Study of Kumaoni Language for Natural Language Processing in End-to-End Conversation Scenario

Bhupesh Rawat[1], Ankur Singh Bist[2], Nidhi Mehra[3], Muhammad Faizal Fazri[4], Yochebed Anggraini Terah[5]

Faculty of Computer Applications[1,3], Faculty of Computer Science and Engineering[2], Faculty Sains and Technology[4,5]
Graphic Era Hill University[1,2,3] , University of Raharja
Road Society Area, Oghal Bhatta, Clement Town, Dehradun, Uttarakhand 248002[1,2,3]
Jl. Jenderal Sudirman No.40, RT.002/RW.006, Cikokol, Kec. Tangerang, Kota Tangerang, Banten 15117[4,5]
India[1,2,3], Indonesia[4,5]
e-mail: bhr222@gmail.com [1], ankur1990bist@gmail.com [2], nidhigehu@gmail.com [3], faizal.fazri@raharja.info [4], yochebed@raharja.info [5]

Abstract

Natural language processing is one of the essential activities of artificial intelligence. There is an excellent need for chatbots that require integration into artificial intelligence applications. The local language makes the process easier. Our research aims to analyze the possibilities and challenges of implementing a chatbot in the Kumaon language. We also provide a detailed survey of the Kumaon language and map it to other languages to make it easier to process it for industrial use. This chatbot can help with various needs and services in the Kumaon language. The method used in this research is a study analysis of the Kumaon language to deal with language extinction. The novelty in this research is a chatbot in the Kumaon language with end-to-end encryption so that the service user has good security.

Keywords: Kumaon, NLP, Chatbots, NLU.

1. Introduction

With the advancement of technology in every area of human life, most tasks have been made easier with automation. This is a result of the development of the times. For example, humans can concentrate more on high-priority tasks than on activities automated by machines or artificial intelligence. Many chatbots appear as part of an application that performs some core activity with several people in this context. For example, many websites have chatbot functions to answer user questions anytime and anywhere, but the language used is still very standard. Many companies use these chatbots to serve their customers 24x7. These chatbots use artificial intelligence to automate their processes. There are chatbots available in many languages to cater to the needs of local users.

■ 143

The information obtained by chatbots is a highly unstructured and free structure that the engine does not directly handle. So, it must be corrected and developed immediately to turn it into an organized structure and make meaning out of this literary information. With this background, language handling is usually widely used by experts and information researchers worldwide. NLP is a subset of informatics, numerical phonetics, AI, and artificial intelligence. NLP helps your chatbot to dissect human language and then generate messages. We have to look at the middle part of Natural Language Processing. Natural Language Processing is used to set up the bot on the different connections and discussions it will go through and provide many examples of the content it will encounter as it will generally give it a much wider range of premises that can also evaluate parse questions more effectively.

The main aim of AI is to improve user-machine interaction. Machine learning is also combined with natural language processing for building an effective chatbot that can understand any input provided by the user and answer the user's query in a timely and efficient manner.

This paper aims to analyze the possibility and challenges of implementing chatbots in the Kumaon language. This paper is organized as follows. Firstly, a detailed survey of the Kumaon language is provided thena discussion on how to map the Kumaon language with other languages is described to make it easy to process it for industrial purposes. In the next section, a detailed discussion is presented on AI-based chatbots, which will mainly focus on how chatbots can be implemented with the help of NLP. In the last section of this paper, we intend to provide future work and conclude our work.

2. Related Literature
2.1 Kumaoni Language

The Kumaon language is a particular form of language peculiar to the Kumaon region people. The people belonging to the Kumaon region of Uttarakhand state are known as Kumaouni's and the language spoken by them is known as the Kumaouni language. Kunayaoni, Kumman, Kumgoni, Kumawani, Kamaoni, Kumau etc are the different synonyms' of Kumaouni. Kumaon division includes six districts of Uttarakhand state i.e., Almora, Nainital, Bageshwar, Pithoragarh, Champawat, Udam Singh Nagar. This region is bounded on the north by Tibet and east by Nepal. Kumaouni language is spoken by the natives and said in some parts of Nepal and Himachal Pradesh. Kumaoni language is one of the 325 recognized languages in India. It is also classified as one of the sub-groups of the Pahari language.

