

From Phonemes to Robot Commands with a Neural Parser

Xavier Hinaut

▶ **To cite this version:**

Xavier Hinaut. From Phonemes to Robot Commands with a Neural Parser. IEEE ICDL-EPIROB Workshop on Language Learning, Sep 2017, Lisbon, Portugal. pp.1-2. hal-01665823

HAL Id: hal-01665823
https://inria.hal.science/hal-01665823

Submitted on 17 Dec 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

From Phonemes to Robot Commands with a Neural Parser

Xavier Hinaut
Inria Bordeaux Sud-Ouest, Talence, France
LaBRI, UMR 5800, CNRS, Bordeaux INP, Université de Bordeaux, Talence, France
Institut des Maladies Neurodégénératives, UMR 5293, CNRS, Université de Bordeaux, Bordeaux, France
xavier.hinaut@inria.fr

Abstract*—The understanding of how children acquire language [1][2], from phoneme to syntax, could be improved by computational models. In particular when they are integrated in robots [3]: e.g. by interacting with users [4] or grounding language cues [5]. Recently, speech recognition systems have greatly improved thanks to deep learning. However, for specific domain applications, like Human-Robot Interaction, using generic recognition tools such as Google API often provide words that are unknown by the robotic system when not just irrelevant [6]. Additionally, such recognition system does not provide much indications on how our brains acquire or process these phonemes, words or grammatical constructions (i.e. sentence templates). Moreover, to our knowledge they do not provide useful tools to learn from small corpora, from which a child may bootstrap from. Here, we propose a neuro-inspired approach that processes sentences word by word, or phoneme by phoneme, with no prior knowledge of the semantics of the words. Previously, we demonstrated this RNN-based model was able to generalize on grammatical constructions [7] even with unknown words (i.e. words out of the vocabulary of the training data) [8]. In this preliminary study, in order to try to overcome word misrecognition, we tested whether the same architecture is able to solve the same task directly by processing phonemes instead of grammatical constructions [9]. Applied on a small corpus, we see that the model has similar performance (even if a little weaker) when using phonemes as inputs instead of grammatical constructions. We speculate that this phoneme version could overcome the previous model when dealing with real noisy phoneme inputs, thus improving its performance in a real-time human-robot interaction.

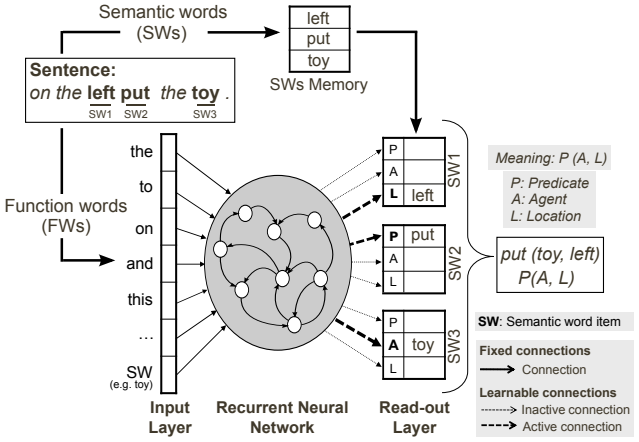


Figure 1. The core part of the model is made of an Echo State Network (ESN) [10], a Recurrent Neural Network (RNN) where only the ouput layer is trained, whereas input and recurrent weights are initialized randomly and kept constant afterwards. Sentences are converted to a grammatical construction (i.e. sentence structure) by replacing semantic words by a Semantic Word (SW) marker. The ESN is given the grammatical construction word by word. Each word activates a different input unit. During training, the connections to the readout layer are modified to learn the mapping between the grammatical construction and the arguments of the predicates. When a sentence is tested, the most active units are bound with the SW kept in the SW memory to form the resulting predicate. (Adapted from [4].)

I. Introduction

Hereafter, we briefly present the previous model in Figure 1 (for more details on the model please refer to [7][8]). Then we describe the modifications performed for the phoneme version of the model. After that, we present the corpus and particular aspects which makes it difficult to learn. Finally, we present preliminary results.

II. Methods

A. Phoneme Extension of the Model

We extend the model to process directly sequence of phonemes instead of sentences structures (i.e. sequence of words with semantic words replaced with SW markers). Hereafter we describe the training procedure.

First, an instance of ESN is created by generating the random weights for the input connections and the recurrent connections. Each word is replaced by its corresponding list of phonemes based on the Carnegie Mellon University word-phoneme correspondence dictionary (CMUdict v0.07). No additional space or any other clue enabling the model to detect boundary of words was added. Each sentence is then encoded as a succession of input unit activations in a localist (i.e. hot-vector) fashion: 1 for the corresponding phoneme, and 0 for others. The output are generated as in the previous model, as if grammatical constructions were processed. Thus, the model have no simple cue indicating that a semantic word is being processed (i.e. activation of the SW input unit, as it would be the case in the previous model).

*This work was partly supported by the PHC PROCOPE (Campus France - DAAD) for LingoRob project.

B. Corpus

The corpus was obtained by asking naive users (agnostic about how the system works) to watch several actions in a video and give the commands corresponding to the motor actions performed, as if they wanted a robot to perform the same action. The video shown to the users is available online (as supplementary material) along with the first experiment we did with robots [4]. Five users provided 38 English commands, thus there is 190 sentences in total. The corpus is the same than in [8] and is available at https://github.com/neuronalX/EchoRob/corpora.

