Translation of Sign Language Glosses to Text Using Sequence-to-Sequence Attention Models

Nikolaos Arvanitis, Constantinos Constantinopoulos, Dimitrios Kosmopoulos
University of Patras
Rion-Patras 26504, Greece

***Abstract*—This work deals with the problem of Sign Language Translation and more specifically with translating Glosses to text. We applied Sequence to Sequence models with attention mechanism to a parallel gloss to English corpus. This is the first work that used these models to translate American gloss sentences to English. We present our experiments on several network architectures with three different attention functions. The results are very promising and can be useful for the further implementation of a full sign language recognition system.**

***Index Terms*—sign language translation, gloss to text, SLT, sequence-to-sequence, encoder-decoder, attention mechanism, GRU**

I. Introduction

Communication among deaf-mute and hearing-impaired people is typically achieved using Sign Language. However, it is of major importance for them to find means of communication with people that can not sign or do not understand Sign Language at all. Of course, texting is a way to overcome this problem, especially nowadays with the common use of smart phones. But typically deaf people have a lot of difficulties in reading or writing texts, due to their poor language experiences and due to their limited exposure to this type of communication [11], [12].

So, there is a need of developing algorithms that can render Sign Language into text or even better voice. Specifically, translating Sign Language to text is a difficult and challenging problem considering that signing includes hand gestures, facial expressions and body pose. In order to model and analyze glosses (sign language words) all of the above channels of information should be utilized, achieving the mapping of the video features to their proper translation.

Sequence-to-sequence models with attention mechanism have been successfully applied to translation from a language to another [2], [3], [9]. In the current work, we applied sequence-to-sequence attention models to solve the problem of gloss sentences to text. The glosses can be the output of a visual sign-language translation system (e.g. [32]). Already having gloss sequences and their corresponding word sequences, makes the translation feasible with the use of the aforementioned models.

email: arvanit@ceid.upatras.gr, dkosmo@upatras.gr, kkonstantino@upatras.gr

The contribution of this work lies in presenting a method for translating Sign Language glosses to text. To our knowledge it is the first time that a parallel corpus dataset of this size is evaluated (American Sign Language-Parallel Corpus 2012 [1]). Previous attempts, e.g. [24], used a much smaller corpus, which may not be enough to demonstrate the potential of a machine translation method. Here we implemented and evaluated experimentally a sequence-to-sequence attention system, using two different architectures with promising results.

In the next two sections, we briefly look into the past works both on sign language recognition or translation and sequence to sequence models emphasizing on how attention mechanism improves the behaviour of these models. In section IV, the specific architectures used are analyzed and the experimental results are provided. Finally section V concludes the paper and describes future steps.

II. Related Work

Sign Language processing has many difficulties when trying to extract text from visual signs. Firstly, the amount of frames that correspond to a gloss is not fixed. Also, sign language includes manual and non-manual cues and we need to capture all the useful channels of information somebody uses to sign and map these features to some text. The most serious efforts on the problem took place during the last decade and were more focused on recognising the gloss out of isolated frames or to a continuous sign recognition, but all these approaches did not have satisfying results. The first attempts were influenced from automatic speech recognition methods using Hidden Markov models [35], [36], [37] and the more recent ones focused on Convolutional (CNN) and Recurrent Neural Networks (RNN) language models [31], [32], [33], [34]. Till the rise of sequence-to-sequence neural models, the results were rather poor.

A. *Sign Language Translation*

Translating Sign Language to a spoken one, means to capture video from signers, process the frames to extract meaningful features and map these features to the corresponding text sentence. However, there are not a lot studies (and also not many datasets) that translate Sign Language directly from video frames to text. In Neural Sign Language Translation [24], Koller et al. continued their previous work applying sequence to sequence models and were the first that created and made freely available PHOENIX 2014T dataset with annotation both on gloss and on German language [28]. They made three groups of experiments, that is translation of gloss sequences to text, mapping frames directly to text and translating gloss sequences to text after having estimated the glosses out of the frames. The advantage of our work compared to Neural Sign Language Translation is that PHOENIX 2014T includes 8257 parallel (train, validation and test) data which is much smaller than ASLG-PC12 that we used and is made up of 87710 parallel sequences.