Sharma, L. [1] has mentioned the 13 sub dialects of Kumaoni dialect; these are Danpuriya, Askoti, Sirali, MajhKumaiya, Johari, Soryali, Chaugarkhyali, Kumaiya, Gangola, Khasparjia, Phaldakoti, Pachhai, and Rauchaubhais. Further[2] 7 more sub dialects Central Kumaoni, South eastern Kumaoni, North eastern Kumaoni, Western Kumaoni, Bhabari, Pashchimi, Phaldakotiya. All dialects are effortlessly understandable by each of the Kumaoni speaking provinces.

Renkovskaya[3] has concluded that Kumaoni is an Indo Aryan language whose vocabulary is derived from Sanskrit but not grammar. Also, it is influenced by dialects of Nepali, dialects of the Tibeto-Burmese family, and dialects of Himachal.

UNESCO's Atlas of the world language [4], has put the Kumaoni language in an unsafe category which means it requires consistent conservation efforts. As per the survey done in 1961, there were 1030254 kumaoni speakers in India. Which is increased to 2.2 million in 2011[4].

Researchers [1][3][5][6]has mentioned that being part of the Indo Aryan dialect sphere Kumaoni language shares its vocabulary with Sanskrit [7]and grammar or descriptive linguistics with its alternative indo Aryan languages, particularly Nepali, Hindi, Kashmiri, Gujarati and so on the grammar is additionally shared with other languages of the central Pahadi regions of Garhwali and Jaunsari. The distinction in the grammar between Kumaoni and its other central Pahadi languages exists solely attributable to the influence of Khasas which is currently extinct and was spoken by the first soul of the region. In this case, the verb

is made from the basis ach, which is the same, just in the case of Rajasthani and Kashmiri. Also with the time, new words are added in the Kumaoni vocabulary, which is influenced by various other dialects of the immigrants who came to the Kumaon regions of Uttarakhand.
Kumaoni language has exceptionally rich writing, and it has been very much addressed in the media and, surprisingly, in society music. By noticing the tunes and songs, one might say that Kumaoni music has a dash of nature also. Lokratna Pant or Gumani Pant (1790-1846), Gauri Dutt Pant "Guarda" (1872-1939), Charu Chandra Pandey (1923-2017), Shailesh Matiyani (1931-2001), Mohan Upreti (1925-1997), Brijen Lal Shah (1928-1998) are not many conspicuous authors of Kumaoni language.

3. Methods
3.1 AI Based Chatbots

A chatbot can be characterized as an application programming which used to lead an internet-based talk by means of text-to-discourse or text for the benefit of giving direct contact with a live human agent[8]. It is additionally named chatterbox by Michael Mauldin. Chatbots are intended to reenact the manner in which a human would act as a conversational accomplice, these frameworks regularly require constant tuning and testing, and numerous underway stays are incapable of sufficient banter or passing the business standard Turing test[9].

A chatbot can be described as an application programming which used to lead a web-based visit through text-to-talk or text for outfitting direct contact with a live human agent[8]. It is furthermore named chatterbox by Michael Mauldin. Chatbots are expected to reenact the way wherein a human would go about as a conversational accessory, these structures consistently require relentless tuning and testing, and various in progress stay ill-suited to acceptably talk or pass the business standard Turing test[10]. Technology accepts a monstrous part in the trade and consistently errands. It serves a kind of limits and is applied in absolutely specific approach in different parts of the world. Lately, the general populace has been fantasizing about Artificial Intelligence. Computer programming reenacts the psychological part gifts of somebodys. To be extra careful and immovably connected with individuals, the AI Chatbots are and by supplanting human responses with this item group.

