Incremental Processing for Neural Conversational Models

Pierre Lison
SAMBA- Statistical analysis, machine learning & image analysis
Norsk Regnesentral, Oslo, Norway

Casey Kennington
Computer Science Department
Boise State University, USA

Key idea

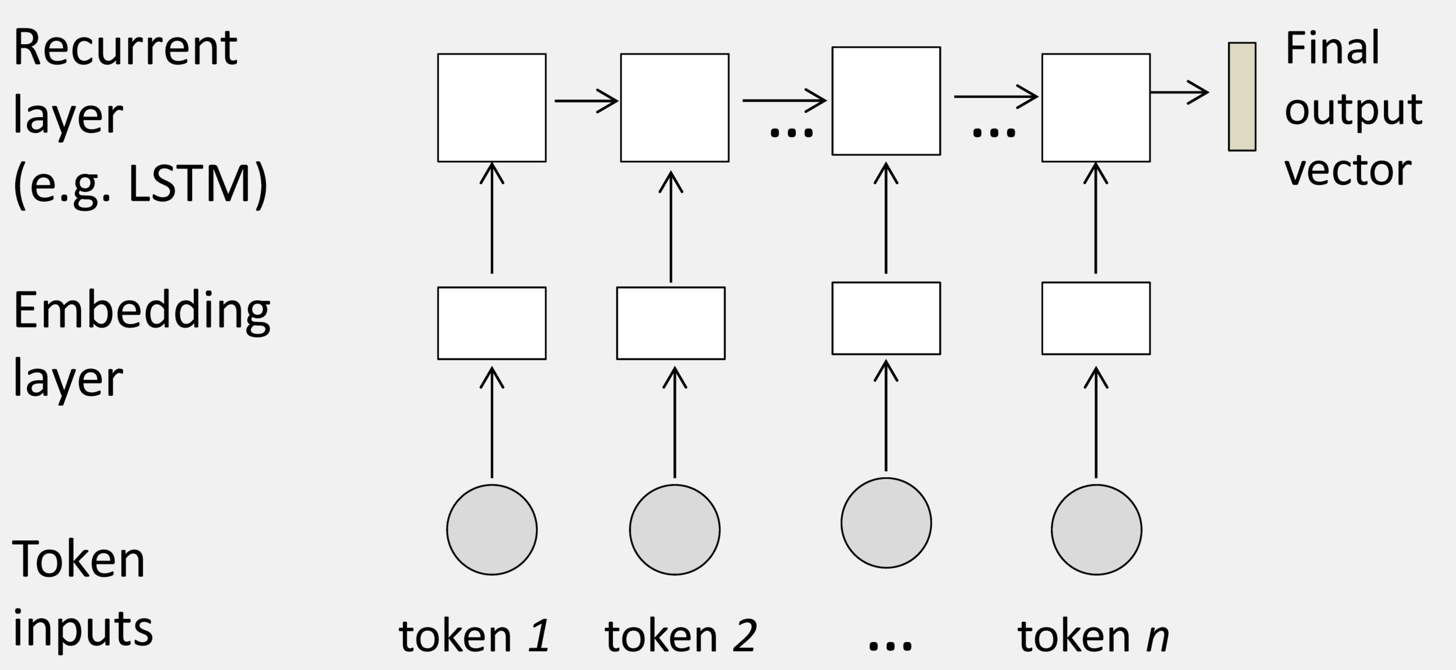
- Increasing popularity of dialogue modelling approaches based on recurrent neural networks
- These neural models construct a latent representation of the dialogue state on a *token-by-token* basis.

  → conceptual proximity with incremental approaches to spoken dialogue processing
- However, in practice, these neural models are always applied to fully fledged sentences.

***Question**: Can we adapt neural conversation models to operate on **incremental units** instead of fixed sequences of tokens?*

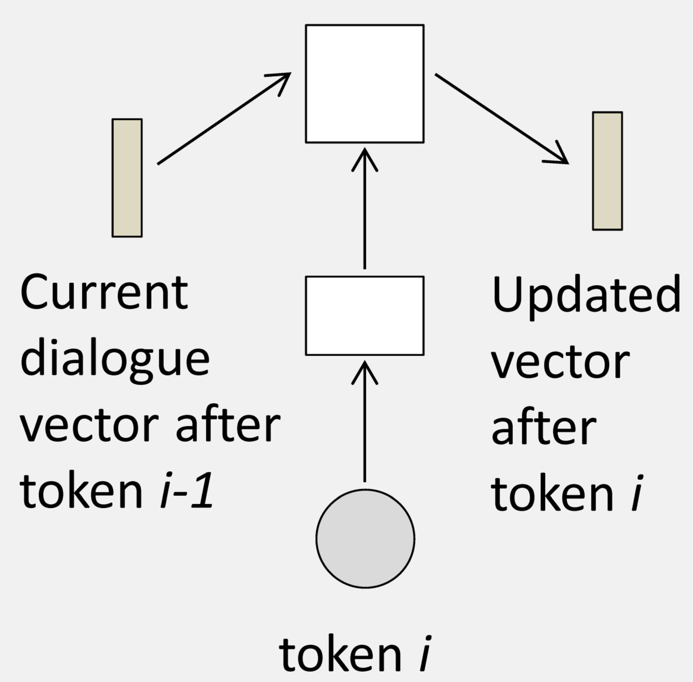
Yes!
The presented model is able to process incremental units (IUs) one at the time, through a sequence of updates
… and commit/revoke IUs at any point during processing