In [25], [29] the authors after creating their own dataset (KETI sign language dataset), also followed the approach of attention sequence models but based on the estimation of human keypoints with the help of OpenPose [26] and they got decent results. Lately, the authors of [27] introduced a hybrid system which combines rule-based and statistical translation approaches in order to translate Turkish sign language.

Our work differs from the aforementioned in the dataset used and in the gloss level of translation. Our objective is similar with the first group of experiments of [24], that is translating glosses to text, but we aim in translating American sign language to English and not in German Sign language to German text.

III. Sequence-to-Sequence Models

Translating text from a language to another using sequence-to-sequence (or encoder-decoder) models was firstly proposed by Kalchbrenner and Blunsom [2], Sutskever et al. [3] and Cho et al. [4]. Let us explain how these models work and achieve translation. These models, try to learn-encode information of the whole input sequence and pass this encoded message to the decoder to produce the expected word in each time step. Having a sequence in the input and output, means that there is a dependency of each time input with each previous. This time-dependency and the variable-length input/output sequence, raises the need for using recurrent neural networks.

A. RNN as Encoder-Decoder

Thus, there is a combination of two recurrent neural networks; one for the encoder and another one for the decoder model. The encoder RNN reads a word, as a word embedding vector step by step. Word embeddings are real-valued vector dense representations that carry information about the meaning of the word and encode semantic similarity among the words of the vocabulary [19], [20]. By the end of reading the whole source sequence, the hidden state of the encoder RNN includes a context vector (c in Figure 1); that is a summary of the input. While encoder operates as an ordinary recurrent network, decoder differs by the fact that apart from the previous output and the hidden state it has an additional input of the context vector in order to predict the next output. Equations (1) and (2) describe RNN-Encoder and RNN-Decoder hidden unit in a sequence-to-sequence model.

$$\mathbf{h_t^{Encoder}} = RNN(\mathbf{x_t}, \mathbf{h_{t-1}^{Encoder}}) \tag{1}$$

$$\mathbf{h_t^{Decoder}} = RNN(\mathbf{y_{t-1}}, \mathbf{ch_{t-1}^{Decoder}}) \tag{2}$$

where in the above equations $\mathbf{x_t}$ is the input at time t, $\mathbf{y_{t-1}}$ is the previous output, $\mathbf{h_{t-1}}^{Encoder}$, $\mathbf{h_{t-1}}^{Decoder}$ are the encoder and decoder hidden outputs respectively and $\mathbf{c}$ is the context vector (encoder output at last input time step). Given a source word sequence in the input, the whole encoder-decoder model aims at maximizing the probability of a correct target word sequence.

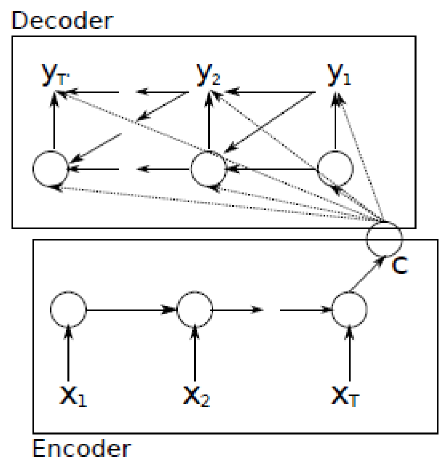

Fig. 1. An encoder-decoder model, where C depicts the context vector of the input encoded information [8].

There are however some problems encountered by the use of RNN and by the structure of the encoder. There is the known problem of the vanishing or exploding gradient during training [5], [6]. This is partially dealt with the use of Gated Recurrent Unit (GRU) RNN which is less computationally expensive, but more importantly less vulnerable to the gradient problem than the Long Short Term Memory (LSTM) [7]. Also, it is a problem that the encoder needs to encode all the input sequence information in a fixed length vector. When the model is tested with an input of a longer sequence than those of the training set, it will be difficult for it to produce acceptable results.

