Gesturely: a Conversation AI based Indian Sign Language Model

Dr. Nupur Giri
Head of Department, Computer Engineering
Vivekanand Education Society's Institute of Technology
Mumbai, India
nupur.giri@ves.ac.in

Piyush Chugeja
Student, Computer Engineering
Vivekanand Education Society's Institute of Technology
Mumbai, India
d2021.piyush.chugeja@ves.ac.in

Sakshi Kirmathe
Student, Computer Engineering
Vivekanand Education Society's Institute of Technology
Mumbai, India
d2021.sakshi.kirmathe@ves.ac.in

Deven Bhagtani
Student, Computer Engineering
Vivekanand Education Society's Institute of Technology
Mumbai, India
d2021.deven.bhagtani@ves.ac.in

***Abstract*—The project, Gesturely, aims to improve communication within educational settings for individuals with hearing impairments. Sign language, notably Indian Sign Language (ISL) in India, serves as a primary mode of expression for the deaf community. The form of expression among the deaf, relies on a rich vocabulary of gestures involving fingers, hands, arms, eyes, head, and face. The research endeavors to develop an algorithm capable of translating ISL into English, focusing initially on words within the education domain. Through the integration of advanced computer vision and deep learning methodologies, the objective is to create a system capable of interpreting ISL gestures and converting them into written text. The project involves the creation of a comprehensive dataset, with 50 number of words and over 2500 videos. The vision is to empower the deaf community with real-time translation capabilities, promoting inclusivity and accessibility in communication.**

***Index Terms*—Indian Sign Language, computer vision, gesture recognition, sign language dataset**

I. Introduction

Communication is vital for human interaction, yet individuals who are mute or deaf face challenges connecting with the hearing community. In India, Indian Sign Language (ISL) is widely used and recognized as the primary mode of communication. It is used by over 5 million deaf people in India. [1] The development of Natural Language Processing (NLP) systems for sign languages such as American Sign Language (ASL) [2], British Sign Language (BSL) [3], and Deutsche Gebärdensprache (DGS) [4] have benefited from the availability of translation datasets. However, there has been relatively limited focus on ISL due to the scarcity of large annotated datasets. This paper aims to address a new translation dataset focused on ISL, with a particular emphasis on the education domain. Additionally, it introduces a deep learning model for classifying gestures.

ISL presents unique challenges due to its limited resources and reliance on bodily gestures for communication, which adds complexity to training machine learning models. Annotating sign language at the gesture level, rather than the sentence level, poses scalability issues. Prior research has explored translating signs into gloss representations and then converting them into written language (Sign2Gloss2Text) [4]. *Glosses* are textual labels assigned to signed gestures, helping translation systems in working at a more detailed level of sign translation. [5] However, generating gloss representations for entire signed sentences presents additional challenges in data annotation. Overall, in this research paper, following contributions are made:

- A comprehensive ISL to English translation dataset which has more than 50 glosses spread across 2500 and more videos. Making this dataset available for the NLP community can help future research in sign languages.
- A deep learning model for ISL to English translation, inspired by SignAll SDK.[1]

II. Literature review

Unlike spoken languages, sign languages rely on body movements like hand shapes, head nods, eye gazes, and facial expressions to communicate. Translating these continuous movements into written text is quite challenging, opening up new opportunities for research in sign language translation. Many studies have looked into recognizing sign language, using different techniques like gloves, Microsoft Kinect sensors to track hand movements, classify frames based on segmentation masks or utilizing Mediapipe pose estimation pipeline[2].

A. Agarwal and M. K. Thakur [6] use Microsoft Kinect sensors to recognise sign language. They make use of depth images that were captured using the sensor and a gesture is viewed as a sequence of frames. T. Pryor et al. [7] developed

[1] https://developers.googleblog.com/2021/04/signall-sdk-sign-language-interface-using-mediapipe-now-available.html

[2] https://blog.research.google/2020/12/mediapipe-holistic-simultaneous-face.html

SignAloud, which incorporates a pair of gloves equipped with sensors. These gloves track hand position and movement, enabling the conversion of gestures into speech. These hardware solutions are reliable and provide good accuracy but are generally expensive and not portable. The system eliminates the need of external hardware by using any embedded camera.

