Research Proposal on Sign Language Processing

Ayush Zenith[1] and Samuel Ji[2]

***Abstract*— We propose to research about the process of sign languages translation based on big data training.**

I. Justification

For those who are hard of hearing (or deaf community in general), sign language is the primary means of communication. With sign language having all the characteristics of a natural language concepts of natural language processing can be applied yet there doesn't seem to be any major effort from academics to model and study sign languages. With multiple different forms of sign languages around the world we want to study and model a couple of them in order to shorten the divide between multiple different languages and allow for translation without the need of human interpreters. By showing that sign languages can also be processed and translated as a natural language we hope to shorten the bridge between sign languages and normal languages and promote research in the different communities of Sign Language and it's relation with technology.

II. Related work

The proposed topic is still uncovered by sufficient research, so we decided to reference the work that is related to the sign-language recognition, as well as the sign-language spoken-language translation, combining these topic organically and conduct a research.

A. SubUNets: End-to-End Hand Shape and Continuous Sign Language Recognition[1]

This article provides an approach to sequence to sequence learning that outperforms previous techniques that recognizes hand-shapes. It also eliminates the necessity of alignment step to segment out the signs for recognition while obtaining comparable recognition rates. We are aiming to use the model to recognize the input gestures and take them for the rest calculation and training.

B. Sign Language Transformers: Joint End-to-end Sign Language Recognition and Translation[2]

This article provides an insight of applying mid-level sign gloss in the translation process. We are aiming to refer the Connectionist Temporal Classification (CTC) loss mentioned in their research to form a unified model architecture. We can also leverage the evaluation they used to measure their model's performance to access ours.

[1]Ayush Zenith is an undergraduate student with the Khoury school of Computer Science, Northeastern University, Boston, MA zenith.a@northeastern.edu

[2]Samuel Ji is an undergraduate student with the Khoury college of Computer Science, Northeastern University, Boston, MA ji.xian@northeastern.org

III. Method

We are trying to translate one sign language used in a country/region to another sign language used in another country/region. The translation process for one particular sign language involves at least 3 layers - sign, gloss, and speech. Given the limited time, we may first only focus on understanding the two way translation process for one sign language, and potentially extend from there to build a bilingual sign language translation system.

The first layer of translation can be modelled by using an Recurrent Neural Network, where its inputs are the equivalent open pose for video clips that presents a sign language sequence, and the outputs are the corresponding gloss representation. From there we take the generated gloss as the input for another RNN and output the speech text. The reverse works similar except we need simulate hand gestures according to the generated open pose. For multilingual translation, we can use gloss as the medium and build model that can interpret as well as translate gloss from one sign language to another as it contains more syntax information for sign language.

IV. Data

For our data set we will be using How2Sign Amanda et. al., 2021[3]. With 16,000 signs spread over 35,000 sentences and over 79 hours of videos, How2Sign is one of the largest and most customize data sets available on American Sign language. How2Sign also conveniently divides it's data for us to train, validate, and test/evaluate. The data also includes RGB videos with English and Gloss subtitles along with coordinates and OpenPose files to help us make pose estimations for the simulated robot. All this can help us convert from English to Gloss, Gloss to ASL/Pose and vice versa.

V. Evaluation

We will be using BLEU as the metric in order to validate the performance of the model. We are aiming for a 0.55-0.60 on the BLEU metric as the probability of two sentences being phrased the same way is pretty low thus a score from 0.4-0.50 is often a really seen as a very high quality transformation.

x	Expected	Real
Random	≤ 0.1	
My Method	$0.55 \geq$	

REFERENCES

[1] N. C. Camgoz, S. Hadfield, O. Koller and R. Bowden, "SubUNets: End-to-End Hand Shape and Continuous Sign Language Recognition," 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3075-3084, doi: 10.1109/ICCV.2017.332.

[2] Camgöz, N.C., Koller, O., Hadfield, S., Bowden, R. (2020). Sign Language Transformers: Joint End-to-End Sign Language Recognition and Translation. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 10020-10030.

[3] A. Duarte, S. Palaskar, L. Ventura, D. Ghadiyaram, K. DeHaan, F. Metze, J. Torres, and X. Giro-i-Nieto, "How2Sign: A large-scale multimodal dataset for continuous American sign language," arXiv.org, 01-Apr-2021. [Online]. Available: https://arxiv.org/abs/2008.08143. [Accessed: 13-Oct-2022].