B. Attention Mechanism

To address the aforementioned problems of the classic encoder-decoder structure, attention mechanism was introduced by Bahdanau et al. in [9] and by Luong et al. in [10]. The goal of attention mechanism is to align encoder and decoder hidden states. Previously, just the last hidden state of the encoder was passed to the decoder as an encoded summary of the input. In attention mechanism each time step a context vector is computed as a linear combination of the alignment vector and the encoder hidden states, that is:

$$\mathbf{c_t} = \sum_{s=1}^{T} \mathbf{a_t}(\mathbf{s})\mathbf{h}_s \tag{3}$$

each time step the alignment vector $\mathbf{a_t}(\mathbf{s})$, is computed by the below equation:

$$\mathbf{a_t}(\mathbf{s}) = \frac{exp(score(\mathbf{h}_t^T, \overline{\mathbf{h}}_s))}{\sum_{s=1}^{T} exp(score(\mathbf{h}_t^T, \overline{\mathbf{h}}_s))} \tag{4}$$

where $\mathbf{h}_t^T$ and $\mathbf{h}_s$ are the current target (decoder) hidden state compared with each source (encoder) hidden state.

Equations (5-7) show the three attention score functions as proposed by Luong [10]:

$$Dot\ function: \quad \mathbf{h}_t^T \overline{\mathbf{h}}_s \tag{5}$$

$$General\ function: \quad \mathbf{h}_t^T \mathbf{W}_\alpha \overline{\mathbf{h}}_s \tag{6}$$

$$Concat\ function: v_\alpha^T tanh(\mathbf{W}_\alpha[\mathbf{h}_t^T; \overline{\mathbf{h}}_s]) \tag{7}$$

It is worth mentioning, that the third score function is very similar to the one suggested by Bahdanau [9]:

$$Bahdanau\ function: v_\alpha^T tanh(\mathbf{W}_\alpha \mathbf{h}_t^T + \mathbf{U}_\alpha \overline{\mathbf{h}}_s) \tag{8}$$

where in the above equations, $v_\alpha^{\mathbf{T}}$, $\mathbf{W}_\alpha$, $\mathbf{U}_\alpha$ are weight parameters. After having been computed, the context vector is combined with the decoder hidden state into a concatenation layer to produce attentional hidden state:

$$\tilde{\mathbf{h}}_t = tanh(\mathbf{W}_c[\mathbf{c}_t; \mathbf{h}_t]) \tag{9}$$

where $\mathbf{W}_c$ is a weight matrix. Finally, attentional hidden states are fed into a softmax layer to produce output predictions.

$$p(y_t|y_<t, x) = softmax(\mathbf{W}_c \tilde{\mathbf{h}}_t) \tag{10}$$

IV. Experimental Results

A. Dataset and Preprocessing

For our experiments, we used the ASLG-PC12 dataset [1]. It was created due to the need of a big parallel corpus for American Sign Language. The authors presented a novel algorithm for creating glosses from English words. It contains about 87710 gloss sequences-word sequences pairs from which we used the first eighty percent of them for the needs of training and the rest twenty percent for extracting the translation results. Both training and testing samples were shuffled in a random manner and specially training samples were also shuffled before every epoch.

Based on this dataset, there has been developed an approach of a probabilistic model that builds a translation memory and with this memory, statistical machine translation was achieved [1]. Also in [23], the authors motivated by the lack of a parallel corpora between English and ASL, presented an algorithm that transforms English speech to ASL gloss.

A system (as a part of Speech2signs project) that translates English text to gloss text was introduced by Manzano in [30]. Our work is the first one using ASLG-PC12 dataset for translating gloss sequences to English word sequences using encoder-decoder models with attention mechanism.