In [8], Joshi et al. address the lack of resources for ISL in sign language processing. They present a dataset designed for word-level recognition in ISL from video recordings, with over 4700 words covering diverse topics. To overcome ISL's resource limitations, they use a prototype-based one-shot learner, leveraging ASL resources to improve ISL predictions. Ketan Gomase et al. [9] discuss the development of a sign language recognition system using the Mediapipe framework, which leverages machine learning to detect and interpret hand gestures. The framework identifies 21 3D landmarks on the hand from a single frame, making real-time hand and finger tracking possible. Joshi et al. in [10] introduce the ISLTranslate dataset and proposes a baseline model named Pose-SLT for ISL to English translation, leveraging pose estimation models and transformer architecture.

The dataset being proposed draws inspiration from CISLR, as outlined in the reference [8]. While CISLR primarily supports a one-shot learner model, the proposed dataset diverges from this methodology. The aim is to enhance the dataset's utility by including a higher number of videos per gloss, thereby enabling a broader spectrum of research and applications.

III. Novelty of work

The dataset is self curated from videos provided by CISLR. It has 50 distinct glosses spanned across 6 categories: education, adjectives, food, greetings, verbs and miscellaneous words. Notably, each gloss is represented by multiple videos capturing diverse angles and lighting conditions, enriching the training data and enhancing model robustness. By prioritizing diversity and model resilience, the dataset aims to advance sign language recognition, particularly in educational settings, fostering inclusivity and accessibility for individuals with hearing impairments. Training was performed on this dataset with Support Vector Machine, Sequential architecture with LSTM and CNN, 3D CNN, and fine tuned VideoMAE.

IV. Proposed Methodology

The primary aim of this study is to develop a framework for translating ISL glosses into English text, thereby improving accessibility and inclusivity for individuals with hearing impairments. A multi-step approach has been proposed to achieve this objective that involves data curation, preprocessing, feature extraction, model development, and evaluation. By building upon recent advancements in machine learning and computer vision, the methodology aims at overcoming the challenges associated with ISL translation, including the lack of annotated datasets and the complexity of sign language recognition. The methodology laid out in [6] is been followed in this approach which involves carefully examining videos frame by frame.

A. Data Curation

The proposed dataset is created from the videos provided by CISLR [8], which has 57 distinct categories. Within the dataset, emphasis has been put on the education domain amongst 6 categories where education has 19 glosses out of 50. All videos maintain a consistent format, adhering to a 1:1 aspect ratio, recorded with a 720p resolution at 30 frames per second. This standardized recording setup ensures uniformity and quality across the dataset. Following are the categories and distributions:

- Education: This category has glosses like *college, blackboard, examination, laboratory*, etc. There are a total of 19 glosses in education category.
- Adjectives: Certain adjectives like *cold, angry, clever, bad*, etc. are included in the adjectives category with a total of 7 adjectives.
- Greetings: Generally used greetings such as *good morning, good afternoon, good night, namaste*, etc. are included in this category. Total glosses in this category are 9.
- Verbs: Simple action words like *call, grow and die* are included in this category.
- Food: Food and food related items like *apple, bread, eat and bowl* are included.
- Miscellaneous: This category has words symbolising objects, nouns, etc.

B. Pre-processing and Feature extraction

Each video undergoes frame-by-frame processing, where pose coordinates are extracted from each frame as shown in Fig. 1. Mediapipe provides 33 coordinates detailing the human body from head to toe. However, as depicted in Fig. 2, body parts below the waist remain unseen, making coordinates below the waistline unusable. This discrepancy in data could introduce inconsistencies. To rectify this issue, all unused coordinates per frame, specifically coordinates 25 to 32 (8 coordinates), are filtered out before feeding the data to the model.

