone**	
	`Naturalness`	`Wrong`	`Naturalness`	`Wrong`
Reference	4.19	0	4.21	0
NMT	4.31	14	4.26	13
XLM-R	4.06	18	4.19	22
Ours	**4.66**	**13**	**4.35**	**9**

Table 3: Human evaluation on the generated Pidgin texts (100 instances) for all models in the few-shot scenario. Annotators were asked to evaluate the *naturalness* (0-5) and *wrong* (i.e. # hallucinated slots w.r.t. the slot-value pairs) of the texts.

Human Evaluations. We also perform human evaluation with two expert who are well-versed in both English and Pidgin and show the results on *naturalness* and *wrong* in Table 3. *Wrong* is a content selection metric that measures how accurate the generated Pidgin texts (See Appendix) are adhering to the dialogue semantic frames. We observe that results on both metrics are consistent with the BLEU-4 scores where our proposed approach consistently generate more natural text with fewer mistakes on both *Naija* and *Yaounde*.


Before Contrast.

After Contrast.

Figure 2: t-SNE projection *before* and *after* contrastive training for dialogue English (0), Naija (1) and general Naija (2).

Embedding Space Separation. To get a clear picture of the sentence representations, we visualize the respective embedding projections in Figure 2. We show the *before* and *after* projections of English and Pidgin utterances, which shows that the clusters of *dialogue English and Pidgin*, and *general Pidgin* sentences are more separated.

8 Conclusion

In this paper, we show that the proposed two-stage training approach can help to adapt the high resource English conversation texts into natural, high-fidelity Pidgin sentences in low resource scenarios. Moreover, we conclude that while Naija and Yaounde are two similar languages, augmenting with either languages provide no further improvements, while separating the representations with contrastive objectives is hugely beneficial.

References

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, et al. Masakhaner: Named entity recognition for african languages. *arXiv preprint arXiv:2103.11811*, 2021.

Miriam Ayafor. Cameroon pidgin english (kamtok): Morphology and syntax. In *A handbook of varieties of English*, pp. 2101–2120. De Gruyter Mouton, 2008.

Ernie Chang, David Ifeoluwa Adelani, Xiaoyu Shen, and Vera Demberg. Unsupervised pidgin text generation by pivoting english data and self-training. *arXiv preprint arXiv:2003.08272*, 2020.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *International Conference on Machine Learning, ICML*, 2020.

Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2005.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, 2020.

Nicholas Faraclas. Nigerian pidgin english: morphology and syntax. *Varieties of English: Africa, South and Southeast Asia*, 4:340–367, 2008.

Nicholas Faraclas. *Nigerian pidgin*. De Gruyter Mouton, 2013.

Charles Gilman. The origin of cameroonian pidgin dialects. *Anthropological Linguistics*, 22(9): 363–372, 1980.

Stephan Gramley and Kurt-Michael Pätzold. *A survey of modern English*. Routledge, 2004.

Melanie Green, Miriam Ayafor, Gabriel Ozón, et al. A spoken corpus of cameroon pidgin english: Pilot study. *Oxford Text Archive Core Collection*, 2016.

Michael U Gutmann and Aapo Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *The journal of machine learning research*, 2012.

Jiaji Huang, Yi Li, Wei Ping, and Liang Huang. Large margin neural language model. *Empirical Methods in Natural Language Processing, EMNLP*, 2018.

Shittu Adeshina Khalil. Artificial intelligence: The impact of chatbots in the nigerian banking sector. 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*, 2018.

Irene Langkilde and Kevin Knight. Generation that exploits corpus-based statistical knowledge. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*, 1998.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Annual Meeting of the Association for Computational Linguistics, ACL*, 2020.

Yang Liu and Maosong Sun. Contrastive unsupervised word alignment with non-local features. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence,*, 2015.

Lajanugen Logeswaran and Honglak Lee. An efficient framework for learning sentence representations. *International Conference on Learning Representations*, 2018.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2016.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 2013.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. The e2e dataset: New challenges for end-to-end generation. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pp. 201–206, 2017.

Kelechi Ogueji and Orevaoghene Ahia. Pidginunmt: Unsupervised neural machine translation from west african pidgin to english. *arXiv preprint arXiv:1912.03444*, 2019.

