A Data-Driven Representation for Sign Language Production

Harry Walsh; Abolfazl Ravanshad; Mariam Rahmani; Richard Bowden

Phonetic representations are used when recording spoken languages, but no equivalent exists for recording signed languages. As a result, linguists have proposed several annotation systems that operate on the gloss or sub-unit level; however, these resources are notably irregular and scarce. Sign Language Production (SLP) aims to automatically translate spoken language sentences into continuous sequences of sign language. However, current state-of-the-art approaches rely on scarce linguistic resources to work. This has limited progress in the field. This paper introduces an innovative solution by transforming the continuous pose generation problem into a discrete sequence generation problem. Thus, overcoming the need for costly annotation. Although, if available, we leverage the additional information to enhance our approach. By applying Vector Quantisation (VQ) to sign language data, we first learn a codebook of short motions that can be combined to create a natural sequence of sign. Where each token in the codebook can be thought of as the lexicon of our representation. Then using a transformer we perform a translation from spoken language text to a sequence of codebook tokens. Each token can be directly mapped to a sequence of poses allowing the translation to be performed by a single network. Furthermore, we present a sign stitching method to effectively join tokens together. We evaluate on the RWTH-PHOENIX-Weather-2014T (PHOENIX14T) and the more challenging meineDGST (mDGS) datasets. An extensive evaluation shows our approach outperforms previous methods, increasing the BLED-1 back translation score by up to 72%.