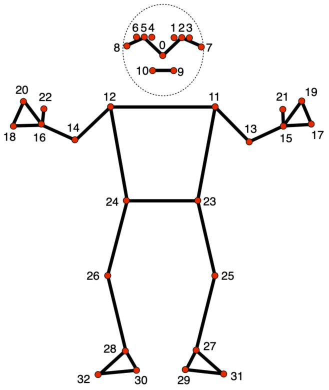


Fig. 1. Mediapipe Pose Landmarks [13]

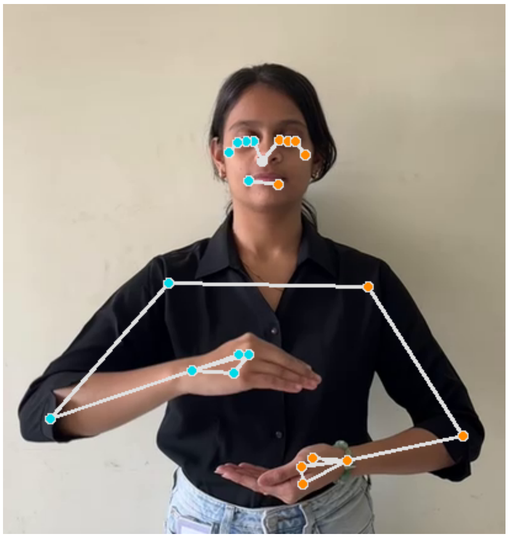


Fig. 2. Extracting coordinates from a frame of video for gloss *'grow'*

C. Model Development

After the coordinates are pre-processed and filtered, classification models will be trained. Following the approach outlined in [6], the focus is on training models capable of accurately distinguishing ISL glosses from the extracted pose data. Convolutional Neural Networks (CNN) are being incorporated, and Long Short-Term Memory (LSTM) networks to capture the complex patterns present in sign language gestures. These models are trained using optimization algorithms like Adam, with carefully selected hyperparameters to ensure effective learning. Throughout the model development process, thorough experimentation and validation were conducted to ensure the reliability and applicability of the approach.

D. Evaluation

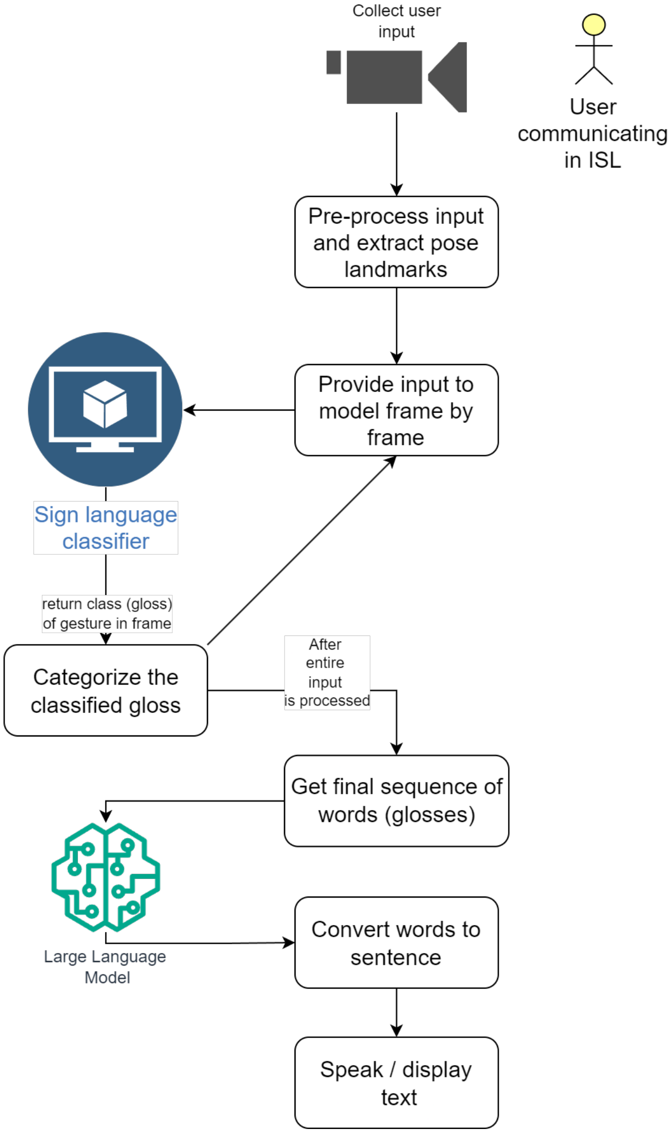
In evaluation, the performance of the developed model has been assessed based on various parameters, including accuracy, precision, recall and test loss. The evaluation is done on how well the model works when the data isn't seen or is retrieved real-time. Tests were conducted to determine if the model is unbiased and doesn't favor a particular class.