It is worth noting some steps of preprocessing in the dataset that helped improve the results. As a normalisation step on each input sequence, we subtracted all the punctuation (commas, dots, multi spaces, exclamation mark etc.). After having the dataset loaded, we search for all the glosses/words of the output/input vocabulary that count less than 5 appearances in all sequences. All these words/glosses are replaced by the symbol 'UNK', meaning unknown word, reducing the vocabulary size to its half (as suggested in [13], [14]). This reduction of the input and output vocabulary size is a fact that helped improve our results. Specifically, our input (gloss) vocabulary is finally consisted by 5316 tokens and out output vocabulary by 6900 tokens. After that, train and test data are shuffled in a random order.

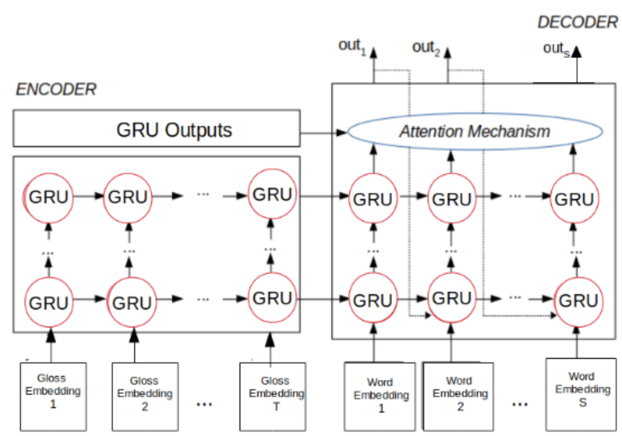

Fig. 2. Our Sign Language Translation system, for two or four encoder-decoder layers.

B. Experiments Setup

In the current work, we implemented a sequence-to-sequence system with attention mechanism for the purpose of translating a gloss input sequence to its corresponding English word output sequence, as shown in Figure 2. We implemented the whole system using PyTorch framework [21]. For the encoder and the decoder we used GRU hidden units, as they perform better than the LSTM. Each time step, GRU encoder gets in the input a gloss in a vector representation of a word embedding (gloss embedding in figure 2; allow us to put it this way). During training, the whole gloss input sequence is mapped with the corresponding output word sequence (where each word is also represented as a word embedding).

In the decoder part, we included all three Luong's attention mechanism functions. Above the decoder hidden layer, there is placed a Softmax layer to choose the proper index of the decoded word in the vocabulary and give the output. As an optimization algorithm we used Adamax [17], a variant of Adam algorithm that uses infinite order moment norm instead of Adam's second order. As the authors of Adam claim, infinite moment norm makes the algorithm more stable to noise in the gradient. As a cost function, cross-entropy measurement was chosen.

We implemented different experiments comparing the translation results for the three different attention mechanism score functions. We applied these experiments on two different encoder-decoder architectures. The first architecture includes four layers, each layer has 800 hidden nodes and was trained for 10 epochs, while the second one includes two layers, each layer having 350 hidden nodes and was trained for 5 epochs. The number of epochs on both architectures was chosen as the one demanded for the system to converge adequately.

Some settings we need to mention are the following. We set the size of batches to 32 for all the experiments and for the Adamax optimization algorithm we set the learning rate to 0.001. We also set a dropout rate of 0.25 both for the encoder and the decoder at the training phase to prevent over-fitting. To help the whole system learn better and quicker, we applied teacher forcing ([38], [39], [40]) to the decoder. Teacher forcing means that every time the decoder gets as an input its previous produced output, instead of this output we feed it with the actual output, that is the true expected word. We applied teacher forcing with a probability of 0.5 to happen per decoder input.

As far the word embedding matrix concerns, we set its dimensions equal to the length of input/output vocabulary size as the number of rows and the GRU hidden size as the number of columns. We did not make use of an existing pretrained word embedding but we used PyTorch embedding module. Pytorch embedding objects are actually parameters that are trained in an end-to-end manner along with our whole sequence-to-sequence system.

C. Evaluation Metrics

Since we have to do with natural language, the results would be better evaluated by a human, considering the fact that there may be many correct translations for a reference sentence. But there is a very useful metric for evaluating results in natural language processing, the BLEU score. BLEU score is a way to compare a translation result to a reference translation [18], [42]. BLEU metric score ranges from 0 to 1; a score of 1 means the sentence is identical to its reference. For simplicity reasons however, it is often stated on a scale of 1 to 100. BLEU uses n-grams to compute BLEU scores, looking for the presence or the absence of a word (or a group of n words, n-grams) in a sentence.

