SPOKEN LANGUAGE ACQUISITION BASED ON REINFORCEMENT LEARNING AND WORD UNIT SEGMENTATION

Shengzhou Gao[†]*, Wenxin Hou*[†]*, Tomohiro Tanaka*[†]*, Takahiro Shinozaki*

Tokyo Institute of Technology

ABSTRACT

The process of spoken-language acquisition has been one of the topics of greatest interest to linguists for decades. By utilizing modern machine learning techniques, we simulated this process on computers, which helps to understand it and develop new possibilities of applying this concept on intelligent robots, among other things. This paper proposes a new framework for simulating spoken-language acquisition by combining reinforcement learning and unsupervised learning methods. Our experiments also show that a spoken language can be acquired considerably faster by identifying potential word segments from collected ambient sounds in an unsupervised manner.

Index Terms— Spoken language acquisition, zero resource word segmentation, reinforcement learning

1. INTRODUCTION

The mechanism that enables human beings to acquire spoken language from scratch still remains mysterious and fascinating for linguists and scientists. One prominent and widely accepted explanation, proposed by Skinner in 1957 [1], states that children acquire language based on behaviorist reinforcement principles, by associating words with meanings. Along with the advance in Artificial Intelligence (AI) and Machine Learning technologies, we are now capable of simulating this complex process on computers. In fact, computer models that simulate cognitive processes are making astonishing progress in many ways, such as playing Go [2] or video games [3], processing speech and natural language [4, 5] and recognising objects [6] or faces [7]. In this paper, we work in the direction of Skinner's idea and provide a proof of concept by running a computer-simulated experiment.

One area of research that is similar to our interest is the grounded language learning problem, which refers to associating an abstract language term to tangible objects such as images or actions or classifying this term [8, 9]. In the early years, Siskind presented a non-statistical language-grounding model that consisted of many handmade logics [10, 11]. Several reinforcement-learning-based methods have been proposed to automatically construct high-performance grounding models. Hermann et al. introduced a language-acquisition model that moves the agent around in a virtual 3D environment according to the instruction from a given sentence [12]. Yu et al. proposed a language acquisition model for question answering and sentence-directed navigation that is trained by interacting with the virtual 2D world [13]. Sinha et al. presented an attention-based language-grounding model that navigates the user to the place specified in a given descriptive sentence [14]. However, those models use texts as inputs and are not simulating the spoken language-learning processes per se.

To ground spoken language, Roy proposed a grounded spoken-language acquisition model that segments continuous speech and associates it to the visual category, using input from a camera and a microphone [15]. Chauhan et al. introduced the categorizing model, with an extended condition that the number of segmentation categories is open-ended and new categories can be added incrementally [16]. These verbally based models use pairs of spoken utterances and single objects to train the models, and they ground based on phoneme sequences. Yu et al. have proposed a co-occurrence-based categorizing model trained by images that contain multiple objects and their spoken descriptions [17]. While these tasks look similar to our research, they are fundamentally different in that they are supervised and carried out in a way that does not reflect the true condition of early-stage spoken language learning for human beings and, thus, does not serve as proof for Skinner's idea.

E. Dupoux (2018) gives a holistic overview of the recent developments in the field of computer-simulated infant language learners [18]. However, his article focuses on providing insights rather than concrete solutions.

The rest of this paper is organized in the following order. Section 2 introduces the fundamentals of Deep Q-Learning. Section 3 introduces the unsupervised word segmentation and ES-KMeans algorithm. A designed task that demonstrates our idea is explained in Section 4. Section 5 gives a detailed description of our proposed method. Section 6 explains how the experiments are carried out. The respective results are stated and analyzed in Section 7. Finally, Section 8 concludes the paper and gives some insights into our idea.

[†]Equal contribution.

2. DEEP Q-LEARNING

The Deep Q-learning (or Deep Q-Network, DQN) method is a variant of Q-learning introduced by Mnih et al. [3] to handle the challenges in complex reinforcement learning environments. It has proven to be effective in many challenging tasks such as computer resource management [19], robotics [20] and even chemistry [21].

DQN applies two deep neural networks (DNN) to estimate the action-value function $Q(s, a; \theta)$. One is the policy Q-network Q with weights θ, which is used to decide the action; and the other one is the target network $\hat{Q}$ with weights θ^-, which is used for generating the Q-learning targets. Every C updates, the target network $\hat{Q}$ copies the weights θ from the policy network Q.

