Cold-Start and Interpretability: Turning Regular Expressions into Trainable Recurrent Neural Networks

Chengyue Jiang[◇], **Yinggong Zhao**[†], **Shanbo Chu**[†], **Libin Shen**[†], **Kewei Tu**[◇*]
[◇] School of Information Science and Technology, ShanghaiTech University
Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences
University of Chinese Academy of Sciences
Shanghai Engineering Research Center of Intelligent Vision and Imaging
[†] Leyan Technology. Inc
{jiangchy, tukw}@shanghaitech.edu.cn
{ygzhao, chushb, libin}@leyantech.com

Abstract

Neural networks can achieve impressive performance on many natural language processing applications, but they typically need large labeled data for training and are not easily interpretable. On the other hand, symbolic rules such as regular expressions are interpretable, require no training, and often achieve decent accuracy; but rules cannot benefit from labeled data when available and hence underperform neural networks in rich-resource scenarios. In this paper, we propose a type of recurrent neural networks called FA-RNNs that combine the advantages of neural networks and regular expression rules. An FA-RNN can be converted from regular expressions and deployed in zero-shot and cold-start scenarios. It can also utilize labeled data for training to achieve improved prediction accuracy. After training, an FA-RNN often remains interpretable and can be converted back into regular expressions. We apply FA-RNNs to text classification and observe that FA-RNNs significantly outperform previous neural approaches in both zero-shot and low-resource settings and remain very competitive in rich-resource settings.

1 Introduction

Over the past several years, neural network approaches have rapidly gained popularity in natural language processing (NLP) because of their impressive performance and flexible modeling capacity. Nevertheless, symbolic rules are still an indispensable tool in various industrial NLP applications. Regular expressions (RE) are one of the most representative and useful forms of symbolic rules and are widely used for solving tasks such as pattern matching (Hosoya and Pierce, 2001; Zhang et al., 2018) and intent classification (Luo et al., 2018). RE-based systems are highly interpretable and therefore support fine-grained human inspection and manipulation. For example, individual RE rules in a system can be easily added, revised, or removed to quickly adapt the system to changes in the task specification. Moreover, RE-based systems do not require a training stage with labeled data and hence can be quickly deployed with decent performance in zero-shot scenarios. However, REs rely on human experts to write and often have high precision but moderate to low recall; RE-based systems cannot evolve by training on labeled data when available and thus usually underperform neural networks in rich-resource scenarios.

How to combine the advantages of symbolic rules and neural networks is an open question and is drawing increasing attention recently. One possible way is to use rules to constrain neural networks, usually in the manner of regularization via knowledge distillation (Hu et al., 2016) and multi-task learning (Awasthi et al., 2020; Xu et al., 2018), or by tuning the output logits of neural networks (Li and Srikumar, 2019; Luo et al., 2018). In this way, information from rules can be injected into neural networks, though the neural networks still require training and remain black boxes that are hard to interpret and manipulate. Another way of utilizing rules is to design novel neural network architectures inspired by rule systems (Schwartz et al., 2018; Graves et al., 2014; Peng et al., 2018; Lin et al., 2019). Models designed based on this idea usually achieve better interpretability, but they must be trained on labeled data and cannot be directly converted from rules or manually specified by human experts because of their structural differences from rule systems.

In this paper, we propose finite-automaton recurrent neural networks (FA-RNN), a novel type of recurrent neural networks that is designed based on the computation process of weighted finite-state automata. Because of the equivalence between

[*] Corresponding author.

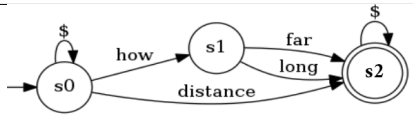
Label	[*distance*]
RE	*$*(how (far | long) | distance) $**
Matched Text	⟨BOS⟩ tell me **how far** is oakland airport from downtown ⟨EOS⟩
FA	

Table 1: RE for matching sentences asking about distance, and a matched sentence. '$' is the wildcard. '|' is the **OR** operator. '*' is the Kleene star operator. We also show the finite automaton converted from the RE. s_2 is the final state.

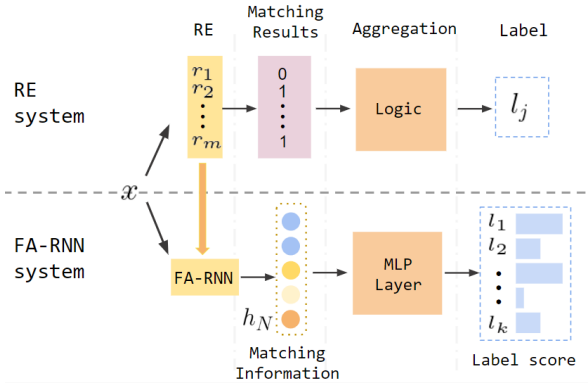


Figure 1: RE and FA-RNN systems for text classification.

REs and finite-state automata, we can convert any REs into an FA-RNN, which can be deployed in zero-shot and cold-start scenarios. When there are labeled data, the FA-RNN can also be trained in the same way as any neural network, which improves its prediction accuracy over the original REs. The FA-RNN has good interpretability. When converted from REs, it is (approximately) equivalent to the REs and is fully interpretable. Even after training, it often remains highly interpretable and can be converted back into REs. The interpretability of FA-RNNs opens the possibility of fine-grained manipulation such as integrating new REs into a trained FA-RNN and disabling old REs that are used to initialize an FA-RNN.

