ENGLISH TO ASL TRANSLATOR FOR SPEECH2SIGNS

Daniel Moreno Manzano

daniel.moreno.manzano@alu-etsetb.upc.edu

ABSTRACT

This paper illustrates the work around the English - American Signs Language (ASL) data generation for the speech2signs system that is devoted to the generation of a signs language interpreter. The current work will be, first, an approximation to the speech2signs system and, second, a video-to-video corpus generator for an end-to-end approximation of speech2signs.

In order to generate the desired corpus data, the Google Transformer [1] (a Neural Machine Translation system based completely on attention) will be trained to translate from English to ASL. The dataset used to train the Transformer is the ASLG-PC12 [2].

Index terms: American sign language, speech2signs, translation, Transformer, ASLG-PC12

1. INTRODUCTION

According to the World Health Organization, hearing impairment is more common than we think, affecting more than 253 million people worldwide [3]. Although recent advancements like the Internet, smartphones and social networks have enabled people to instantly communicate and share knowledge at a global scale, deaf people still have very limited access to large parts of the digital world.

For most of deaf individuals, watching online videos is a challenging task. While some streaming and broadcast services provide accessibility options such as captions or subtitles, but these are available for just a part of the catalog and often in a limited amount of languages. However, accessibility is not guaranteed for every commercial video.

Over the last years, Machine Learning and Deep Learning have had increasingly advances and so it is also with the Machine Learning Tasks. After years of Statistical Machine Translation predominance, the Neural Machine Translation began having more prominence with the good results of the Recurrent Neural Networks (RNN) with some Attention mechanism but they are hard to train, a lot of time and computational effort. Lately, the Google implementation of the Transformer [1] is state of the art in this field and it is just based in Attention, no RNN what means that is fast and does not require much computations. Nowadays, very impressive progresses are taking place in the Multimodal Machine Translation field that takes advantage of different ways to represent the same concept in order to learn about it and its translation.

Surprisingly, in these advances from the Machine Learning field, the ones with respect to the deaf community problems have focused more effort in us understanding their sign language than the other way [4, 5, 6, 7]. On the contrary, speech2signs aims to bring the Machine Learning and Deep Learning advances to the deaf community watching videos difficulties.

1.1. speech2signs

The speech2signs project is a video-to-video translation system that given a video of some person talking, the system will generate a puppet interpreter video to translate the speech signal into American Sign Language.

Fig. 1. An example of the ideal result of the speech2signs project

The final system is planned to be an end-to-end Neural Network that process the data itself. Despite of the absence of a proper database to train that NN, the first step of the project is to generate data. In order to do that, the system has been split in three different blocks.

1. An Automatic Speech Recognition (ASR) block that extracts the audio from the video and transcribes it to text.
2. A Neural Machine Translation (NMT) module that this paper concerns, translating from english to American Sign Language.

3. A Video Generator that creates the puppet interpreter avatar[1] [8, 9].

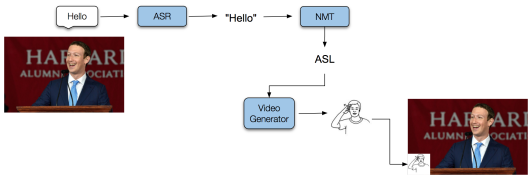

Fig. 2. The speech2signs blocks architecture

1.2. Sign language and sign language annotation

The sign language vocabulary amount and grammar is not exactly the same as in its origin language. For example, a sentence is not exactly equally constructed as it can be seen in Fig. 3. The verbs conjugation has no sense and the subject pronouns are different depending on its meaning in each context.

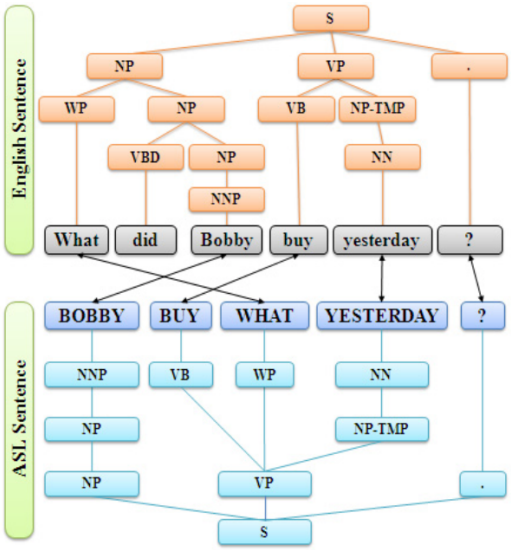

Fig. 3. Sign language grammatical structure example [10]

There are as much sign languages as the spoken ones, as soon as each spoken language has its own sign version. Depending on the country it may variate, even. For example, the ASL is quite diverse than the Britain one (BSL). There also exist an International Sign Language, but there is not much people that uses it. This is a very big problem for developing a solution for the whole deaf community.

Moreover, in order to describe or write a sign to be simply understood by a computer there are different annotation approaches (Stokoe notation, Hamburg notation System (HamNoSys), Prosodic Model Handshape Coding (PMHC), Sign Language Phonetic Annotation (SLPA)) giving more or less information about the gesture, fingers, ... of the sign [11].

