ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

Towards Better Transition Modeling in Recurrent Neural Networks: the Case of Sign Language Tokenization

Pierre Poitier, Jérôme Fink and Benoît Frénay

University of Namur - NaDI - Faculty of Computer Science - PReCISE
rue Grandgagnage 21, B-5000 Namur - Belgium

Abstract. Recurrent neural networks can be used to segment sequences such as videos, where transitions can be challenging to detect. This paper benchmarks strategies to better model the transition between states. The specific task of SL video tokenization is chosen for the evaluation, as it remains challenging. Tokenizers are the cornerstone of natural language processing pipelines. There exist powerful tokenizers for text data, but sign language (SL) video tokenizers are still under development. Benchmarked strategies prove to be useful to improve SL videos tokenization, but there is still room for improvement to better model state transitions.

1 Introduction

Recurrent architectures have proven to be effective at processing sequential data. However, such models suffer from limitations that may prevent them to correctly model state transitions. Several architectures have been developed to mitigate those issues. This paper aims to assess their effectiveness, with a specific focus on sign language tokenization. This task was chosen as it remains challenging and correct modeling of state transitions would greatly improve the performance for real-world videos. Section 2 introduces sequential models and strategies to better model state transitions, Section 3 discusses the complexity of sign language tokenization and presents the real-world dataset used in the experiments. Section 4 presents and discusses the experiments performed on the retained models. Finally, Section 5 concludes with future perspectives.

2 Sequential Models and Transition Modeling

Hidden Markov models (HMMs) [1] were among the first successful models for sequential data. However, they suffer from some limitations. First, they implicitly use a geometric distribution to approximate the duration of each state. This led to the creation of hidden semi-Markov models (HSMMs) [1, 2], where each state duration is explicitly modelled with a probability density function (PDF). A second related limitation is an assumption that state transitions only depend on the current state. Furthermore, observations are assumed to be conditionally independent, given the current state. These assumptions hinder the performance of HMMs on complex tasks. This led to the creation of recurrent neural networks (RNNs) [3] such as the long short-term memory (LSTM) [4].

Despite their achievements, LSTMs may fail to properly model state transitions. In this paper, two architectures that address this limitation are considered:

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

(i) the explicit duration recurrent network (EDRN) [5] that aims to better model state durations by relying on sub-states transitions and (ii) the Mogrifier LSTM (mLSTM) [6] that tackles the assumption that state transitions only depend on the current state. They were chosen as they represent the state of the art.

In an EDRN [5], each hidden state has D sub-states. Before moving to the next hidden state, sub-state transitions must be performed, i.e., transition to the next hidden state can be performed when the last sub-state is reached. Similar to an LSTM [4], each sub-state has a duration approximated by a geometric distribution [5]. Therefore, the duration of a state, decomposed into sub-states, is approximated by a mixture of geometric distributions and can model a complex duration distribution [5]. The standard LSTM is a specific case of EDRN.

The Mogrifier LSTM (mLSTM) [6] improves the LSTM expressiveness by using a context-dependent transition function. This is done by applying a gate between the input and the hidden state before the LSTM cell for several rounds. Consequently, the transition function depends on the current hidden state and the input. This process results in a contextualised representation of the input.

A benchmark of EDRN and mLSTM models against the standard LSTM architecture is conducted to evaluate **the improvement in transition modeling**. This benchmark is done in the specific case of sign language tokenization.

3 Sign Language Segmentation

In recent years, significant progress has been made in natural language processing (NLP). Tokenizers are important parts of NLP pipelines: this mandatory preprocessing step divides a document into a list of sequential tokens. While it is easy to find efficient tokenizers for written languages, it is not the case in sign language (SL). Unlike textual data where splits are based on white spaces or punctuation symbols, transitions between tokens are harder to detect in SL video recordings. For example, the end of a sign often overlaps with the beginning of the next one. The creation of a good SL tokenizer would be a great step toward SL translation. It could also speed up the currently manual and tedious annotations of SL video recordings, leading to even larger datasets.