The weights θ are updated by gradient descent. The loss function $L(\theta)$ is calculated as follows:

$$y = r + \gamma max_{a'} \hat{Q}(s', a'; \theta^-), \tag{1}$$

$$L(\theta) = (y - Q(s, a; \theta))^2, \tag{2}$$

where y is the target, r is the reward of the current action a, γ is the discounting factor, and s' and a' are the expected state and action of the next step, respectively.

3. ES-KMEANS WORD SEGMENTATION

Unsupervised word segmentation aims to tackle the problem of identifying word units (word boundaries) from a relatively long utterance under a zero resource condition. Several research attempts have been made in this direction. [22] attempts to model the problem in a Bayesian way. [23] tries to sort out a solution at the syllable level. [24] proposes an embedded segmental model that embeds audio segments into fixed-length embeddings and then applies K-means algorithms on them.

Amongst several existing unsupervised word segmentation techniques, this paper uses the ES-KMeans method proposed in [24], which is one of the most sophisticated ones with low computational complexity while maintaining a relatively low unsupervised Word Error Rate (WER). We briefly discuss the general idea of the ES-KMeans in the rest of this section.

Given an audio clip consisting of frames (e.g., MFCC feature vectors): $y_{1:M} = y_1, y_2, ..., y_M$, the aim is to break this sequence down into different sub-segments of meaningful words. We first define an embedding function f_e which maps an arbitrary-length segment (e.g., $y_{t1} : y_{t2}$) into a fixed-dimension embedding x_i of $x \in R^D$:

$$f_e(y_{t1} : y_{t2}) = x_i. \tag{3}$$

There are multiple choices for this embedding function. For this paper, we use a simple down-sampling technique in the toolset from S. Bhati [23][1].

[1] https://github.com/ramesh720/recipe_zs2017_track2_phoneme

The long utterance is first randomly cut into segments q. The embeddings of the segments are then clustered by a K-means algorithm under a fixed set of segmentation boundaries q. We then fix cluster assignments z and optimize q. This optimization loop is repeated on a joint target function as follows until convergence:

$$min_z \sum_{c=1}^{K} \sum_{x \in X_c} ||x - \mu_c||^2, \tag{4}$$

where $\{\mu_{c=1}^K\}$ are the cluster means, X_c are all vectors assigned to cluster c, and element z_i in z indicates to which cluster x_i belongs. The main idea behind this approach is that acoustically similar segments should get geometrically closer in the embedding space after the clustering. The full ES-KMeans method involves more complex and rigorous proof than the general idea presented here. Detailed elaboration can be found in [24].

4. LANGUAGE ACQUISITION TASK

We designed a language acquisition task to demonstrate our idea. In this task, the agent is set to be in a 3D space and to have the motivation to be at or closer to the origin as possible. The agent has to learn by itself how to use the correct words to move efficiently. To model the task mathematically, we use 3-dimensional coordinates (x, y, z) to represent agent's position. The agent is initialized at a random position (x_0, y_0, z_0), where x_0, y_0, z_0 are integers in the range of $[-k, k]$, and the origin is set to be the final destination. The agent is given a long speech clip that contains word segments. Some of the words contained are meaningful (e.g. up, $down$, $left$, $right$, $forward$, $backward$) and the environment responds by pushing the agent along in the corresponding direction by one unit when the agent pronounces the word while the other words have no meaning (the environment does not respond to such words). First, the agent has to identify the words from the long clip and learn to choose the correct words based on its current position, thereby finding a way to the origin.

5. PROPOSED METHOD

The structure of our proposed system is represented in Fig. 1. The agent is an entity that has its own internal motivations and tries to acquire spoken language that can facilitate its communication with the external environment to achieve its desires. The agent is assumed to have zero prior knowledge of the language, like a newborn baby. It first makes an observation. In our case, this is a speech segment from the environment. (This can be interpreted as what infants hear from the surroundings when learning a language.) It then identifies the word units within this long segment and makes use of unsupervised word segmentation techniques under zero resource conditions. This part is realized by the ES-KMeans algorithm