Non-incremental model



Incremental model

- Once the model parameters are learned, we can construct an equivalent, incremental version of the same neural model
- The network architecture is modified to take two inputs: a single input token + the previous dialogue state

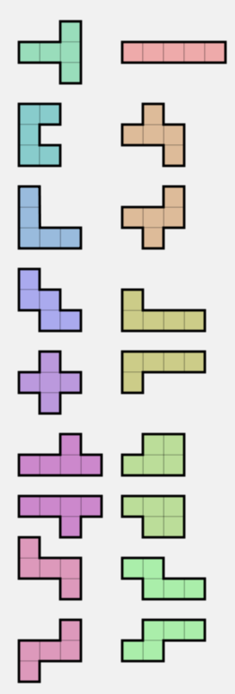


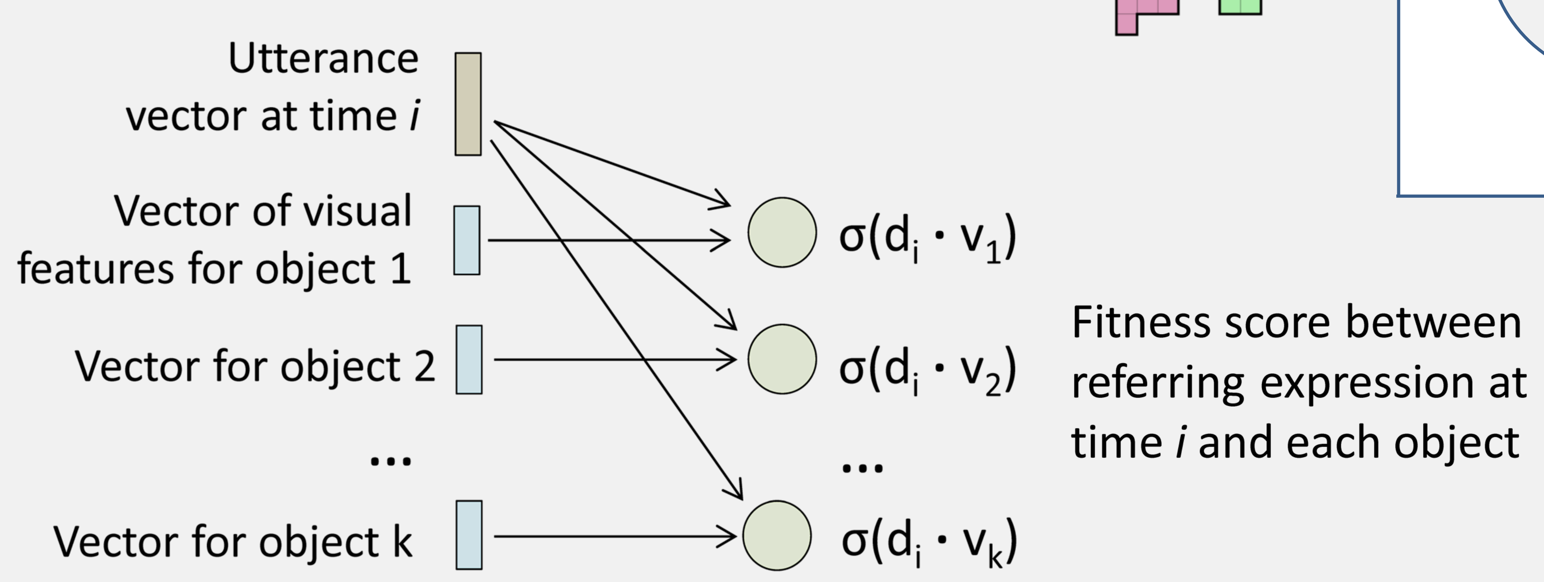
- The network outputs the updated state vector
- The model parameters remain unchanged
- The history of previous state vectors is kept in memory. This allows the system to backtrack to previous (not-yet-committed) state vectors whenever incremental units are revoked.
- To deal with uncertainty/ambiguities (coming from e.g. speech recognition), we can interpolate the vectors: If d_{i-1} represent the dialogue vector at time t-1 and w_i a new word hypothesis with probability p_i, the updated vector d_i can be defined as