A Chatbot might be a handled program that behaves like a colloquist between the human and the test system or a bot, a remote helper that has become extraordinarily standard as of late in the primary in light of emotional upgrades inside the areas like software engineering, AI and different fundamental advances, for example, neural organizations and normal language process. These chatbots adequately speak with any person with intelligent questions. As of late, there's been an enormous expansion in a few cloud-based visiting test systems that are made out there for the turn of events and improvement of the chatbot sector[11] like IBM Watson, Cleverbot, ELIZA chatbot, and a few others. These everyday specialists have turned into extra responsive, and furthermore, the craft of language among people and robots throughout recent years has improved radically.
Below Table depicts a comparison between two Indo Aryan dialects i.e, Kumaoni and Dogri.

S.No	English	Devanagari	Kumaoni	Kashmiri (Dogri)
1	I	Mein	Mi/ mei/ mai	Auun
2	He	Woh/ Usne	Wou/ ull/ wu	O

3	She	Woh/ Usne	Wu	O
4	You	Tum/ Aap (respect)	Tu/ ter	Tus
5	It	Yeh	Ya	Eh
6	This	Yaeh	Yo	Aay
7	That	Vaeh	Wo	Oh
8	A	EK	A	Ek
9	Yes	Haan	Hoye	Aan
10	Come	Aao/ Aaiye (respect	Aoh	Aa
11	Came	Aaya (he)/ Aayee (she)/ Aaye (plural)	Aai	Aaya
12	Will come	Aayega (he)/ Aayegi (she)/ Aayenge (plural)	Uul	Aauna
13	Open	Kholo/ Kholiye (respect)/ Kholna	Khulul/ kholna	Kholna
14	Opened	Khola (he)/ Kholee (she)/ Khole (plural)	Khol halo	Kholudya
15	Will open	Kholega (he)/ Kholegi (she)/ Kholenge (plural)	Kholul	Kholudge
16	Sit	Baitho/ Baithiye (respect)/ Baithna (to sit)	Baithnou	Baauna
17	Walk	Chalo/ Chaliye (respect)/ Chalna (to walk)	Chalul/ hituul	Chalna
20	Eat	Khao/ Khaiye (respect)/ Khana (to eat)	Khan	Khana
23	Go	Jaao/ Jaaiye (respect)/ Jaana (to go)	Jaan/ janai	Jana
24	Went	Gaye	Jaan/ janai	Uthiiya
25	Will go	Jayenge	Jauol	Jaage
26	Run	Daudna	Bhagun/ daudun	Nassna
29	He ate an apple.	Usnesev khaya.	Ullsaibkheihaaali.	Ossebkhada.
30	I saw the film last week.	Meinepichhlesaftah film dekhi	Mi/ meylpichadihafta picture dekhi.	Me pishlehafte film dikhi.
31	She came by bus yesterday.	Woh kal bus se aayee.	Wu byal bus/ gaadilbatiaai	O kal bus uparaayi.
32	They went to the temple.	Ve mandir gaye.	Wou mandir gyein	O mandregede hey.

33	He slept the whole night.	Woh raatbhar soya.	Wu sariiraatsito.	O puriraat sutta ha.
34	He wrote well in the examination.	Usne exam achcha likha	Ullpariksha me bhallikhoh	Os paper chshaellikhya.
35	He has eaten.	Usnekhaaliya.	Ullkhaiy halo.	Oskhauudya.
36	What is your name?	Tumhara/Aapka naam kyahai?	Teyrnam key chu?	Thohdakenaa ye?
37	What	Kya	Kei/ key	Ke
38	Is	Hai		Ae
39	Your	Tumhara/ Aapka (respect)	Ter/ tumhar	Thohda
40	Name	Naam	Nam	Naa
41	What did you do?	Tumne/ Aapne (respect) kyakiya?	Tumulkeikarouh/ tumul key karrakhoh?	Toskekarde?
42	What should I do?	Mein kyakaroon?	Miki kekarunchann?	Me kekarnachayida?
43	What can I do?	Mein kyakarsaktahoon	Mai key karsakoon?	Aaukekarisakna?
44	What are the questions?	Kya sawalhein?	Prasan/ sawaal key chin?	Sawaalkeaen?
45	What is written in the letter?	Khat meinkyalikhahai?	Chitti me key likhrakho?	Darkhaastuparkelikha ha?
46	What you had been told?	Tumheinkya kaha gayatha?	Tumun key batanrakchii?	Tukkikesunayagya?