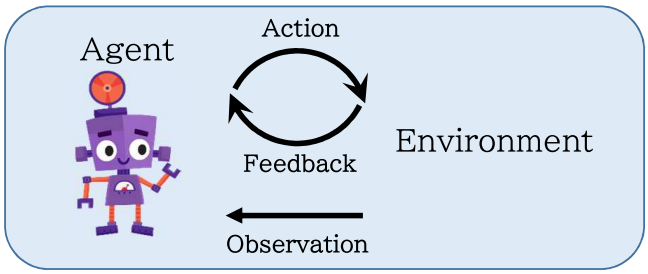


Fig. 1. Overview of the proposed system

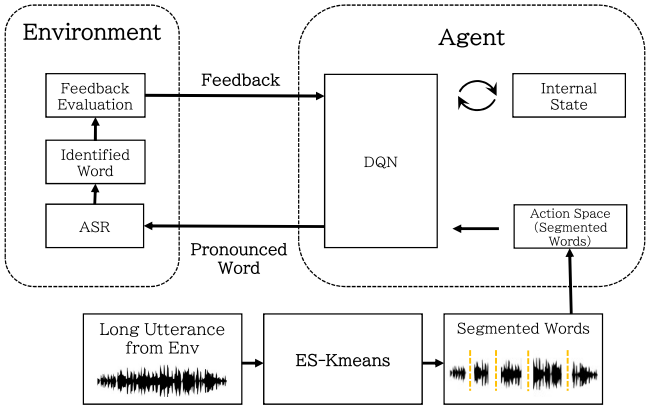


Fig. 2. System structure

proposed by [24] which has been discussed in detail in section 3.

Action refers to the utterance that the agent makes to the environment. The agent repeatedly attempts to speak to the environment. The environment then gives the agent corresponding feedback, which is picked up by the agent. The agent evaluates the current action's reward based on this feedback and its current internal states to evaluate the attempt made—whether it was a good or a bad attempt. With multiple repetitions of this loop, the agent learns how to choose what to say under which circumstance and finally acquires the ability to speak to some extent as illustrated in section 2.

This stage is achieved through the Deep Q-learning technique. The detailed system implementation is illustrated in Fig. 2. The long utterance is first fed into the ES-Kmeans algorithm, whose output represents the boundaries of identified words. The agent then takes the segmented words as action space for the Deep Q-learning algorithm to start the learning loop. An action (a word to pronounce) is first decided by the initial Deep Q-Network and output to the environment. In real use cases, the environment interacting with the agent can eventually be replaced with a collection of real humans. But for the illustration's purpose, we adopt a trained ASR (Automatic Speech Recognition) model to do the action evaluation task. The ASR model in the environment then evaluates the received waveform and attempts to recognize the pronounced word. The recognized word is then sent to the feedback evaluation algorithm to determine what kind of action must be posed back onto the agent. After receiving feedback from the environment, the agent evaluates the reward for the current iteration based on the feedback and its current internal state. The reward is then used to tune the decision-making Deep Q-Network to a better state. As this loop goes on, the agent gradually learns to make proper decisions on what to speak (in other words, the agent starts to appreciate the meaning of the words)

6. EXPERIMENT SETUP

6.1. Task Setup

The task described in section 4 is set up as follows. The k is set at 22, and the maximum number of steps taken for each round is set at 5,000. If the agent does not come back to the origin after 5,000 steps, the round fails, and the agent position is reset.

6.2. Dataset

For the initial long utterance, we used the Google Speech Commands Dataset: an English voice command dataset with 65,000 one-second-long utterances of 35 short words by thousands of different people. As explained in Section 4, there are six meaningful words: up, down, left, right, forward, backward. These make the agent move in the corresponding direction in the scenario we designed. Therefore, we pick 200 samples of each of these six words and an additional 200 samples of the word "marvin," which serves as noise in the input data. In total, there are 1,400 samples, which we then concatenate into a single wave file. To demonstrate the importance and effectiveness of unsupervised word segmentation in the system, we performed it in two forms. The first one is a random-cut baseline. For this, we cut the long utterance randomly with an average duration of approximately one word (e.g., 500–1,200ms). As it is cut randomly, broken word segments are expected to be more frequent as a result. For the second one, we pass the long utterance into the unsupervised word segmentation algorithm.

6.3. ASR Model

The ASR model used is Google Speech-to-Text API [2], which is a trained general-purpose ASR system. We aimed to make the simulation as close to a real-world situation as possible by making this choice.