To find BLEU-4, we need to compute each n-gram BLEU score individually by comparing each reference n-gram to each

TABLE I
TRANSLATION EXAMPLES

gloss seq.	DESC-RE NEED TO BE SOME FORM SUPPORT THAT PEOPLE CAN DESC-LIVE ON IF X-Y LOSE X-Y JOB .
ground truth	there needs to be some form of support that people can live on if they lose their jobs
translation	there needs to be some form of support that people can live on if they lose their jobs
gloss seq.	X-WE WILL DECIDE DESC-LATER X-IT BE DESC-NOT DESC-NECESSARY TO DECIDE THAT DESC-NOW .
ground truth	we will decide later it is not necessary to decide that now
translation	we will decide later it is not necessary to decide that now
gloss seq.	X-WE DESC-STILL HAVE DESC-VERY DESC-IMPORTANT MOMENT FOR REFLECTION BEFORE X-WE .
ground truth	we still have a very important moment for reflection before us
translation	we still have a very important moment for reflection before us
gloss seq.	X-IT WOULD DESC-REFORE BE DESC-INCONSISTENT WITH X-WE DESC-EARLIER POSITION TO GIVE CONSENT WITHOUT DESC-FURR ADO .
ground truth	it would therefore be inconsistent with our earlier positions to give consent without further UNK
translation	it would therefore be covered with our earlier position to give the but without further UNK
gloss seq.	DESC-SOCIAL MARKET ECONOMY BE DESC-SUCCESSFUL MODEL BEHIND GERMANY X-POSS DESC-ECONOMIC MIRACLE .
ground truth	the social market economy was the successful model behind UNK economic miracle
translation	social market economy is successful a model of behind UNK economic
gloss seq.	X-I DESC-REFORE CONSIDER DESC-MANDATORY QUALITY LABEL TO BE DESC-IMPORTANT OPPORTUNITY FOR X-WE FARMER .
ground truth	i therefore consider mandatory quality labelling to be an important opportunity for our farmers
translation	i therefore consider the of quality to to be an important opportunity for us
gloss seq.	WOMAN MUST HAVE DESC-UNIVERSAL AND DESC-EASY ACCESS TO INFORMATION ON HEALTH ASPECT SEX , REPRODUCTION AND DESC-MEDICAL SERVICE .
ground truth	women must have universal and easy access to information on health aspects of sex UNK and medical services
translation	women must have a and and to to information on health the aspects aspects UNK UNK
gloss seq.	DESC-PARI CONVENTION REGULATE FREQUENCY , QUALITY AND DESC-ORGANISATIONAL PROCEDURE DESC-INTERNATIONAL EXHIBITION .
ground truth	the paris convention lays down rules on frequency quality and procedure for international exhibitions within its remit
translation	the UNK convention down down rules on the quality quality and procedures for international UNK within its

hypothesis and then we will compute their weighted geometric mean as:

$$BLEU = min(1, \frac{l_{hyp}}{l_{ref}}) \prod_{n=1}^{4} (bleu_i)^{1/4} \tag{11}$$

where l_{hyp}, l_{ref} are the hypothesis sentence length and reference sentence length accordingly and this first term is introduced to penalise sentences with length shorter than that of the reference. An example of calculating BLEU-4 score for a reference and a hypothesis sentence would be helpful to better understand this metric. Assume the next two sentences as the reference and the hypothesis accordingly:

- reference: ”Today I woke up too early”
- hypothesis: ”Today I woke up very early”

In table II, BLEU scores for the hypothesis sentence are shown:

TABLE II
SCORES FOR DEMONSTRATING BLEU COMPUTATION

BLEU-1	BLEU-2	BLEU-3	BLEU-4	BLEU
0.83	0.6	0.5	0.33	0.537

For computing the BLEU score, we used NLTK's ([16]) open source BLEU score functions. We computed BLEU-1, BLEU-2, BLEU-3 and BLEU-4 cumulative scores. We computed each BLEU score for each sentence according to its reference and then we computed the mean value (macro-average precision).