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

Experiment

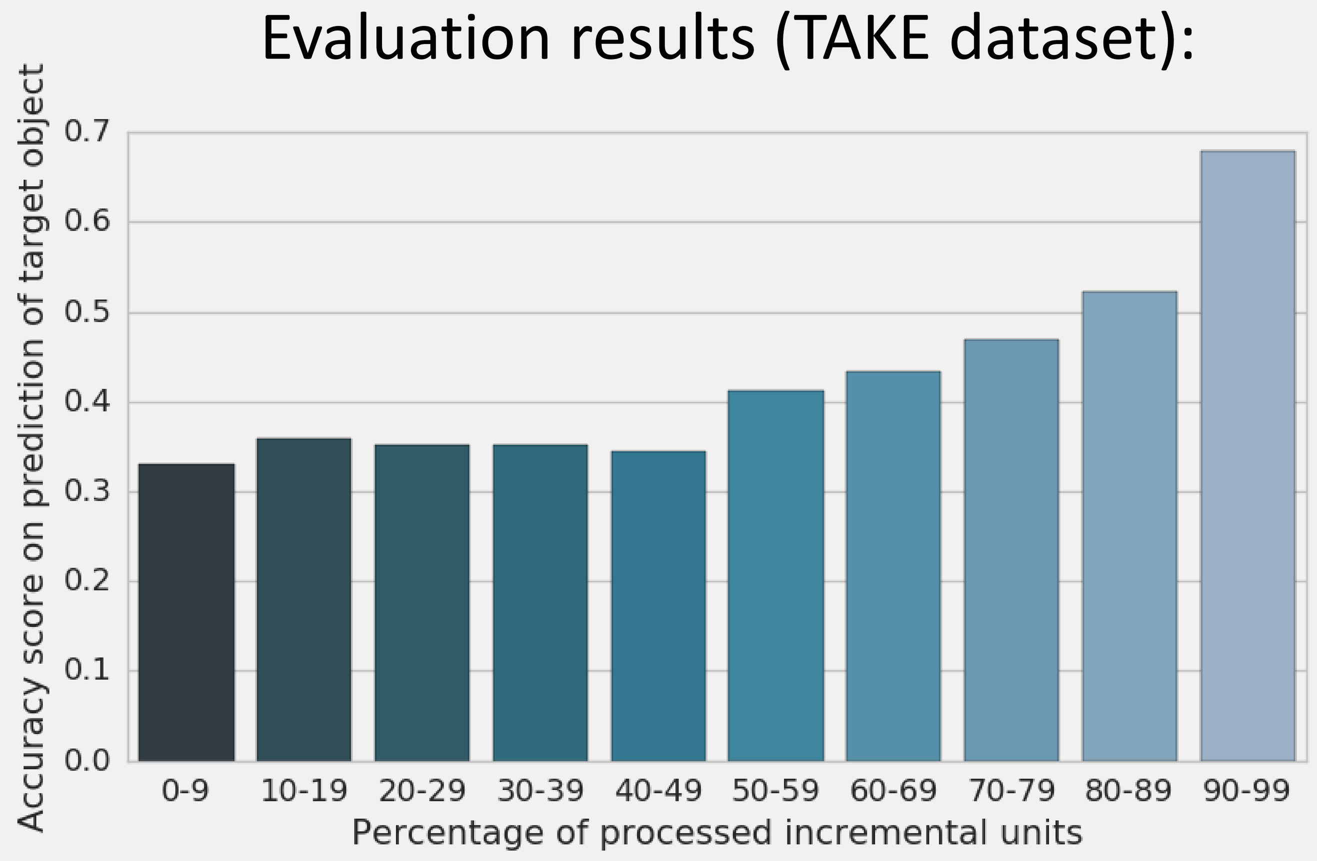
- Proof-of-concept experiment with the TAKE corpus (Wizard-of-Oz study where participants had to instruct the system to select one tile from a virtual Pentomino board through verbal descriptions and pointing gestures)
- The neural network for this visual reference resolution task relies on the dot product of visual and utterance vectors:





- Training on positive and negative examples (the distractors in each scene)
- The streaming Google Speech API was used to obtain incremental operations from the TAKE episodes (insertions, revoke and commit operations).
- After each operation, the neural model was triggered to update the utterance vector and determine the fitness scores of each visual object
- The accuracy (defined as the selection of the right target object among 15 objects in each scene) increases as more words are processed.

Evaluation results (TAKE dataset):



[Final accuracy after processing the full utterances: 0.67 for noisy ASR, 0.87 for manual transcriptions]