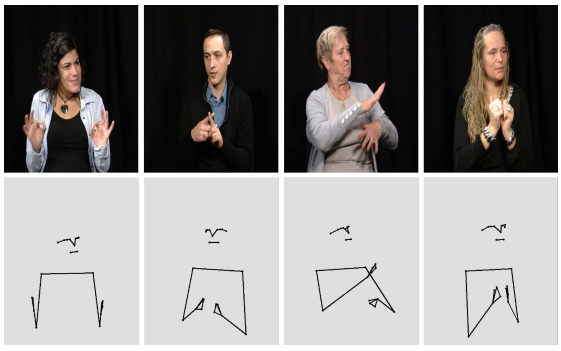


Fig. 1: Examples of frames with corresponding skeletons in LSFB-CONT.

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

Models are trained here with the large, real-world LSFB-CONT [7] dataset. It contains 50 FPS videos of individual people in real-life discussions without speed or vocabulary constraints. They are extracted from the LSFB Corpus [8], a large effort since 2012 by researchers at the University of Namur to collect and annotate LSFB (French Belgian Sign Language) conversations, with the aim to better understand this sign language. Consequently, it offers a large vocabulary (more than 6,000 words) with fast signs and short periods between them. One of the challenges is the imprecise nature of signs. Indeed, the boundaries of signs are highly dependent on the surrounding signs. The transition between two signs is called a *coarticulation* [9]. Conversations are annotated with the start and the end of all signs along with their labels. In order to avoid issues due to the large number of different signs and serious imbalance (10 most frequent signs account for 22% of annotations), we consider two simplified settings where annotations are replaced by a much smaller set of labels that only take into account the sign boundaries. In the two-class setting, labels are *talking* or *waiting*. In the three-class setting, labels are *talking*, *coarticulation* for the segments of less than 1 second between two signs and *waiting* for the others.

In this paper, tokenization is performed on the basis of pre-processed upper-body skeletons. Each skeleton consists of 23 landmarks extracted with MediaPipe [10] (see Figure 1). A model input is thus a sequence with n frames and 23 coordinates (x and y). Landmarks have been linearly interpolated to avoid discontinuity and coordinates are smoothed using a Savitzky-Golay filter [11] with a window length of 7 and a polynomial order of 2.

4 Benchmarking Transition Modeling in RNNs

This section evaluates the interest of models that aim to better model transitions, either through duration modeling (EDRNs) or contextual transitions (mLSTMs). The LSFB-CONT dataset [7] is used as described in Section 3 to create two segmentation tasks. Open research directions are highlighted.

4.1 Experimental Setting

Two experiments are carried out with two and three classes, respectively. LSTMs, EDRNs and mLSTMs are trained with a weighted cross-entropy loss, where frame weights are the inverse of the frequency of each class (see Table 1 for all details). The dataset consists of a subset of the LSFB-CONT dataset [7]: the training set and the test set contain randomly selected videos with, respectively, 3.5M and 2.2M frames. In addition, the signers are different in each set to avoid overfitting. As sequences have varying dimensions, they are windowed for batches with a window size of 1500 and a stride of 800.

All the models have 128 features in their hidden states and end with a linear layer. The optimizer is a stochastic gradient descent with a learning rate of 0.01 and a momentum with a factor of 0.9. The EDRN is trained with 4 sub-states per hidden state and the Mogrifier LSTM uses 5 rounds [6] (see Table 2).

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

	classes	frames	frequency	weight
two-class setting	talking	1,836,565	31.72%	3.15
	waiting	3,953,383	68.28%	1.46
three-class setting	talking	1,836,565	31.72%	3.15
	waiting	3,176,663	54.87%	1.82
	coarticulation	776,720	13.41%	7.45

Table 1: Summary of the two experiments with two and three classes.

	input features	hidden features	sub-states	rounds
LSTM	46	128		
EDRN	46	128	4	
mLSTM	46	128		5

Table 2: Configuration of the models used in the two experiments.