The absence of a global standard in sign language makes very difficult to create systems or develop a corpus that could solve the proposed task. In this work the ASL is chosen despite of the amount of people that can understand it and because it has a richer state of the art than others.

[1] http://asl.cs.depaul.edu/

2. RELATED WORK

As explained before, the research community working on the sign language context is mainly focused on the fields of Sign Language Recognition.

Few works are devoted to the relationship and translation of spoken language to the sign one [12, 13, 14, 15, 16] and they are very old and based on Statistical Machine Translation. On the other hand, this paper describes the commitment of giving a NMT state of the art for english to sign language translation.

3. ARCHITECTURE

In NMT the most used model is the Encoder-Decoder one...

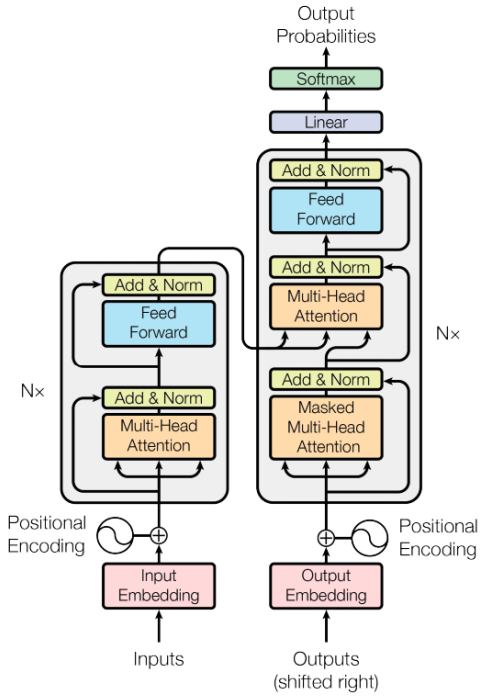

Fig. 4. The Transformer - model architecture [1].

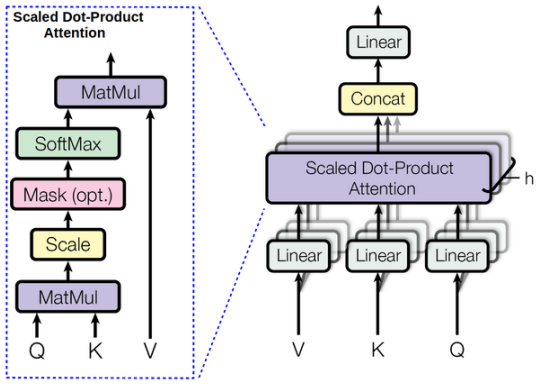

Fig. 5. (left) Scaled Dot-Product Attention. (right) Multi-Head Attention consists of several attention layers running in parallel [1].

4. TRAINING

In this section...

4.1. Dataset

The main problem of this project is the data retrieval. There is not a proper dataset for sign language translation and very difficult to find. Moreover, the existing ones are very small and force researches to resign themselves with a narrow domain for training [16].

The database used is the ASLG-PC12[2] [2, 10]. It is not annotated in any sign language notation by convention. They decide that the meaning of a sign is the written correspondence to the talking language to avoid complexity [10].

As it can be seen in Table 1, the ASLG-PC12 corpus ...

Table 1. English - ASL Corpus Analysis

Characteristics	Corpus's English set	Corpus's ASL set
# sentences	87710	87710
Max. sentence size	59 (words)	54 (words)
Min. sentence size	1 (words)	1 (words)
Average sent. size	13.12 (words)	11.74 (words)
# running words	1151110	1029993
Vocabulary size	22071	16120
# singletons	8965 (39.40%)	6237 (38.69%)
# doubletons	2855 (12.94%)	1978 (12.27%)
# tripletons	1514 (6.86%)	1088 (6.75%)
# othertons	9007 (40.81%)	6817 (42.29%)

By convention, the dataset was randomly split in a development and test sets of $\sim$ 2000 sentences each (Table 2).

[2] http://achrafothman.net/site/asl-smt/

Table 2. Database split for training

Train set length	Development set length	Test set length
83618 sentences (95.4%)	2045 sentences (2.3%)	2046 sentences (2.3%)

4.2. Preprocessing

In order to preprocess the raw data and tokenize it, the Moses tools [17] have been used. As it will be seen in the Table 3, a tokenization problem of ASL special words as the pronouns will appear. They will not be properly tokenized despite of the ASL is not a language discerned by the Moses Project and, thus, has not the correct tokenizer rules.