In the list below are some sentence examples from the noisy English corpus: first is indicated the specific aspect of the sentence, and then the sentence is given. For instance, one can see that the order of actions to be performed does not necessarily correspond to the semantic word order in the sentence: i.e. the chronological order is reversed. Note also that some sentences provided by users are grammatically incorrect.

- Sequence of actions: "Touch the circle **after** having pushed the cross to the left"
- Sequence of actions: "Put the cross on the left side and **after** grasp the circle"
- Implicit reference to verb: "**Move** the circle to the left **then the cross to the middle** "
- Implicit reference to verb and object: "**Put** first the triangle on the middle **and after on the left** "
- "Crossed reference": "***Push*** **the triangle** and *the circle* ***on the middle***"
- Repeated action: "Hit **twice** the blue circle"
- Unlikely action: "Put the cross to the right and **do a u-turn**"
- Particular function word: "Put **both** the circle and the cross to the right"

III. Results

We performed a random search focusing on the three main hyper-parameters (spectral radius, input scaling and leak rate)[1] and among the best results we took the parameter set that contain the less number of recurrent units. During this search, we choose to fix the spectral radius to 1 for two reasons: firstly people in the community tend to use values close to one for various reasons (for instance the Echo State Property[10]), secondly because the spectral radius parameter is interacting with the two other parameters. The parameter set we found is the following: 200 reservoir units, leak rate: 0.1667, input scaling: 0.073. Input and recurrent matrices were fully connected.

Over 50 runs of different instances of the network (i.e. random weight for input and recurrent matrices) with a 10-fold cross-validation, we obtained a sentence error[2] of 28.0% (+/- 2.3). Even if a weaker, this error is not so far (for a preliminary result without preprocessing) compared to the 21.4% (+/- 2.2) obtained in [8] (with different hyper-parameters: input scaling: 0.03, leak rate: 0.2; the spectral radius was still 1). In fact, the current experiment needed less reservoir units (200 instead of 500) compared to previous experiment with grammatical constructions and preprocessing of the input (i.e. replacement of infrequent function words by a IF marker, which helps generalisation). Moreover, the quick parameter search we did may not be sufficient to find the best results.

[1] 300 parameter sets evaluated.

[2] The sentence error indicates the number of sentences that have not been fully recognized (i.e. at least one role for one semantic word is wrong).

IV. Discussion

It is important to note that we did not yet include other specific pre- or post-processing that could have given better results: this would be a first dimension on which we could improve the model to have better results. Moreover, the corpus used for training is only 200 sentences, which is too small to learn and extract enough regularities on phoneme sequences. Thus, we aim to repeat the experiment on a much larger corpus in order to demonstrates better performance: this would be a second dimension on which we could improve the performances. Nevertheless, the current performances are already interesting and useful for small corpus applications in Human-Robot Interaction experiments. As the core part of the model is a generic neural architecture, it could be easily reused or adapted for other computational or robotic experiments in language acquisition.

References

[1] M. Tomasello. *Constructing a language: A usage based approach to language acquisition*. Cambridge, MA: Harvard University Press, 2003.

[2] Deb Roy. New horizons in the study of child language acquisition. *INTERSPEECH-2009, Brighton, United Kingdom*, 2009.

[3] Angelo Cangelosi, Giorgio Metta, Gerhard Sagerer, Stefano Nolfi, Chrystopher Nehaniv, Kerstin Fischer, Jun Tani, Tony Belpaeme, Giulio Sandini, Francesco Nori, et al. Integration of action and language knowledge: A roadmap for developmental robotics. *IEEE Transactions on Autonomous Mental Development*, 2(3):167–195, 2010.

[4] X. Hinaut, M. Petit, G. Pointeau, and P. Dominey. Exploring the acquisition and production of grammatical constructions through human-robot interaction with echo state networks. *Frontiers in Neurorobotics*, 8, 2014.

[5] Deb Roy. Grounded spoken language acquisition: Experiments in word learning. *IEEE Transactions on Multimedia*, 5(2):197–209, 2003.

[6] J. Twiefel, T. Baumann, S. Heinrich, and S. Wermter. Improving domain-independent cloud-based speech recognition with domain-dependent phonetic post-processing. In *Twenty-Eighth AAAI. Québec City, Canada*, pages 1529–1535, 2014.

[7] X. Hinaut and P. Dominey. Real-time parallel processing of grammatical structure in the fronto-striatal system: a recurrent network simulation study using reservoir computing. *PloS one*, 8(2):e52946, 2013.

[8] X. Hinaut, J. Twiefel, M. Petit, P. F. Dominey, and S. Wermter. A recurrent neural network for multiple language acquisition: Starting with english and french. In *NIPS 2015 Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches*, 2015.

[9] A. Goldberg. *Constructions: A construction grammar approach to argument structure*. University of Chicago Press, 1995.

[10] H. Jaeger. The "echo state" approach to analysing and training recurrent neural networks. *Bonn, Germany: German National Research Center for Information Technology GMD Technical Report*, 148:34, 2001.