Incremental Processing for Neural Conversation Models

Pierre Lison
Norwegian Computing Center
Oslo, Norway
plison@nr.no

Casey Kennington
Boise State University
caseykennington@boisestate.edu

Abstract

We present a simple approach to adapt neural conversation models to incremental processing. The approach is validated with a proof-of-concept experiment in a visual reference resolution task.

1 Introduction

The last recent years have witnessed the emergence of new dialogue modelling approaches based on recurrent neural networks (Vinyals and Le, 2015; Lowe et al., 2017). One neglected aspect of these neural models is that they effectively construct a latent representation of the dialogue state on a token-by-token basis. However, despite this conceptual proximity with incremental approaches to dialogue processing (Schlangen and Skantze, 2011), these neural models have so far always been applied to fully fledged utterances.

We present in this abstract a simple approach for adapting neural conversation models to process incremental units instead of fixed sequences of tokens. This model is able to not only process words one at the time, but also commit or revoke these words at any point during processing.

2 Incremental model

Assume a neural model such as the one illustrated in Figure 1(a). The model takes a sequence of tokens as inputs and transforms this sequence with an embedding layer followed by a recurrent layer (such as an LSTM or a GRU). The sequence length must typically be fixed in advance (by e.g. determining a maximum length and using padding to encode shorter utterances). At the end of the sequence, the model outputs a fixed-size vector representing the dialogue. The model parameters comprise both the embeddings themselves and the weights of the recurrent units. These parameters are optimised on a particular task such as predicting the next utterance in the dialogue.

Once the network parameters are learned, one can construct an equivalent, incremental version of the same model using the following approach. Instead of taking a sequence of tokens as inputs, we adapt the network by reducing the input length to one single token, and adding a new type of input, namely a fixed-size vector representing the dialogue processed so far. The network outputs a new, updated vector after each token. The embeddings and the weights of the recurrent units remain identical to the ones in the non-incremental model. The resulting model is illustrated in Figure 1(b).

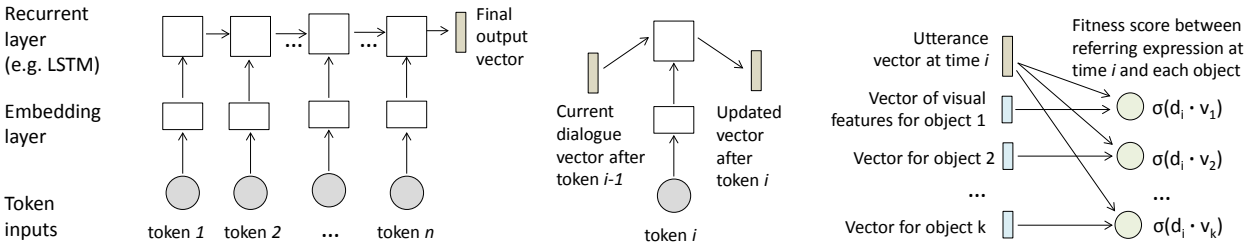
When a new word is inserted into the incremental system, the neural model is triggered to produce another vector expressing the updated dialogue state. The history of previous state vectors is kept in memory until their corresponding words are committed by other modules. This allows the system to "backtrack" to previous state vectors whenever incremental units are revoked.

Thanks to the continuous nature of the vectors generated by the neural network, uncertain inputs (for instance incremental units associated with confidence scores from speech recognition) can be handled by simple algebraic operations. Let d_{i-1} represent the fixed-size vector for the dialogue at time $t-1$ and w_i a new word hypothesis with probability p_i. The updated vector after processing w_i can be defined as an interpolation between the previous vector d_{i-1} and the output of the neural model $N(d_{i-1}, w_i)$:

$$d_i = p_i N(d_{i-1}, w_i) + (1 - p_i) d_{i-1} \qquad (1)$$

3 Experiments

This neural incremental model has been implemented and evaluated in a simple proof-of-concept



(a) Standard neural conversation model. (b) Incremental version. (c) Visual reference resolution task.

Figure 1: On the left, a standard neural conversation model taking a token sequence as inputs and producing a fixed-size output vector. In the middle, an incremental version of the same neural model, taking two inputs (the current dialogue vector and a new token) and producing an updated vector. On the right, the application of the neural model for the visual reference resolution task described in Section 3.

experiment with the TAKE corpus from the PentoRef collection (Zarrieß et al., 2016). The corpus includes 1045 utterances recorded through a Wizard-of-Oz study where the participants had to choose one Pentomino title among 15 titles on a game board and then instruct the system to select it through verbal descriptions and pointing gestures.

To apply the model to this visual reference resolution task, the model was extended with another layer computing the dot products of the utterance with a list of vectors encoding the visual features of each tile in the scene, normalised with a sigmoid function. The model was trained on both positive and negative examples (the distractors in each scene). The model is similar to a Dual Encoder model (Lowe et al., 2017), except the dot products are here computed between referring expressions and visual objects. The utterance vector can therefore be viewed as encoding a "prediction" on the visual features of the target object. The neural model is illustrated in Figure 1(c).

The speech recordings of all TAKE episodes were then transcribed by the streaming Google Speech API in order to obtain a list of incremental operations (comprising not only insertions, but also revoke and commit operations). After each incremental operation, the neural model was triggered to obtain an updated vector and determine the fitness scores between each object and the utterance observed so far. The accuracy on the task of selecting the right target object was measured at each incremental step. The results, shown in Figure 2, show that the accuracy increases as more words are processed. The final accuracy after processing the full utterances is 0.669 when applied to the noisy ASR transcriptions, and 0.87 when applied to the manual transcriptions.

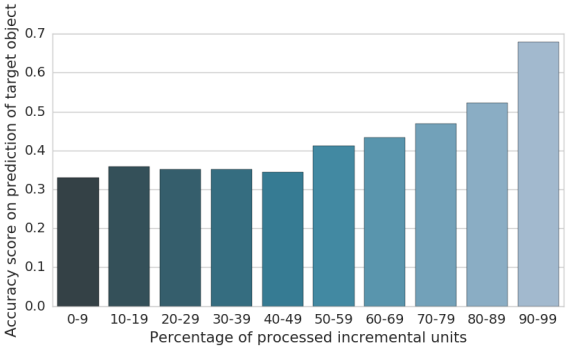


Figure 2: Evaluation results on the visual reference resolution task on the TAKE dataset.

4 Conclusion

We presented a simple approach to make neural dialogue models "incremental" – that is, able to operate on incremental units instead of on complete utterances. The model can handle insertions, commit and revoke operations as well as incremental units associated with probabilities. A proof-of-concept experiment on a visual reference resolution task shows the promise of the approach.

References

R. Lowe, N. Pow, I. Serban, L. Charlin, C.-W. Liu, and J. Pineau. 2017. Training end-to-end dialogue systems with the Ubuntu Dialogue Corpus. *Dialogue & Discourse*, 8(1):31–65.

D. Schlangen and G. Skantze. 2011. A general, abstract model of incremental dialogue processing. *Dialogue & Discourse*, 2(1):83–111.

Oriol Vinyals and Quoc Le. 2015. A Neural Conversational Model. *CoRR*, abs/1506.05869.

S. Zarrieß, J. Hough, C. Kennington, R. Manuvinakurike, D. DeVault, R. Fernandez, and D. Schlangen. 2016. PentoRef: A Corpus of Spoken References in Task-oriented Dialogues. In *Proceedings of LREC 2016*, Portoro, Slovenia.