4. Result and Discussion

Although Natural language processing can be used to make chatbots in regional language as well, it offers certain limitation also which must be handled to make full potential of NLP in the development of effective chatbots. One of the major motivations to build chatbots in local language is that customers prefer to receive company's service in their local language. In this section, we focus on some limitation that are faced in the implementation of chatbot using NLP. We also provide suggestion to overcome these limitations.

A. Language Understanding

Despite the fact that Chatbots have at least preparation, we can sort out what clients mean for most dialects -one of the fundamental difficulties confronting chatbots today is their capacity to comprehend different local accents or semantic varieties (Dialects), which are novel by their own doing. In spite of the fact that there are a few proposed answers for

this issue, one that is viable is to prepare language understanding models with examples of various phonetic assortments. This arrangement tackles a large portion of the issues, however assuming that the's designer will likely form a chatbot in numerous nation dialects, then, at that point, a language understanding assistance can be utilized to help various dialects. One more significant part of Language Understanding with regards to multilingual chatbots is Entity preparation. Despite the fact that clients will quite often compose things for the most part in their individual dialects, that is not the case constantly. For instance, while certain words are verbally expressed in English, they likewise have their own language variations[12].

B. Language detection

Among other parameters, one of the most significant ones is for a chatbot to have the ability to detect the right word in a regional language as the same word is found in multiple languages, so it essentially becomes more important to understand the context of conversation so that the program can generate appropriate response to its user. Although there are several services in the market to solve this issue of disambiguates, they have their limitations. For example, it becomes very difficult for a translator service to differentiate between words in a different language.

C. Text Parsing

In addition to the above-mentioned issues, the issue of parsing also comes up which needs to be dealt with[12]. Parsing is used to understand the syntactic structure of a text by analyzing its constituent words based on underlying grammar. The main goal of parsing is to come up with a program that can determine the structure of sentences written in Kumaoni language.Majority of existing parsing approaches are based on statistical, probabilistic, and machine learning. In one such work[13] the authors have have taken a set of pre-existing Kumaoni sentences and derived rules of grammar from them[14][15], which have been converted to a mathematical model using Early's algorithm, suitably modified by them.This mathematical model can be used for parsing new Kumaoni language sentences, thus providing researchers a new parsing tool[16][17][12][18].

5. Conclusion

This study aims to analyze the chat bots for Kumaoni dialect and the possibility and implementation challenges as Kumaoni language learning medium. Chat bots provide various advantages for language learners in terms of (1) comfort {trainee feel more comfortable in talking to bots/computer than a teacher}(2) repeating same content with learners endlessly (3) learning methods {provides both text and synthesized speech} (4) creating interest (5)provide quick and effective feedback in both spelling and grammar [14].

Despite research pointing to the strengthsand weaknesses of casual chatbot conversation partners, scant progress towards chatbots as substantive learning partners has been made. Chatbots are a brand new, revolutionary stage for foreign acquisition. They're currently a useful gizmo that continues to grow and develop. This paper proposes that chatbots will be used more effectively at once with relatively small adjustments and therefore the application of Kumaoni language learning tool which will help in saving language for being extinct.

References

[1]   L. Sharma, "Entrepreneurial intentions and perceived barriers to entrepreneurship among youth in Uttarakhand state of India: A cross-cultural investigation across genders," *Int. J. Gend. Entrep.*, 2018.

[2]   V. Ghirardi, "Hariyā Hercules kī Hairānī: Reading the Novel Through the Lens of the

Kumaoni Community," 2020.

