7PERFECTION: Community Based Development of Commonsense Dataset for Seven Nigerian languages

Anonymous EMNLP submission

Abstract

Communications require understanding of physical activity and relative knowledge of the context of how the physical referents are encoded. This is what has been called common sense. Many studies have examined the process of using current neural networks to learn abstract concepts through bias and, some studies have examined and demonstrated how the large language models can learn physical activities in English. However, less attention has been paid to learning physical activities/commonsense in African languages. So, this study presents a data set on physical activities in 7 Nigerian languages. The languages are Yoruba, Hausa, Igbo, Idoma, Urhobo, Naija (Nigerian Pidgin English), and Isoko. The choice of language is randomly based on the community efforts of Linguistics Island.

1 Introduction

Languages have different displacements, which are filled in from the context of discussion. While it is very common to induce the meaning of a given expression within a communication situation, it is always challenging to make inferences from communication that require physical knowledge, for machines (Hespos and Spelke, 2004; Gao, 2020). Physical knowledge is described as commonsense wherein the interpretation of a given proposition is grounded. Communications require an understanding of physical activity and a relative knowledge of the context of how the physical references are encoded. This is what has been called common sense. Many studies have examined the process of using current neural networks to learn abstract concepts through bias and, some studies have examined and demonstrated how the large language models can learn physical activities in English. However, less attention has been paid to learning physical activities/commonsense in African languages. So, this study presents a data set on physical activities in 7 Nigerian languages. The languages are Yoruba, Hausa, Igbo, Idoma, Urhobo, Naija (Nigerian Pidgin English), and Isoko. The choice of language is randomly based on the community efforts of Linguistics Island[1].

This study is organized as follows; section 2 discusses the selected languages, section 3 discusses the methodology, section 4 presents the language data and the process of developing the dataset and section 5 discusses the expectation of the model results, given the selected languages.

2 Selected Languages

The languages for this study are randomly selected because they are not based on linguistics or other considerations. Seven languages are chosen for this study, of which about all of them are from languages spoken in Nigeria. Specifically, the languages are Yoruba, Hausa, Igbo, Idoma, Isoko, Urhobo, and Nigerian Pidgin English (Nigerian Pidgin English). Figure 1 shows the distribution and approximate locations of the languages on the Nigeria map. As shown in , the locations of the languages cover mostly southern Nigeria and places

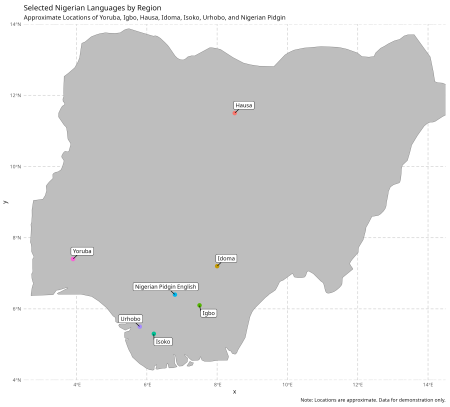

Figure 1: Location of where chosen languages are spoken

[1] Linguistics Island

Language	Phylum	Branch
Yoruba	Niger-Congo	Volta-Niger
Hausa	Afroasiatic	Chadic
Igbo	Niger-Congo	Volta-Niger
Idoma	Niger-Congo	Atlantic-Congo
Isoko	Niger-Congo	Atlantic-Congo
Urhobo	Niger-Congo	Benue-Congo
Naija	N/A	Creole

Table 1: Typological Classification of the languages

where Hausa is spoken. In terms of language family, Table 1 shows the languages in terms of their typological classifications. As shown in Table 1, most of the languages belong to Niger-Congo phylum but their branches differ. Yoruba and Igbo belong to the same branch. Idoma and Isoko belong to the same phylum and branch. Hausa do not share the same phylum with other languages ,but belongs to the chadic branch. Lastly, Naija is a creole and does not belong to any phylum or branch.

The motivation for explaining and providing the typological distribution of languages is that we expect to get similar results in languages that share the same phylum (Jeff, 2009). For example, the evaluation results for Yoruba and Igbo should be similar when compared with the Hausa result. Similarly, the result between Yoruba and Igbo should be more similar than the result compared to Urhobo or Isoko, because, although they belong to the same phylum their branches are different. We turn to a discussion of the methodology.

3 Methodology

The methodology adopted for this are prompt generation and their completions. The prompts are a mixture of questions and adjacency constructions. We have defined an adjacency construction as a pair of sentences that depend on each other for completion. The question prompts were divided into two broad types. These are the how and what question prompt. The How-prompt explains the quantification and measurements of any physical activity. The What-prompt explains the description of any given physical activity.

In the same way, the adjacency prompt has two types. The first is the complementizer adjacency prompt, which explains dependent clauses prompts because they need the main clause for completion. The second is the gerundive adjacency prompt,

Prompt	Types	Sample
How-prompt	how did you move?	I move about 2ft south
What-prompt	what is it like?	it has rough edges.
Complementizer	if you go	speak facing right
Gerundive	singing involves	opening mouth

Table 2: Sample of types of prompts and their completion

which explains a prompt that starts with gerundive verbs and equally needs an independent clause for its completion. We ensured that all of these types explained and / or described one or more physical activities.

Additionally, the completions were mostly answers to the question or a pair of the completions of the prompt. A sample of the question and adjacency pair prompts is given in Table 2.

The structure of the prompt and the completion is equally important. For the prompt, the ones that are question have a structure like what and how. The what shows a description of physical activities while the how explains the degree to which something is done or the how-to process of doing a given thing. Since we want both instances to express common sense of physical knowledge, we keep what and how within physical activities.