4.3. Parameters and implementation details

The Transformer implementation[3] used was programmed in Pytorch [18, 19]. The used optimizer for the training is the Adam optimizer [20] with $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 10^{-9}$ Following [1], it has been configured with:

- $batch_{size} = 64$,
- $d_{\text{inner hid}} = 1024$,
- $d_k = 64$,
- $d_{model} = 512$,
- $d_v = 64$,
- $d_{\text{word vec}} = 512$,
- $dropout = 0.1$,
- $epochs = 50$,
- $max_{\text{token seq len}} = 59$,
- $n_{head} = 8$,
- $n_{layers} = 6$,
- $n_{\text{warmup steps}} = 4000$
- $l_{rate} = d_{model}^{-0.5} min(step^{-0.5}, step^{-1.5} \cdot n_{\text{warmup steps}})$

5. RESULTS

The results in translation tasks are very difficult to be asserted. The most "precise" way nowadays is human evaluation, but can take long time to finish and for this sign language task will require concrete experts what makes the problem even

[3] https://github.com/jadore801120/attention-is-all-you-need-pytorch

harder. In order to try to have a simple-to-achieve and objective measure of how good a Machine Translation (MT) system behaves, the BLEU score was created.

In order to try to show qualitative results, some examples from the test set translation can be shown in Table 3. As commented in Section 4.1, the ASL is not annotated and it use special words (*X-I*, *DESC-OPEN*, *DESC-CLOSE*) Also, as said in the previous section, the vocabulary size is not as big as it should be and some words appears just once. In the translation results some unknown words (*<unk>)* appear as an example. Neither the concrete digits nor *MOBILIATION* are not learned, as it can be seen. The mentioned tokenization errors should be noticed too ("*X-I*" $\neq$ "*x @-@ i*").

Table 3. Some qualitative result examples

English:	*i believe that this is an open question .*
ASL Gloss:	*X-I BELIEVE THAT THIS BE DESC-OPEN QUESTION .*
Translation:	*<s> x @-@ i believe that this be desc @-@ open question . </s>*
English:	*mobiliation of the european globalisation adjustment fund lear from spain*
ASL Gloss:	*MOBILIATION EUROPEAN GLOBALISATION ADJUSTMENT FUND LEAR FROM SPAIN*
Translation:	*<s> <unk> european globalisation adjustment fund <unk> from spain </s>*
English:	*the sitting closed at 23.40*
ASL Gloss:	*SIT DESC-CLOSE AT 23.40*
Translation:	*<s> sit desc @-@ close at <unk> </s>*

Finally, to show an objective measure for this task results, the **BLEU score is 17.73**.

6. CONCLUSIONS AND FUTURE WORK

7. REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017.

[2] A. Othman and M. Jemni, "English-asl gloss parallel corpus 2012: Aslg-pc12," 05 2012.

[3] World Health Organization, "Deafness and hearing loss," tech. rep., 2017.

[4] 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)*, Oct 2017.

[5] O. Koller, J. Forster, and H. Ney, "Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers," *Computer Vision and Image Understanding*, vol. 141, p. 108–125, Dec 2015.

[6] R. Cui, H. Liu, and C. Zhang, "Recurrent convolutional neural networks for continuous sign language recognition by staged optimization," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017.

[7] O. Koller, S. Zargaran, and H. Ney, "Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent cnn-hmms," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017.

[8] M. J. Davidson, "Paula: A computer-based sign language tutor for hearing adults."

[9] R. Wolfe, E. Efthimiou, J. Glauert, T. Hanke, J. McDonald, and J. Schnepp, "Special issue: recent advances in sign language translation and avatar technology," *Universal Access in the Information Society*, vol. 15, pp. 485–486, Nov 2016.

[10] A. Othman, Z. Tmar, and M. Jemni, "Toward developing a very big sign language parallel corpus," *Computers Helping People with Special Needs*, p. 192–199, 2012.

[11] K. Hall, S. Mackie, M. Fry, and O. Tkachman, "Slpannotator: Tools for implementing sign language phonetic annotation," pp. 2083–2087, 08 2017.

[12] A. Othman, O. El Ghoul, and M. Jemni, "Sportsign: A service to make sports news accessible to deaf persons in sign languages," *Computers Helping People with Special Needs*, p. 169–176, 2010.

[13] L. Zhao, K. Kipper, W. Schuler, C. Vogler, N. I. Badler, and M. Palmer, "A machine translation system from english to american sign language," in *Proceedings of the 4th Conference of the Association for Machine Translation in the Americas on Envisioning Machine Translation in the Information Future*, AMTA '00, (London, UK, UK), pp. 54–67, Springer-Verlag, 2000.

[14] M. Rayner, P. Bouillon, J. Gerlach, I. Strasly, N. Tsourakis, and S. Ebling, "An open web platform for rule-based speech-to-sign translation," 08 2016.

[15] A. Othman and M. Jemni, "Statistical sign language machine translation: from english written textto american sign language gloss," vol. 8, pp. 65–73, 09 2011.

[16] M. E. Bonham, "English to ASL Gloss Machine Translation," Master's thesis, Brigham Young University, 2015.

[17] P. Koehn, R. Zens, C. Dyer, O. Bojar, A. Constantin, E. Herbst, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, and et al., "Moses," *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions - ACL '07*, 2007.

[18] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[19] Facebook, "Pytorch." `https://github.com/pytorch/pytorch`, 2016.

[20] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," 2014.