V. Proposed System

The system aims to address the challenge of translating ISL. It works by collecting the user input, processing the frames and classifying them into respective glosses. Once the video is processed and glosses are identified, the sentence is constructed. The proposed system can be seen in Fig. 3, based on the methodology followed in [11] where they leverage *gpt-3.5-turbo* for sentence generation from glosses.

A. Model selection and evaluation

After evaluating various model architectures and their applications, the model type outlined in [12] was adopted, which leverages CNN and LSTM. Table I outlines the various classification techniques used, along with their specifications



Collect user input
User communicating in ISL
Pre-process input and extract pose landmarks
Provide input to model frame by frame
Sign language classifier
return class (gloss) of gesture in frame
Categorize the classified gloss
After entire input is processed
Get final sequence of words (glosses)
Large Language Model
Convert words to sentence
Speak / display text

Fig. 3. Block diagram of proposed system

and corresponding test metrics. The VideoMAE model and 3D CNN model were trained on video data whereas the SVM model and LSTM model were trained on coordinate data collected using Mediapipe Pose. The models trained on video data couldn't efficiently capture the complexity of ISL and spatial data as compared to the models trained on mathematical coordinate data. The Conv1D model was selected as the final model as this model achieved the highest accuracy and the lowest test loss score among all the models considered. For the model, labelled as Sign Language Classifier in Fig. 3, a sequential architecture is being employed, comprising Conv1D, LSTM, Flatten, and Dense layers. This architecture processes each frame's 25 coordinates provided by Mediapipe, enabling the model to classify each video into specified glosses. The model learns patterns from the sequential data to accurately assign gloss labels to the videos during training.

TABLE I
COMPARISON OF DIFFERENT CLASSIFICATION MODELS
A - ACCURACY, P - PRECISION, R - RECALL, L - TEST LOSS

Model name	***Comparison metrics***												
	Specifications	***train = 0.8, test = 0.2***				***train = 0.75, test = 0.25***				***train = 0.7, test = 0.3***			
		A	**P**	**R**	**L**	**A**	**P**	**R**	**L**	**A**	**P**	**R**	**L**
Fine Tuned VideoMAE	learning rate = 0.001, sample rate=4, image resolution=224x224	0.79	0.75	0.74	0.49	0.78	0.74	0.74	0.52	0.75	0.74	0.72	0.52
Support Vector Machine (SVM)	Regularization parameter (C) = 1.0, kernel = 'linear'	0.84	0.84	0.83	-	0.82	0.81	0.81	-	0.80	0.80	0.79	-
Neural network model using 3D CNN	4 hidden NN layers, optimizer='adam', activation='softmax', loss='categorical _crossentropy', epochs=10, validation_split=0.2, batch_size=32	0.88	0.75	0.74	1.08	0.85	0.73	0.72	1.85	0.80	0.72	0.72	2.01
Neural network model using CNN & LSTM	4 hidden NN layers, optimizer='adam', activation='ReLU', loss='sparse _categorical _crossentropy', epochs=10, validation_split=0.2, batch_size=32	0.88	0.88	0.87	0.34	0.89	0.89	0.88	0.31	0.86	0.87	0.85	0.39

