Towards Speech to Sign Language Translation

Amanda Duarte[1,2], Gorkem Camli[1], Jordi Torres[1,2], and Xavier Giro-i-Nieto[1]

[1] Universitat Politecnica de Catalunya - UPC, Spain
[2] Barcelona Supercomputing Center - BSC, Spain
{amanda.duarte, xavier.giro}@upc.edu

Abstract

Sign Language (SL) is the primary means of communication for a majority of the hearing-impaired community. Current computational approaches in this research area have focused specifically on Sign Language Recognition[1] (SLR) and Sign Language Translation (from SL to text) [2] (SLT). However, the reverse problem of translating from spoken language to sign language has so far been unexplored. The goal of our ongoing project is to provide to people with hearing disabilities the audio tracks from online videos, by automatically generating a video-based speech to sign language translation. In this paper, we will point out the shortcomings that limit the advances of this research area and propose first steps towards this end.

Sign Language Recognition approaches have mostly focused on automatic recognition[1] and classification [3] of signs. SLR methods have used hand-crafted intermediate representations [4, 5] and the temporal changes in these features have been designed using classical graph based approaches, such as Hidden Markov Models [6], Conditional Random Fields [7] or template-based methods [8, 9]. Nowadays, with the advance of deep learning, researchers have adopted Convolutional Neural Networks for manual [10, 11] or non-manual [12] feature representation, and Recurrent Neural Networks for temporal modeling [13]. However, most of these systems treat the problem as a simple recognition task ignoring the rich grammatical and linguistic structures of sign language that differs from spoken language[14]. It is also interesting to note that sign languages have their own linguistic rules [15], and spoken languages are not translated into sign languages on a word-by-word basis.

Previous studies [16] have proposed to construct sentences by recognizing an isolated set of signs without taking into account the special linguistic structure of sign language. In contrast, [2] addressed the problem of SLT in the framework of Neural Machine Translation and formulated the task as a sequence-to-sequence problem, resulting in the first end-to-end system to translate sign language into text. However, both these approaches use sign-language as their inputs and generate natural language text as output. Here, we propose to use a similar translation-based end-to-end model but for speech-to-sign language translation. To the best of our knowledge this venue is not yet explored but it would be the necessary methodology to enable a real-time speech to sign language translation.

Nonetheless, in order to build an end-to-end speech-to-signs system, it is necessary to be able to combine components for speech recognition, machine translation and sign language/video synthesis. Furthermore, a large dataset that

includes the speech signal and its respective interpretation in sign language is needed. Although there are Sign Language datasets available [17–22, 22–24] they are usually weakly annotated and lack the human pose information which legacy sign language methods heavily rely on. Moreover, most of them are recorded in controlled environments with limited vocabulary which inhibits the end goal of SLT.

In Table 1 we list some of the most common sign language datasets along with their language ids, segmentation level, public availability and principal content provided. All datasets presented here contain videos including single or multiple signers interpreting general or specific subjects. These content are segmented by *sentence-level*, *word-level* or just *letter*. The content fall into sentence-level category when the continuous sings interpretation is provided, in the word-level when the dataset provides just isolated signs or just letters(or finger spelling) when it contains just letters, numbers or specific signs. We show in the table that some datasets also provide gloss-level sign-by-sign written information along with notations to account the facial and body grammar that goes with the signs. Some of them also provide the correspondent text translation of the respective spoken sign language. We would like to bring the readers attention to the fact that none of these datasets have a speech component in them. This is one of the major challenges and one reason for lack of prior work in this area.

Table 1. Sign Language standard datasets: DGS and ASL stands for *German Sign Language* and *American Sign Language* respectively. *Trans* designates the translation/transcription of the content into the respective language.

Dataset Name	Language ID	Segmentation	Public?	Content		
				Video	Gloss	Trans
RWTH-Phoenix-2014 [17]	DGS	Sentence	✓	✓	✓	
RWTH-Phoenix-2014T [2]	DGS	Sentence	✓	✓	✓	✓
RWTH Fingerspelling [19]	DGS	Letter		✓	N/A	N/A
DGS Kinect 40 [16]	DGS	Word		✓		✓
ASL-LEX [25]	ASL	Word	✓	✓	✓	✓
ASLLVD [22]	ASL	Word	✓	✓	✓	✓
RWTH-Boston-104 [20]	ASL	Sentence	✓	✓		✓
RVL-SLLL [21]	ASL	Word		✓		✓
Dicta-Sign [24]	Multilingual	Word	✓	✓	✓	✓
ATIS Corpus [23]	Multilingual	Sentence		✓	✓	✓

To address this problem, we are currently collecting a video dataset of ASL containing its corresponding speech translation and annotation. However, this type of translation is usually done by SL interpreters. Having such experts for data collection/annotation is a difficult and also a very expensive task.

There are certain TV broadcasters, government organizations, public and private events where every broadcast or talk is also translated into sign language by experts. This expert translation is also often recorded and stored as videos,

but is rarely publicly available. This data would be useful to the community if made publicly available or under license for research purposes. We welcome any and all collaborations and leads that would help us in our efforts towards this data collection or procurement, and finally towards the proposed project goals.