D. Results

In Table III and Table IV the results of our experiments are presented for the two encoder-decoder architectures accordingly. Somebody may notice that the BLEU-1 score gives very good results, however we should take into consideration BLEU-4 as it is the default BLEU score of NLTK library and is actually a lot more meaningful for evaluating machine translation results.

A few typical examples of the translated results, their gloss sequence and their ground truth sentence are given in Table I, as a more intuitive and demonstrative way to evaluate them.

The first three examples can be considered as qualitative translations as they have no wrong word translated. In contrast the rest examples have three or more errors. There were a lot fully correct translation results but we chose to point a little more to the wrong results and comment them.

As somebody can notice by the Translation Examples table, most of the sentences include the symbol 'UNK'. This happens due to the fact that all the replaced words with the 'UNK', actually constitute an important part of the counted words of the vocabulary. As a result when the system was about to predict a rare word, in many cases it was making the wrong choice giving 'UNK'. It was more easy for the algorithm to translate correctly words that had more appearance counts in the output vocabulary. If the vocabulary counts were more equalised (thinking the word counts as a histogram), this phenomenon would be less significant.

TABLE III
LAYERS: 2, EPOCHS: 5, HIDDEN SIZE: 350

Attention Score Func.	BLEU Score			
	BLEU-1	***BLEU-2***	***BLEU-3***	***BLEU-4***
dot	0.789	0.691	0.596	0.498
general	0.811	0.718	0.635	0.544
concat	0.788	0.690	0.601	0.503

TABLE IV
LAYERS: 4, EPOCHS: 10, HIDDEN SIZE: 800

Attention Score Func.	BLEU Score			
	BLEU-1	***BLEU-2***	***BLEU-3***	***BLEU-4***
dot	0.867	0.795	0.732	0.659
general	0.848	0.778	0.707	0.630
concat	0.863	0.790	0.725	0.651

Therefore, in order to provide better results we would need a bigger dataset with less infrequent words. In that case there would be no need to make use of the trick with the 'UNK' replacement word. Also, a Beam search decoding method probably would give a closer to the ground truth translation result [43], [44].

Considering the results on Table III and Table IV, we can be satisfied. BLEU-4 score gave the best results for the concatenation attention function with a value of 0.65 for the second architecture. Close to concatenation result is the dot one. But, the general attention function performs about the same, resulting about 0.65 score both 4-layer encoder-decoder architecture, while dot function had the poorest performance. If the first network architecture was trained for more epochs and could succeed better results, it would be preferable to choose it, combined with the general attention score for computational cost reasons.

V. CONCLUSION AND FUTURE WORK

The whole encoder-decoder system with its amount of layers and corresponding parameters did have a good performance, but is time and space expensive. The so promising Transformer attention model [45], [25] is less computational expensive and seems to give better results than the classic encoder-decoder attention models [15]. As shown above, the results are promising. A different dataset for evaluation and test purposes and a Transformer model, would construct an improved and more trustworthy combination as the major part of a Sign Language translation system.

ACKNOWLEDGEMENT

Cofinanced by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH CREATE - INNOVATE (project code: T1EDK-01299 - HealthSign).

REFERENCES

[1] Achraf Othman,Mohamed Jemni, English-ASL Gloss Parallel Corpus 2012: ASLG-PC12,proceeding of: 5th Workshop on the Representation and Processing of Sign Languages LREC12, May 2012.

[2] Kalchbrenner, N. and Blunsom, P. (2013). Recurrent continuous translation models. In Proceedings of the ACL Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 17001709. Association for Computational Linguistics

[3] Sutskever, I., Vinyals, O., and Le, Q. (2014). Sequence to sequence learning with neural networks. In Advances in Neural Information Processing Systems (NIPS 2014).

[4] Cho, K., van Merrienboer, B., Bahdanau, D., and Bengio, Y. (2014b). On the properties of neural machine translation: EncoderDecoder approaches. In Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation.