B. *Gesture Classification*

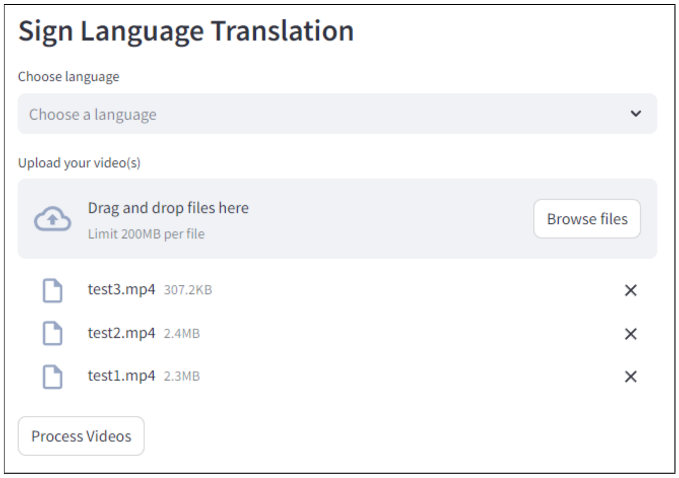
User input is obtained either through computer vision or recorded video, wherein the user performs a series of ISL gestures to communicate. The collected input undergoes pre-processing, wherein each frame is analyzed, and the extracted landmarks are forwarded to the classifier model, depicted in Fig. 3. Landmarks generated per frame, acquired using Mediapipe, are refined to include only upper body landmarks (landmarks 0 to 24, as illustrated in Fig. 1). These refined landmarks are then sequentially sent to the model, which classifies the data into glosses.

C. *Sentence Construction*

Following the identification of a sequence of glosses, the list of words is transmitted to a Large Language Model (LLM). Shahin and Ismail [15] explored ChatGPT's ability to convert glosses into sign language. Similarly, Google's Gemini[3] is being incorporated to build sentences. The LLM generates a grammatically correct and contextually appropriate sentence that conveys the user's message. Following is the prompt given to Gemini: *You are a sign language translator. You are being given a set of words collected from people using sign language, the words are what sentence they want to speak. Generate a grammatically correct and meaningful sentence using those words. Do not add context to the sentence. Return only the sentence and nothing more.*

D. *Outputs of system*

Based on the methodology and flow of system explained in the preceding sections, the functioning of the system is as follows: A basic user interface (UI) shown in Fig. 4 has been developed for users to upload the videos they wish to translate into sentences. As shown in Fig. 4, three videos are uploaded: these correspond to glosses *please, call* and *evening*.



Sign Language Translation

Choose language

Choose a language

Upload your video(s)

Drag and drop files here
Limit 200MB per file

Browse files

test3.mp4 307.2KB ✕
test2.mp4 2.4MB ✕
test1.mp4 2.3MB ✕

Process Videos

Fig. 4. Uploading videos for translation

These videos are then processed using Mediapipe Pose and their coordinates are extracted, as seen in Fig. 5. Once the videos are processed, coordinates are sent to the model and it classifies them into *glosses*. This collection of glosses is

[3] https://deepmind.google/technologies/gemini/

sent to Gemini for sentence construction. Once the sentence is ready, it is translated into the language chosen by the user. The system supports four languages which are Hindi, Marathi, Gujarati and English. Fig. 6 displays the outputs of the system in corresponding languages. Once the sentence is ready, it is converted to speech using Google Text to Speech (gTTS) engine.

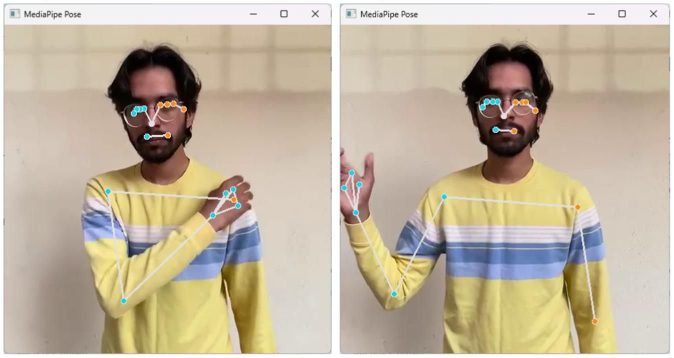
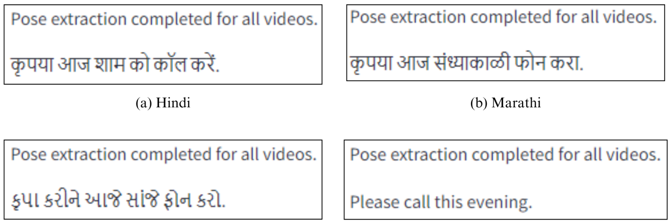


Fig. 5. Processing uploaded videos



Pose extraction completed for all videos.

कृपया आज शाम को कॉल करें.

(a) Hindi

Pose extraction completed for all videos.