Acknowledgements

We thank Ozan Caglayan, Shruti Palaskar and Roma Patel for the valuable feedback on this writeup. This work has been supported by a Facebook Caffe2 Research Award 2017 and the project TEC2016-75976-R funded by the Spanish Ministerio de Economa y Competitividad and the European Regional Development Fund (ERDF).

References

1. Cooper, H., Holt, B., Bowden, R.: Sign language recognition. In: Visual Analysis of Humans. Springer (2011) 539–562
2. Cihan Camgoz, N., Hadfield, S., Koller, O., Ney, H., Bowden, R.: Neural sign language translation. In: Conference on Computer Vision and Pattern Recognition. (2018) 7784–7793
3. Joudaki, S., Mohamad, D.b., Saba, T., Rehman, A., Al-Rodhaan, M., Al-Dhelaan, A.: Vision-based sign language classification: a directional review. IETE Technical Review **31**(5) (2014) 383–391
4. Cooper, H., Ong, E.J., Pugeault, N., Bowden, R.: Sign language recognition using sub-units. Journal of Machine Learning Research **13**(Jul) (2012) 2205–2231
5. Koller, O., Forster, J., Ney, H.: Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. Computer Vision and Image Understanding **141** (2015) 108–125
6. Vogler, C., Metaxas, D.: Parallel hidden markov models for american sign language recognition. In: 7th IEEE International Conference on Computer Vision. Volume 1., IEEE (1999) 116–122
7. Yang, H.D., Sclaroff, S., Lee, S.W.: Sign language spotting with a threshold model based on conditional random fields. IEEE Transactions on Pattern Analysis and Machine Intelligence **31**(7) (2009) 1264–1277
8. Ong, E.J., Cooper, H., Pugeault, N., Bowden, R.: Sign language recognition using sequential pattern trees. In: Conference on Computer Vision and Pattern Recognition (CVPR), IEEE (2012) 2200–2207
9. Buehler, P., Zisserman, A., Everingham, M.: Learning sign language by watching tv (using weakly aligned subtitles). In: IEEE Conference on Computer Vision and Pattern Recognition., IEEE (2009) 2961–2968
10. Koller, O., Ney, H., Bowden, R.: Deep hand: How to train a cnn on 1 million hand images when your data is continuous and weakly labelled. In: Conference on Computer Vision and Pattern Recognition. (2016) 3793–3802
11. Koller, O., Zargaran, O., Ney, H., Bowden, R.: Deep sign: hybrid cnn-hmm for continuous sign language recognition. In: British Machine Vision Conference 2016. (2016)
12. Koller, O., Ney, H., Bowden, R.: Deep learning of mouth shapes for sign language. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. (2015) 85–91
13. Cui, R., Liu, H., Zhang, C.: Recurrent convolutional neural networks for continuous sign language recognition by staged optimization. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017)

14. Emmorey, K.: Perspectives on classifier constructions in sign languages. Psychology Press (2003)
15. Stokoe, W.C.: Sign language structure. Annual Review of Anthropology **9**(1) (1980) 365–390
16. Ong, E.J., Cooper, H., Pugeault, N., Bowden, R.: Sign language recognition using sequential pattern trees. In: Conference on Computer Vision and Pattern Recognition (CVPR), Providence, Rhode Island, USA (June 16 – 21 2012)
17. Forster, J., Schmidt, C., Hoyoux, T., Koller, O., Zelle, U., Piater, J.H., Ney, H.: Rwth-phoenix-weather: A large vocabulary sign language recognition and translation corpus. In: LREC. (2012) 3785–3789
18. Ong, E.J., Cooper, H., Pugeault, N., Bowden, R.: Sign language recognition using sequential pattern trees. In: Computer Vision and Pattern Recognition (CVPR), IEEE (2012) 2200–2207
19. Dreuw, P., Deselaers, T., Keysers, D., Ney, H.: Modeling image variability in appearance-based gesture recognition. In: ECCV Workshop on Statistical Methods in Multi-Image and Video Processing. (2006) 7–18
20. Dreuw, P., Forster, J., Deselaers, T., Ney, H.: Efficient approximations to model-based joint tracking and recognition of continuous sign language. In: International Conference on Automatic Face & Gesture Recognition, 2008., IEEE (2008) 1–6
21. Martínez, A.M., Wilbur, R.B., Shay, R., Kak, A.C.: Purdue rvl-slll asl database for automatic recognition of american sign language. In: 4th IEEE International Conference on Multimodal Interfaces, 2002., IEEE (2002) 167–172
22. Athitsos, V., Neidle, C., Sclaroff, S., Nash, J., Stefan, A., Yuan, Q., Thangali, A.: The american sign language lexicon video dataset. In: CVPRW'08., IEEE (2008) 1–8
23. Bungeroth, J., Stein, D., Dreuw, P., Ney, H., Morrissey, S., Way, A., van Zijl, L.: The atis sign language corpus. 6th International Conference on Language Resources and Evaluation.
24. Matthes, S., Hanke, T., Regen, A., Storz, J., Worseck, S., Efthimiou, E., Dimou, A.L., Braffort, A., Glauert, J., Safar, E.: Dicta-sign–building a multilingual sign language corpus. In: 5th LREC, Istanbul (2012)
25. Caselli, N., Sevcikova, Z., Cohen-Goldberg, A., Emmorey, K.: Asl-lex: A lexical database for asl. Behavior Research Methods (2016)