[5] Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. IEEE Transactions on Neural Networks, 5(2), 157-166.

[6] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. In Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28 (ICML'13), Sanjoy Dasgupta and David McAllester (Eds.), Vol. 28. JMLR.org III-1310-III-1318.

[7] Chung, Junyoung I& Gulcehre, Caglar I& Cho, KyungHyun I& Bengio, Y. (2014). Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling.

[8] Cho, K., van Merrienboer, B., Gulcehre, C., Bougares, F., Schwenk, H., and Bengio, Y. (2014a). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014).

[9] Bahdanau, Dzmitry I& Cho, Kyunghyun I& Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. arXiv:1409.0473v7

[10] Luong, Minh-Thang I& Pham, Hieu I& Manning, Christoper. (2015). Effective Approaches to Attention-based Neural Machine Translation. In Conference on Empirical Methods in Natural Language Processing (EMNLP) 2015.

[11] Liddell, Nathalie N. Blanger, Shari R. Baum I& Rachel I. Mayberry (2012) Reading Difficulties in Adult Deaf Readers of French: Phonological Codes, Not Guilty!, Scientific Studies of Reading, 16:3, 263-285, DOI: 10.1080/10888438.2011.568555

[12] DiFrancesca, S. 1972. Academic achievement test results of a national testing program for hearing-impaired students, Washington, DC: Gallaudet College, Office of Demographic Studies.

[13] Thng, Lng I& Sutskever, Ilya I& V. Le, Quoc I& Vinyals, Oriol I& Zaremba, Wojciech. (2014). Addressing the Rare Word Problem in Neural Machine Translation. 10.3115/v1/P15-1002.

[14] Neubig, Graham. (2017). Neural Machine Translation and Sequence-to-sequence Models: A Tutorial. arXiv:1703.01619v1

[15] Lakew, Surafel Melaku I& Cettolo, Mauro I& Federico, Marcello. (2018). A Comparison of Transformer and Recurrent Neural Networks on Multilingual Neural Machine Translation.

[16] Edward Loper and Steven Bird. 2002. NLTK: the Natural Language Toolkit. In Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics - Volume 1 (ETMTNLP '02), Vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, 63-70. DOI: https://doi.org/10.3115/1118108.1118117, https://www.nltk.org/

[17] Diederik P. Kingma and Jimmy Lei Ba. Adam : A method for stochastic optimization. ICLR 2015, Ithaca

[18] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02). Association for Computational Linguistics, Stroudsburg, PA, USA, 311-318. DOI: https://doi.org/10.3115/1073083.1073135

[19] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'13), C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Eds.), Vol. 2. Curran Associates Inc., USA, 3111-3119.

[20] Globerson, Amir (2007). "Euclidean Embedding of Co-occurrence Data". Journal of Machine Learning Research

[21] https://pytorch.org/

[22] Othman, Achraf I& Jemni, Mohamed. (2013). A Probabilistic Model for Sign Language Translation Memory. 182. 317-324. 10.1007/978-3-642-32063-7_33.

[23] Tmar, Zouhour I& Othman, Achraf I& Jemni, Mohamed. (2013). A rule-based approach for building an artificial English-ASL corpus. 1-4. 10.1109/ICEESA.2013.6578458.

[24] Camgoz, Necati I& Hadfield, Simon I& Koller, Oscar I& Ney, Hermann I& Bowden, Richard. (2018). Neural Sign Language Translation. 10.1109/CVPR.2018.00812.

[25] Sang-Ki Ko, Chang Jo Kim, Hyedong Jung, Choongsang Cho, Neural Sign Language Translation based on Human Keypoint Estimation, arXiv:1811.11436v2

[26] https://github.com/CMU-Perceptual-Computing-Lab/openpose

[27] Kayahan, D., I& Gungor, T. (2019). A Hybrid Translation System from Turkish Spoken Language to Turkish Sign Language. 2019 IEEE International Symposium on INnovations in Intelligent SysTems and Applications (INISTA). doi:10.1109/inista.2019.8778347

[28] O. Koller, H. Ney, and R. Bowden. Deep Hand: How to Train a CNN on 1 Million Hand Images When Your Data Is Continuous and Weakly Labelled. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 3793-3802, Las Vegas, NV, USA, June 2016.