4.2 Experimental Results

Table 3 and 4 show the results of the two-class and three-class experiments, respectively. The metrics include the accuracy, the balanced accuracy and the recall for each class. In order to better apprehend the behaviour of the models, Figure 2 shows duration distributions for the *signing* and *coarticulation* classes in the three-class experiment. For reasons of readability, durations greater than 2s are not shown as we focus on the most significant part of the distribution. Figure 3 shows two examples of mistakes that occur for SL segmentation.

	acc.	bal. acc.	recall	
			talking	waiting
LSTM	81.22	80.53	63.10	**97.96**
EDRN	**82.12**	**81.06**	**64.41**	97.71
mLSTM	82.03	81.02	64.26	97.80

Table 3: Results of the two-class experiment with the LSFB-CONT dataset.

	acc.	bal. acc.	recall		
			talking	waiting	coarticulation
LSTM	73.82	66.19	70.13	97.21	31.24
EDRN	74.41	67.11	**72.02**	**97.32**	31.99
mLSTM	**77.65**	**67.78**	71.54	95.91	**35.89**

Table 4: Results of the three-class experiment with the LSFB-CONT dataset.

4.3 Discussion

In the two-class setting, EDRN outperforms both LSTM and mLSTM in terms of accuracy and balanced accuracy. However, all three models tend to ignore coarticulation and thus merge signs together as shown by Figure 3a. The three-class settings aims to mitigate this issue with the additional *coarticulation* class

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

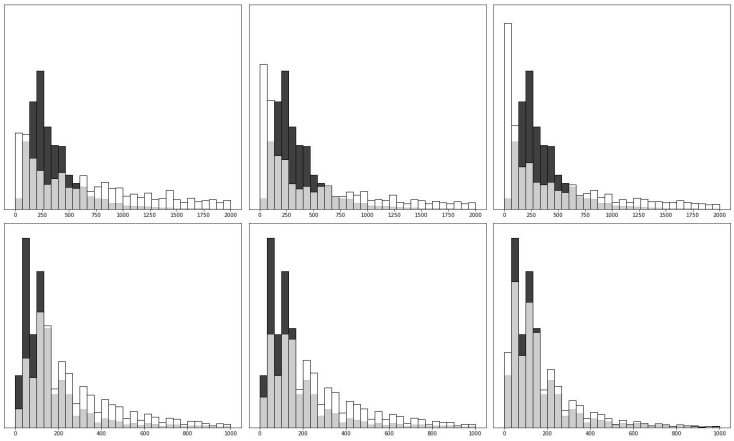
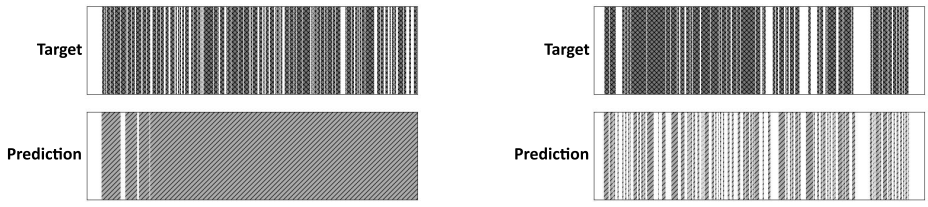


Fig. 2: Comparison of duration distributions in the dataset (black) and in the segmentation (white), for *signing* (first row) and *coarticulation* (second row) in the three-class experiment. From left to right: LSTM, EDRN and mLSTM.

that reflect how signers move from one sign to the next. In that case, mLSTM performs better in terms of accuracy and has a better recall for *transition*. EDRN has a better recall for other classes. In both settings, EDRNs and mLSTMs outperform LSTMs, showing the effect of better transitions modeling .