Baolin Peng, Chenguang Zhu, Chunyuan Li, Xiujun Li, Jinchao Li, Michael Zeng, and Jianfeng Gao. Few-shot natural language generation for task-oriented dialog. *arXiv preprint arXiv:2002.12328*, 2020.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015.

Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. Context-aware captions from context-agnostic supervision. *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017.

Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 2009.

Zonghan Yang, Yong Cheng, Yang Liu, and Maosong Sun. Reducing word omission errors in neural machine translation: A contrastive learning approach. *Annual Meeting of the Association for Computational Linguistics, ACL*, 2019.

Zulipiye Yusupujiang and Jonathan Ginzburg. Data collection design for dialogue systems for low-resource languages. In *Conversational Dialogue Systems for the Next Decade*, pp. 387–392. Springer, 2021.

A Training Configurations

All models were trained on 1 Nvidia V100 GPUs (32GB and CUDA Version 10.2) for 4k steps and averaged over 10 initialization runs for approximately 20 minutes each. All models are selected based on optimal validation BLEU4. Sentences are generated with a beam size of 5.

B Pidgin (Naija) Generation

Reference
If na pub wey get rating of 5 over 5 you wan pick Bue Spice.
Blue Spice be pub for riverside near Rainbow Vegetarian Café.
For riverside near Rainbow Vegetarian Café be one children-friendly pub wey dey get English food wey dem dey call Blue Spice.
Blue Spice na better pub along riverside. Blue Spice be family friendly pub wey dey serve Chinese for Riverside near Rainbow Vegetarian Café.
Blue Spice be restaurant wey dey provide Chinese food for riverside.
One family friendly fast food pub dey near Rainbow Vegetarian Café for riverside wey dem dey call Giraffe.
Blue Spice be fusion of restaurant and Chinese.
Pub wit fast food, Giraffe be child friendly wey dey riverside near Rainbow Vegetarian Café.
Pub wey dem dey call Giraffe wey be family friendly for riverside area near Rainbow Vegetarian Café.
Ours
Giraffe be pub wey dey riverside near Rainbow Vegetarian Café .
Blue Spice be family friendly pub wey dey riverside near Rainbow Vegetarian Café .
Blue Spice be pub wey dey near Burger King dey near Burger King .
Blue Spice be pub wey dey serve English food .
Giraffe be one family friendly pub wey dey riverside near Rainbow Vegetarian Café .
Blue Spice be pub wey dey riverside near Rainbow Vegetarian Café .
Blue Spice be restaurant wey dey provide Chinese food e dey riverside.
For riverside near Rainbow Vegetarian Café, one child friendly pub wey dem de Near Rainbow Vegetarian Café.
Giraffe be kids friendly pub wey dey riverside near Rainbow Vegetarian
If you want pub rated 5 over 5 pick Blue Spice.
NMT
Blue Spice be family friendly pub wey dey serve Chinese for riverside near Rainbow
One pub near Rainbow Vegetarian Café wey dem dey call Blue Spice.
Blue Spice dey pub dey riverside near Rainbow Vegetarian Café.
One coffee shop for city centre area wey dem dey call Blue Spice.
Near Burger King be pub wey dem dey call Blue Spice wey get average
Blue Spice be pub wey dey riverside near Rainbow Vegetarian Café .
Giraffe be one family friendly pub wey dey riverside near Rainbow Vegetarian Café .
Blue Spice be restaurant wey dey riverside .
Blue Spice be family friendly pub wey dey riverside near Rainbow Vegetarian Café for riverside .
Blue Spice be pub wey dey riverside .
XLM-R
Blue Spice wey dey serve English food.
One pub wey dem dey call Giraffe wey dey family friendly de
For riverside near Rainbow Vegetarian Café, one child friendly pub wey dem de
Blue Spice be restaurant wey dey provide Chinese food e dey riverside.
Giraffe be children friendly pub wey dey riverside near Rainbow Vegetarian
Blue Spice pub wey dey near Rainbow Vegetarian Café.
Blue Spice dey pub dey riverside near Rainbow Vegetarian Café.
Blue Spice pub wey dey near Rainbow Vegetarian Café.
A pub wey dem dey call Giraffe be child friendly dey riverside
Situated near Rainbow Vegetarian Café for riverside area of city.

Table 4: Samples of generation outputs and references for the restaurant domain.