[29] Ko, Sang-Ki I& Gi Son, Jae I& Jung, Hyedong. (2018). Sign language recognition with recurrent neural network using human keypoint detection. 326-328. 10.1145/3264746.3264805.

[30] Manzano, D., English to Asl Translator for Speech2signs", 2018.

[31] Cui, Runpeng I& Liu, Hu I& Zhang, Changshui. (2017). Recurrent Convolutional Neural Networks for Continuous Sign Language Recognition by Staged Optimization. 1610-1618. 10.1109/CVPR.2017.175.

[32] Camgoz, Necati I& Hadfield, Simon I& Koller, OscarI & Bowden, Richard. (2017). SubUNets: End-to-End Hand Shape and Continuous Sign Language Recognition. 10.1109/ICCV.2017.332.

[33] Koller, Oscar et al. Re-Sign: Re-Aligned End-to-End Sequence Modelling with Deep Recurrent CNN-HMMs. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017): 3416-3424.

[34] Oscar Koller, Sepehr Zargaran, Hermann Ney, and Richard Bowden. 2018. Deep Sign: Enabling Robust Statistical Continuous Sign Language Recognition via Hybrid CNN-HMMs. Int. J. Comput. Vision 126, 12 (December 2018), 1311-1325. DOI: https://doi.org/10.1007/s11263-018-1121-3

[35] Forster, Jens I& Oberdrfer, Christian I& Koller, Oscar I& Ney, Hermann. (2013). Modality Combination Techniques for Continuous Sign Language Recognition. 10.1007/978-3-642-38628-210.

[36] Koller, Oscar I& Forster, Jens I& Ney, Hermann. (2015). Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. Computer Vision and Image Understanding. 141. 108-125. 10.1016/j.cviu.2015.09.013.

[37] Forster, Jens et al. Improving Continuous Sign Language Recognition: Speech Recognition Techniques and System Design. SLPAT (2013).

[38] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent Neural networks. In Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'15), C. Cortes, D. D. Lee, M. Sugiyama, and R. Garnett (Eds.), Vol. 1. MIT Press, Cambridge, MA, USA, 1171-1179.

[39] Ferenc Huszr, How (not) to Train your Generative Model: Scheduled Sampling, Likelihood, Adversary?.arXiv:1511.05101v1

[40] Anirudh Goyal, Alex Lamb, Ying Zhang, Saizheng Zhang, Aaron Courville, and Yoshua Bengio. 2016. Professor forcing: a new algorithm

for training recurrent networks. In Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16), Daniel D. Lee, Ulrike von Luxburg, Roman Garnett, Masashi Sugiyama, and Isabelle Guyon (Eds.). Curran Associates Inc., USA, 4608-4616.

[41] Freitag, Markus and Al-Onaizan, Yaser. Beam Search Strategies for Neural Machine Translation. 2017 Proceedings of the First Workshop on Neural Machine Translation, 56-60.

[42] Callison-Burch, C., Osborne, M. and Koehn, P. (2006) "Re-evaluating the Role of BLEU in Machine Translation Research" in 11th Conference of the European Chapter of the Association for Computational Linguistics: EACL 2006 pp. 249256

[43] Sam Wiseman, Alexander M. Rush. Sequence-to-Sequence Learning as Beam-Search Optimization. Association for Computational Linguistics, Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 1296–1306

[44] Collobert, R., Hannun, A. I& Synnaeve, G.. (2019). A fully differentiable beam search decoder. Proceedings of the 36th International Conference on Machine Learning, in PMLR 97:1341-1350

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, ukasz Kaiser and Illia Polosukhin. 2017. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17), Ulrike von Luxburg, Isabelle Guyon, Samy Bengio, Hanna Wallach, and Rob Fergus (Eds.). Curran Associates Inc., USA, 6000-6010.