For adjacency completion, we sometimes convert the completions to make them have two pairs in which one complements the other. In some other cases, we used the answers only to create both the prompt and their completion. A sample of this is given in Table 2. We turn to the language data.

4 Language Data

4.1 Language Description

The structures of most languages are mostly similar. Syntactically, they are all subject verb object word orders. The structure of the phrases are mostly similar, too, though they slightly differ. The differences in the phrasal constructions do not affect the prompt structure described here. The linguistics generalization is that

I. *A question prompt is a statement having a wh-phrase, a noun phrase and a prepositional phrase such that its completion needs to provide answers to the prompt by beginning with the noun phrase, adjectival phrase and small clauses.*

The steps to the creation of all the question prompts and their completion are given as follows:

1. begin with a how/what question
2. provide the degrees of a given phenomenon
3. write an independent sentence
4. Create your answer by having a noun and verb phrase.

Although words and some structural patterns vary between the different languages, they are all in the format provided. In addition, we tried to mitigate the variation in the prompts as much as possible by adding dependent clauses in questions and independent clauses in the completions, as expressed in the generalizations.

Similarly, the adjacency prompts follow a pattern of dependent clauses and their completions as the independent clause. The complementizer prompt is a dependent clause, and the gerundive prompt is a gerund verb sentences.

II. *An adjacency prompt is a statement having a complementizer phrase and/or a gerund, a noun phrase and a prepositional phrase such that its completion needs to provide answers to the prompt by beginning with the verb phrase, noun phrase, adjectival phrase and small clauses.*

The step is given as follows:

1. begin with a complementizer or gerund verb
2. provide the specific noun phrase
3. write an independent sentence explaining the noun
4. Create your completion by verb phrase or full independent clause.

Throughout our prompt and completion, we tried to maintain the format by keeping track of the number of prompt types used in the dataset creation. We track that by noting them in a column but we removed the column before submitting the final paper draft. In the next section, provide a result of the dataset.

4.2 Result

In all the data[2], Yoruba has the highest number of sentences of about 1000 sentences.Igbo has more than 400 sentences. There are at least 200 sentences for Hausa and Naija while Idoma, Isoko and Urhobo have at least 100 sentences each. Each of the sentences was reviewed at least once by a member of the same language group. In cases in which the sentences were not up to 100, we combined the sentences as a single contribution and the person with the shorter sentences reviewed the document.

Langs.	Types	Expressions
Hausa	Prompt	Idan mutum ya yi tuntuße me zai faru?
	Sol0	Zai kurje ƙafarsa ko ya fadi
	Sol1	Zai kurje ƙafarsa ko ya tashi sama
Idoma	Prompt	A lẹ ọch ẹ ma igbo afleyi ọchochi na lẹ okunu hẹ tu ọ ku nma ọchi
	Sol0	Ọch ẹ ọ ma ge chilawọ ohi ku nma ọchi chọọ
	Sol1	Ọchẹ ọma ge chilawọ ohi ku nma ọ chi duu
Igbo	Prompt	Kedu ngwa digasi iche iche e ji esi nri ọha n'ala Igbo?
	Sol0	E ji ite ọla, osi ite ya na nkụ na nnukwu eku esi nri ọha na-ala Igbo
	Sol1	E ji iko ọla, osi ite ya na nkụ na nnukwu eku esi nri ọha na-ala Igbo
Isoko	Prompt	Eme ọ rẹ lẹliẹ emọ ewhurẹ kpohọ uwouewhurẹ?
	Sol0	Emọ ewhurẹ a re kpohọ uwouewhurẹ re a whurẹ eriarịẹ gbe iruo obọ,..
	Sol1	Iwhurẹ a re kpohọ uwouewhurẹ re a whurẹ eriarịẹ gbe iruo obọ,...
Yoruba	Prompt	báwo lo se lé gbe ẹrù tó bá wà lórí òkè àjà
	Sol0	ó lè gbé ẹrù náà nípà gígùn orí ijòkó
	Sol1	ó lè gbé ẹrù nípa wíwọ bàtà gíga

Table 3: Sample of Language Data

For the length, we varied the prompts and their completions. To extend the prompts and/or completion, we either add a prepositional phrase or a full

[2]The data for the project can be found here: Linguistics Island Data

dependent clause. This works for all languages.

5 Discussion and Conclusion

The purpose of this paper is to report and explain the strategies explored in creating a common sense physical activity data set for 7 selected Nigerian languages. The number of dataset is given in Table 4. The languages are arranged based on the size of their dataset.

The chosen languages are mostly from southern Nigeria. The form of creating prompts and completion was through the two linguistic generalizations. We find that two different prompts will help the dataset creation and that grouping them into themes will accelerate the process of prompt generation. So, we also segment our prompt within food, culture, education and technology themes. There are two broad types of prompts, question prompt and adjacency prompts. The question prompt has two parts, how and what question. The adjacency also has two parts, complementizer and gerundive. The completion starts mostly with a noun for the question prompts but adjacency prompts completion start with verbs or gerunds. For all the evaluations,

Language	Dataset
Yoruba	999
Igbo	455
Naija	249
Hausa	215
Urhobo	120
Isoko	109
Idoma	101
Total	2,248

Table 4: Total Language Data

our prediction is that there will be similarities in the results for languages with similar phylum.

References

Choi Y. Le Bras R. Zellers R. Bisk Y. Gao, J. 2020. Piqa: Reasoning about physical commonsense in natural language. *Computing Research Repository*, arXiv:1503.06733. Version 2.

S. J. Hespos and E. S Spelke. 2004. Conceptual precursors to language. *Nature*, 430:453–456.

Good Jeff. 2009. African languages and linguistic typology. *University of Buffalo*.