6.4. Deep Q-Network

The Deep Q-learning model is designed to have an action space of around 2,000 which corresponds to the number of

[2]https://cloud.google.com/speech-to-text/docs/reference/rest/

	Recog. Meaningful Words	Accuracy
Rand. Baseline	236	19.7%
Unsup. Seg.	447	37.3%

Table 1. Performance comparison between random cut baseline and unsupervised word segmentation

word segments identified by either unsupervised or random segmentation. Out of all of these, only a portion can be considered valid actions, namely, those that can be recognized by ASR. The Q-net is trained to output higher Q values on those valid actions given the current state of the agent.

The state of the agent is set as its current position and the satisfaction level SL is set to be the minus euclidean distance between the current position of the agent and the origin. The reward function r is designed to be the change in the satisfaction level between two consecutive steps:

$$SL(t) = -(x_t^2 + y_t^2 + z_t^2), \tag{5}$$

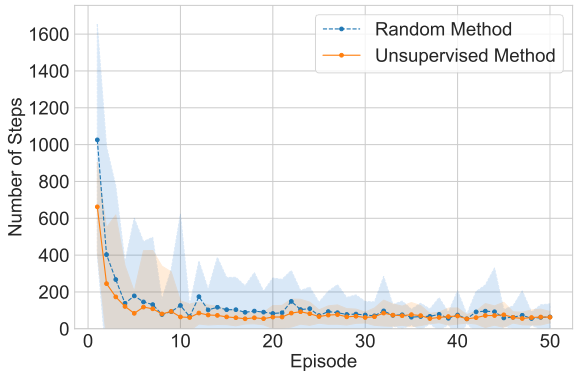
$$r(t) = SL(t) - SL(t-1), \tag{6}$$

where x_t, y_t, and z_t are the coordinates of the agent at step t.

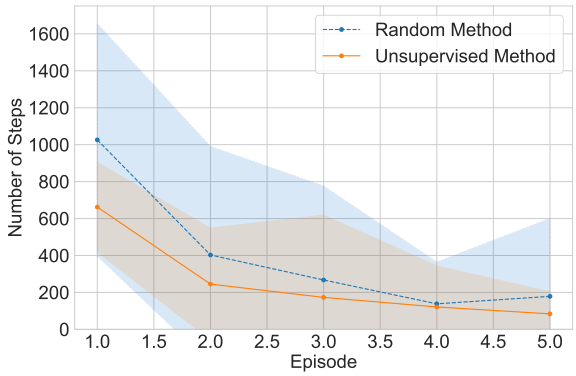
7. RESULTS AND ANALYSIS

The word recognition rate is compared in Table 1. This is calculated by the number of recognized meaningful words over the total number of actual words (i.e. 1,200). We blindly obey the results from ASR but ignore the true quality of the identified words based on the fact that the only information source for the agent is the environment (ASR in this case). We observed that the unsupervised word segmentation method is 18% more accurate than the random-cut method.

Fig. 3 shows the results of running the Deep Q-learning model on the designed task in section 4 . The vertical axis represents the steps taken by the agent to return to the origin for each episode. The results are obtained by running 50 episodes over 100 random seeds. From (a), we can safely conclude that the acquisition of spoken language is successful in both methods as the number of steps taken eventually converges. The random-cut does work as well because as long as there are meaningful candidates of valid words, the agent is able to learn to speak correctly by reinforcement learning. However, from (b) we do observe a difference in their speed of learning. Unsupervised-learning-supported reinforcement learning does excel with a 35.45% reduction in the average number of steps taken for the first episode, and this difference would be further amplified in more complex and larger-scale spoken language acquisition tasks. The unsupervised method also shows greater stability in performance, considering the standard deviation shown in the results.



(a) Result of 50 episodes



(b) Result of first 5 episodes

Fig. 3. The curves represent the means of numbers of steps taken per episode. The shadows represent the standard deviation.

8. CONCLUSION

In this paper, we simulated the process of spoken language acquisition of human beings on computers and offered strong evidence for the hypothesis that the process of acquiring spoken language is fundamentally a combination of observing the environment, processing the observation, and grounding the observed inputs with their true meaning through a series of reinforcement attempts. While the idea is conceptually straightforward, our results demonstrated the feasibility of unsupervised-learning-supported reinforcement learning in solving problems under zero resource circumstances. Our future works will include: further confirmation of the idea on larger and more realistic data sets; thinking of possible real-world scenarios where this idea can be useful; and extending our research to the ideas from [25], which gives detailed and evidenced insights on real-world infant language learning processes.

