Table 2: Experiment Accuracy Results of Various Algorithms on Two Complex Reasoning Tasks involving Set Output and Set of Sequences.
Mean F_1 score (mF1) and mean Edit Distance (mED) are Used.

	Metric	Multi-Label	SSG-S	SSG-RNN	SSG-CNN
Task 1	$mF1$, the higher the better	0.64	0.19	0.42	**0.70**
Task 2	mED, the lower the better	N/A	8.10	3.75	**2.00**

As one can see from Table 1, SSG substantially outperforms the simple sigmoid network for multi-label classification. Although one might use different or more complex architectures than the sigmoid network, we believe the relative improvement would be consistent (which supported in the following more complex tasks).

Synthetic Datasets

We conduct two experiments to compare the proposed methods: a number problem that predicts sets, and another problem that predicts a set of sequences. We first describe each problem, with the aim of tackling complex reasoning tasks that traditional machine learning methods cannot handle.

Task 1: Predicting Sets. In this task, the input is a positive integer read as a string of digits. Let the leading digit be m. The output is the set of m leading digits of the input string, with duplicates counted only once. For example, if $X = 33874$, then $Y = \{3, 8\}$. We call this Task-1. We again use $mF1$ as the accuracy score to compare the ground truth label set and the learned set.

Task 2: Predicting Set of Sequences. In the second task, the input is a digit string of length 20. Let the string be evenly split into two halves. The first 10 digits are grouped into five pairs: (s_1, e_1), ..., (s_5, e_5); and the last 10 digits constitute a string a. The output set consists of (at most) 5 subsequences of a: $a[s_1, e_1)$, ..., $a[s_5, e_5)$. Whenever $s_i \geq e_i$ for some i, the substring is empty and hence it does not count as an element of the output set. Similar to the first data set, duplicate strings are removed. For example, if $X = 00490000349172105519$, then $Y = \{2, 10551\}$. The elements of Y are substrings $a[4, 9)$ and $a[3, 4)$, where $a = 9172105519$. Note that 0-based indexing is used here. Treated as a multi-label classification problem, the number of classes is 10^{10}, which is impossible to handle. We call this Task-2. We use mean edit distance, mED, as the accuracy score to compare the ground truth set of sequences and the learned set of sequences. For ground truth set and learned set, we compute ED distance between every pair of sequences and divided by the total number of pairs. The lower the score mED, the better.

System Architecture: Since both tasks have sequence inputs, we use an encoder-decoder architecture (**?**). We use a one layer LSTM with 60 encoder hidden units and 120 decoder hidden units. An embedding layer of size 60 is used for appropriate discrete inputs and outputs. We use Adam optimizer (**?**) with a batch size of 15, and cross entropy as loss function. We generate 1000 samples and randomly split 70% as training and the rest as testing.

We compare three methods SSG-S, SSG-RNN, and SSG-CNN with the baseline multi-label sigmoid network for these two tasks. Table 2 shows the results. In both tasks, we can see that SSG-CNN is the best method, outperforming the second best SSG-RNN by a large margin (28% $mF1$ and 1.75 mED). Moreover, the neural-network-based SSG-CNN and -RNN outperform SSG-S, showing that it is very important to consider the complexity of reasoning tasks. Note that we did not tune or search for the best hyper-parameters and it is reasonable to assume that these performance figures can be further improved. SSG-CNN also outperforms the multi-label method on Task 1, and the multi-label method is not applicable to Task 2 due to the extreme modeling complexity.

Conclusion

We proposed a general framework, SSG, along with three variants, designed to solve set-valued output problems. We developed a sequential generation approach that can efficiently learn set relationships from data, as demonstrated on benchmark and reasoning tasks. Experiments show that the sequential generation procedure can improve performance on traditional multi-label tasks and can handle more complex sets such as set of sequences, where traditional methods are not readily applicable.

Further work will include theoretical analysis on the relationships between the set size and the learning performance, investigation on better training methods for SSG, and testing on a wider variety of set components, including sets of sets. We believe set-valued outputs have many applications such as theorem proving in AI and are foundational for systems that perform reasoning in particular, making their general treatment an important research direction to address.

Acknowledgments

We thank for colleagues at AISR for helpful discussion and anonymous reviewers for insightful comments.

Nesciunt mollitia porro, sed accusantium earum aut quaerat architecto incidunt, id consectetur labore omnis ipsa autem consequuntur cumque maxime, temporibus consequuntur eos, dolorum maxime ea modi sed adipisci a perferendis ipsum.Ducimus explicabo praesentium minus nam, tenetur iure accusamus voluptatibus expedita nisi illum cupiditate saepe mollitia voluptas laudantium, ut amet dolorem ipsa veniam voluptate sed nostrum reiciendis molestiae eaque, corporis esse consectetur ad architecto modi officiis dolor, natus officia tenetur assumenda quia consectetur ipsam dignissimos ex sit nobis magni.Deserunt ex tempora ab iste, autem delectus quos natus ipsum vel maxime

incidunt debitis sit ex libero, harum ratione quaerat perferendis id saepe vitae officiis nihil quis omnis, quod fugit mollitia qui commodi.