कृपया आज संध्याकाळी फोन करा.

(b) Marathi

Pose extraction completed for all videos.

કૃપા કરીને આજે સાંજે ફોન કરો.

(c) Gujarati

Pose extraction completed for all videos.

Please call this evening.

(b) English

Fig. 6. Outputs of the system in various languages

VI. Conclusion

A novel approach to Indian Sign Language (ISL) translation has been presented using advanced machine learning techniques. By using Mediapipe for pose estimation and a sequential model architecture comprising Conv1D and LSTM layers, a system capable of classifying ISL gestures into specified glosses has been developed with an 88% accuracy. The approach, inspired by previous studies, shows promise in connecting the deaf and hearing communities better. Looking ahead, the work paves the way for more improvements in ISL translation tech, which could make life easier and more inclusive for people with hearing challenges.

References

[1] Advaith Sridhar, Rohith Gandhi Ganesan, Pratyush Kumar, and Mitesh Khapra. 2020. INCLUDE: A Large Scale Dataset for Indian Sign Language Recognition. In Proceedings of the 28th ACM International Conference on Multimedia (MM '20). Association for Computing Machinery, New York, NY, USA, 1366-1375. https://doi.org/10.1145/3394171.3413528

[2] Li, Dongxu & Rodríguez, Cristian & Yu, Xin & Li, Hongdong. (2020). Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison. 1448-1458. 10.1109/WACV45572.2020.9093512.

[3] Samuel Albanie, GülVarol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, and Andrew Zisserman. 2021. BOBSL: BBC-Oxford British Sign Language Dataset.

[4] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In 2018 IEEE/CVFConference on Computer Vision and Pattern Recognition, pages 7784–7793.

[5] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden, 'Sign Language Transformers: Joint End-to-end Sign Language Recognition and Translation', arXiv [cs.CV]. 2020.

[6] A. Agarwal and M.K. Thakur, "Sign Language Recognition using Microsoft Kinect," Sixth International Conference on Contemporary Computing (IC3), September 2013.

[7] MailOnline, "SignAloud gloves translate sign language gestures into spoken English," 2016. [Online]. Available: http://www.dailymail.co.uk/sciencetech/article-3557362/SignAloudgloves-translate-sign-language-movements-spoken-English.html. [Accessed: 06-03-2024].

[8] Abhinav Joshi, Ashwani Bhat, Pradeep S, Priya Gole, Shashwat Gupta, Shreyansh Agarwal, and Ashutosh Modi. 2022. CISLR: Corpus for Indian Sign Language Recognition. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 10357–10366, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

[9] Ketan Gomase, Akshata Dhanawade, Prasad Gurav, Sandesh Lokare, and Jyoti Dange. Hand Gesture Identification using Mediapipe. In March 2022 International Research Journal of Engineering and Technology, pages 1199-1202.

[10] Abhinav Joshi, Susmit Agrawal, Ashutosh Modi. 'ISLTranslate: Dataset for Translating Indian Sign Language'. arXiv [Cs.CL], 2023, http://arxiv.org/abs/2307.05440. arXiv.

[11] O. M. Sincan, N. C. Camgoz, and R. Bowden, 'Using an LLM to Turn Sign Spottings into Spoken Language Sentences', arXiv [cs.CV]. 2024. https://arxiv.org/pdf/2403.10434.pdf

[12] K. B. Tran, U. D. Nguyen and Q. T. Huynh, "Continuous Sign Language Recognition Using MediaPipe," 2023 International Conference on Advanced Technologies for Communications (ATC), Da Nang, Vietnam, 2023, pp. 493-498, doi: 10.1109/ATC58710.2023.10318855.

[13] Pose landmark detection guide — MediaPipe — Google for Developers [Online]. Available: https://developers.google.com/mediapipe/solutions/vision/pose_landmarker [Accessed: 07-03-2024]

[14] Z. Tong, Y. Song, J. Wang, and L. Wang, 'VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training', arXiv [cs.CV]. 2022.

[15] N. Shahin and L. Ismail, 'ChatGPT, Let us Chat Sign Language: Experiments, Architectural Elements, Challenges and Research Directions', arXiv [cs.CL]. 2024.