9. REFERENCES

[1] B.F Skinner, "Verbal behavior," in *Verbal behavior*. New York: Appleton-Century-Crofts, 1957.

[2] D. Silver and A. Huang, "Mastering the game of Go with deep neural networks and tree search," in *Nature*. Google DeepMind, 2016, vol. 529, pp. 484–489.

[3] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.

[4] D. Amodei, S. Ananthanarayanan, and R. Anubhai ... Z. Zhu, "Deep Speech 2 : End-to-end speech recognition in English and Mandarin," in *Proceedings of the 33rd international conference on machine learning*. 2016, ICML '08, p. 173–182, USA: PMLR.

[5] D. A. Ferrucci, "Introduction to 'This is Watson'," *IBM Journal of Research and Development*, vol. 56, no. 3, pp. 235–249, May 2012.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the ieee international conference on computer vision*. Microsoft Research, 2015, p. 1026–1034.

[7] C. Lu and X. Tang, "Surpassing human-level face verification performance on LFW with GaussianFace," Tech. Rep., The Chinese University of Hong Kong, 2014.

[8] Stevan Harnad, "The symbol grounding problem," *Physica D: Nonlinear Phenomena*, vol. 42, no. 1-3, pp. 335–346, 1990.

[9] Cynthia Matuszek, "Grounded language learning: Where robotics and NLP meet.," in *IJCAI*, 2018, pp. 5687–5691.

[10] Jeffrey Mark Siskind, "Grounding language in perception," *Artificial Intelligence Review*, vol. 8, no. 5-6, pp. 371–391, 1994.

[11] Jeffrey Mark Siskind, "Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic," *Journal of artificial intelligence research*, vol. 15, pp. 31–90, 2001.

[12] Karl Moritz Hermann, Felix Hill, Simon Green, Fumin Wang, Ryan Faulkner, Hubert Soyer, David Szepesvari, Wojciech Marian Czarnecki, Max Jaderberg, Denis Teplyashin, et al., "Grounded language learning in a simulated 3D world," *arXiv preprint arXiv:1706.06551*, 2017.

[13] Haonan Yu, Haichao Zhang, and Wei Xu, "Interactive grounded language acquisition and generalization in a 2D world," in *International Conference on Learning Representations*, 2018.

[14] Abhishek Sinha, B Akilesh, Mausoom Sarkar, and Balaji Krishnamurthy, "Attention based natural language grounding by navigating virtual environment," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, pp. 236–244.

[15] Deb Roy, "Grounded spoken language acquisition: Experiments in word learning," *IEEE Transactions on Multimedia*, vol. 5, no. 2, pp. 197–209, 2003.

[16] Aneesh Chauhan and Luís Seabra Lopes, "Using spoken words to guide open-ended category formation," *Cognitive processing*, vol. 12, no. 4, pp. 341, 2011.

[17] Chen Yu and Dana H Ballard, "On the integration of grounding language and learning objects," in *AAAI*, 2004, vol. 4, pp. 488–493.

[18] E. Dupoux, "Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner," *Cognition*, vol. 173, pp. 43–59, 2018.

[19] H. Mao, M. Alizadeh, I. Menache, and S. Kandula, "Resource management with deep reinforcement learning," Tech. Rep., Massachusetts Institute of Technology, Microsoft Research, 2016.

[20] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," Tech. Rep., University of California Berkeley, 2016.

[21] Z. Zhou, X. Li, and R. N. Zare, "Optimizing chemical reactions with deep reinforcement learning," Tech. Rep., Stanford University,, 2017.

[22] H. Kamper, A. Jansen, and S. J. Goldwater, "A segmental framework for fully-unsupervised large-vocabulary speech recognition," *Comput. Speech Lang.*, vol. 46, pp. 154–174, 2017.

[23] O. Räsänen, G. Doyle, and M. C. Frank, "Unsupervised word discovery from speech using automatic segmentation into syllable-like units," Tech. Rep., Aalto University, Stanford University, 2016.

[24] H. Kamper, K. Livescu, and S. Goldwater, "An embedded segmental K-means model for unsupervised segmentation and clustering of speech," Tech. Rep., Stellenbosch University, Toyota Technological Institute, University of Edinburgh, 2017.

[25] Anne Cutler, *Native listening: Language experience and the recognition of spoken words*, Mit Press, 2012.