Figure 2 shows that the duration distributions are better modeled by the EDRN and the mLSTM. In particular, mLSTM is more accurate for the *coarticulation* class, with most of the probability mass put in the left part of the distribution and a smaller distribution tail. Yet, there is still room for improvement. The mLSTM model tends to produce very short signs as shown by the segmentation example in Figure 3b.

The EDRN also favours short signs but the problem is less important. However, the *coarticulation* duration distribution has a rather heavy tail in Figure 2, i.e., predicted transitions are too long. Notice that those observations are consistent with the higher transition recall and higher balanced accuracy measured for the mLSTM.



(a) Too long predicted signs with an LSTM in the two-class experiment.

(b) Too short predicted signs with an mLSTM in the three-class experiment.

Fig. 3: Examples of target (top) and predicted (bottom) segmentation. Black segments correspond to *talking*, white space indicates *waiting* or *coarticulation*.

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

5 Conclusion

In this paper, two experiments are carried out on the real-world task of sign language tokenization. Three RNN models are compared, which differ in the modeling of state durations and state transitions: LSTM, EDRN and mLSTM.

The first experiment highlights the difficulty of the prediction of coarticulation between signs. Both the EDRN and the mLSTM outperform the LSTM in the second experiment. mLSTM is more accurate, but it predicts too short signs. The EDRN is more accurate than the LSTM and predicts longer signs compared to the mLSTM. However, predicted coarticulations are too long.

This paper demonstrates the interest of developing better models of state transitions and state durations for RNNs through the case of SL tokenization. However, it also demonstrates that there is room for improvement. Future work includes proposing new powerful models in that direction. To the best of our knowledge, there exist only few works [12, 13], that tackle SL tokenization, but they do not focus on the problem of state transitions and durations.

Acknowledgments

This work is supported by the Funds InBev-Baillet Latour and the F.R.S.-FNRS EOS VeriLearn project n. 30992574. The authors thank Valentin Delchevalerie, Mohammed El Adoui and Géraldin Nanfack for their comments.

References

[1] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[2] S.-Z. Yu. Hidden semi-markov models. *Artificial intelligence*, 174(2):215–243, 2010.

[3] D. E. Rumelhart, G. Hinton, and R. J. Williams. Learning internal representations by error propagation. Technical report, California University San Diego, 1985.

[4] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[5] S.-Z. Yu. Explicit duration recurrent networks. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–11, 2021.

[6] G. Melis, T. Kočiskỳ, and P. Blunsom. Mogrifier LSTM. *arXiv:1909.01792*, 2019.

[7] J. Fink, B. Frénay, L. Meurant, and A. Cleve. LSFB-CONT and LSFB-ISOL: Two new datasets for vision-based sign language recognition. In *Proc. IJCNN*, pages 1–8, 2021.

[8] L. Meurant. Corpus LSFB. Corpus informatisé en libre acces de vidéo et d'annotations de langue des signes de Belgique francophone. Namur: Laboratoire de langue des signes de Belgique francophone (LSFB Lab), FRS-FNRS, Université de Namur, 2015.

[9] L. Naert, C. Larboulette, and S. Gibet. Coarticulation analysis for sign language synthesis. In *Proc. UAHCI*, pages 55–75, 2017.

[10] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann. Mediapipe: A framework for building perception pipelines. *arXiv:1906.08172*, 2019.

[11] A. Savitzky and M. J. E. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639, 1964.

[12] J. Zheng, Z. Zhao, M. Chen, J. Chen, C. Wu, Y. Chen, X. Shi, and Y. Tong. An improved sign language translation model with explainable adaptations for processing long sign sentences. *Computational Intelligence and Neuroscience*, 2020, 2020.

[13] A. Orbay and L. Akarun. Neural sign language translation by learning tokenization. In *Proc. IEEE FG*, pages 222–228, 2020.