[3]   E. Renkovskaya, "New Indo-Aryan associative plural markers derived from Old Indo-Aryan 'other'and their further grammaticalization," *Ling. Posnan.*, vol. 62, no. 2, pp. 65–80, 2020.

[4]   E. Dahl, "Pathways to split ergativity: The rise of ergative alignment in Anatolian and Indo-Aryan," *Diachronica*, vol. 38, no. 3, pp. 413–456, 2021.

[5]   E. Dafouz and U. Smit, *ROAD-MAPPING English medium education in the internationalised university*. Springer, 2020.

[6]   N. Chatterjee, "Colonialism and Knowledge in Grierson's Linguistic Survey of India, by Javed Majeed Nation and Region in Grierson's Linguistic Survey of India, by Javed Majeed." Oxford University Press UK, 2021.

[7]   A. P. Chaves and M. A. Gerosa, "How should my chatbot interact? A survey on social characteristics in human–chatbot interaction design," *Int. J. Human–Computer Interact.*, vol. 37, no. 8, pp. 729–758, 2021.

[8]   L. A. Faza, P. M. Agustini, S. Maesaroh, A. C. Purnomo, and E. A. Nabila, "Motives For Purchase of Skin Care Product Users (Phenomenology Study on Women in DKI Jakarta)," *ADI J. Recent Innov.*, vol. 3, no. 2, pp. 139–152, 2022.

[9]   E. A. Nabila, S. Santoso, Y. Muhtadi, and B. Tjahjono, "Artificial Intelligence Robots And Revolutionizing Society In Terms Of Technology, Innovation, Work And Power," *IAIC Trans. Sustain. Digit. Innov.*, vol. 3, no. 1, pp. 46–52, 2021.

[10]   M. Prawira, H. T. Sukmana, V. Amrizal, and U. Rahardja, "A Prototype of Android-Based Emergency Management Application," *2019 7th Int. Conf. Cyber IT Serv. Manag. CITSM 2019*, 2019, doi: 10.1109/CITSM47753.2019.8965337.

[11]   S. Barve, S. Gaikwad, and D. Nimbane, "Chat-Bot For College Management System," *Arti Nimbane, Dinesh, Chat-Bot Coll. Manag. Syst. (MAY 7, 2021)*, 2021.

[12]   L. K. Fryer, K. Nakao, and A. Thompson, "Chatbot learning partners: Connecting learning experiences, interest and competence," *Comput. Human Behav.*, vol. 93, pp. 279–289, 2019.

[13]   N. Lutfiani, U. Rahardja, and K. T. Khasanah, "The Development Viewboard As an Information Media at Official Site Asosiation," *APTISI Trans. Manag.*, vol. 6, no. 1, pp. 10–18, 2022.

[14]   W. Zhou, "Vocabulary analysis of English major textbooks based on range software," in *Journal of Physics: Conference Series*, 2021, vol. 1856, no. 1, p. 12020.

[15]   J. C. Farah, V. Sharma, S. Ingram, and D. Gillet, "Conveying the Perception of Humor Arising from Ambiguous Grammatical Constructs in Human-Chatbot Interaction," in *Proceedings of the 9th International Conference on Human-Agent Interaction*, 2021, pp. 257–262.

[16]   L. Fryer, D. Coniam, R. Carpenter, and D. Lăpușneanu, "Bots for language learning now: Current and future directions," *Lang. Learn. Technol.*, vol. 24, no. 2, pp. 8–22, 2020.

[17]   D. Kent, "A room with a VUI–Voice user interfaces in the TESOL classroom," *Teach. English with Technol.*, vol. 20, no. 3, pp. 96–123, 2020.

[18]   J. Kuang, S. Ashraf, U. Das, and C. Bicchieri, "Awareness, risk perception, and stress during the COVID-19 pandemic in communities of Tamil Nadu, India," *Int. J. Environ. Res. Public Health*, vol. 17, no. 19, p. 7177, 2020.