We apply FA-RNNs to the text classification task and compare them with neural network baselines as well as existing approaches of integrating REs and neural networks. Our experiments find that FA-RNNs show clear advantages in both zero-shot and low-resource settings and remain very competitive in rich-resource settings.

2 Background

2.1 Regular Expressions

Regular expressions (RE) are patterns usually used for searching or matching a string and are a succinct way to denote regular languages. We show a simple example RE for matching sentences[1] in Table 1.

2.2 RE System for Text Classification

The text classification task aims to assign a class label to an input sentence. Let $\boldsymbol{x} = \langle x_0, \cdots, x_N \rangle$ be a sentence and $\mathcal{L} = \{l_1, \cdots, l_k\}$ be the label set. One common and straight-forward way to use REs for classification is as follows. Firstly, write m REs $\mathcal{R} = \{r_1, \cdots, r_m\}$, where each RE corresponds to some label in $\mathcal{L}$. Then, for each sentence $\boldsymbol{x}$, apply these REs to get matching results. Finally, aggregate the matching results to produce a final label for sentence $\boldsymbol{x}$ based on a set of propositional logic rules. Each rule specifies a logical expression of matching results that implies a specific label. For example, let M_i represent whether RE r_i is matched, then we may have a rule: $(M_i \vee M_j) \wedge \neg M_k \rightarrow l_p$. The whole procedure is shown in the top half of Figure.1.

2.3 Finite-State Automaton

Finite-state automata (FA) are machines with finite numbers of states. An FA can transit from one state to another in response to an input. It has a start state s_0 and a set of final states $\mathcal{S}_\infty$. Every RE can be converted into an FA expressing the same language by Thompson's construction algorithm (Thompson, 1968). For a sequence $\boldsymbol{x} = \langle x_1, \cdots, x_N \rangle$, an RE matches the sequence if and only if the converted FA starts from s_0 and finally reaches a final state after consuming $\boldsymbol{x}$. Table 1 shows an FA converted from the example RE. Further, for every RE, there exists a *unique* FA with a minimum number of states and deterministic transitions (m-DFA) such that they express the same language (Hopcroft et al., 2001). Deterministic transitions mean that given a current state and an input, there is a unique next state. The m-DFA can be obtained by running the powerset construction algorithm (Rabin and Scott, 1959) and the DFA minimization algorithm (Hopcroft, 1971).

[1] Example taken from the ATIS intent classification dataset.

2.4 Weighted Finite-State Automaton

A weighted finite-state automaton (WFA) assigns a weight to each transition, which is formally defined as a 5-tuple: $\mathcal{A} = \langle \Sigma, \mathcal{S}, \boldsymbol{T}, \boldsymbol{\alpha}_0, \boldsymbol{\alpha}_\infty \rangle$.

- Σ: a finite input vocabulary. $|\Sigma| = V$.
- $\mathcal{S}$: a finite set of states. $|\mathcal{S}| = K$.
- $\boldsymbol{T} \in \mathbb{R}^{V \times K \times K}$: a tensor of transition weights. $\boldsymbol{T}[\sigma, i, j]$ is the weight of transiting from s_i to s_j in response to input σ. $\boldsymbol{T}[\sigma] \in \mathbb{R}^{K \times K}$ denotes the transition matrix of σ.
- $\boldsymbol{\alpha}_0 \in \mathbb{R}^K$: initial weights. $\boldsymbol{\alpha}_0[i]$ is the weight of staying at state s_i at time $t = 0$.
- $\boldsymbol{\alpha}_\infty \in \mathbb{R}^K$: final weights. $\boldsymbol{\alpha}_\infty[i]$ is the weight of staying at state s_i after reading all the inputs.

An FA can be seen as a WFA with 0/1 weights. $\boldsymbol{T}[\sigma, i, j]$ is 1 if s_i can transit to s_j in response to σ and 0 otherwise. $\boldsymbol{\alpha}_0[i] = \mathbb{1}\{s_i \in \mathcal{S}_0\}$ and $\boldsymbol{\alpha}_\infty[i] = \mathbb{1}\{s_i \in \mathcal{S}_\infty\}$, where $\mathbb{1}()$ is the indicator function and $\mathcal{S}_0$ denote the set of start states[2].

For sequence $\boldsymbol{x}$, the score of WFA $\mathcal{A}$ accepting $\boldsymbol{x}$ can be calculated using the forward (Baum and Petrie, 1966) and Viterbi (Viterbi, 1967) algorithms. Let path $\boldsymbol{p} = \langle u_1, \cdots, u_{N+1} \rangle$ be a sequence of indexes of the states that we visit when consuming $\boldsymbol{x}$. The score $\mathcal{B}(\mathcal{A}, \boldsymbol{p})$ of path $\boldsymbol{p}$ can be computed by:

$$\boldsymbol{\alpha}_0[u_1] \cdot \left(\prod_{i=1}^{N} \boldsymbol{T}[x_i, u_i, u_{i+1}] \right) \cdot \boldsymbol{\alpha}_\infty[u_{N+1}] \tag{1}$$

Let $\pi(\boldsymbol{x})$ be the set of all paths that start from start state s_0 and reach a final state $s_i \in \mathcal{S}_\infty$ after consuming sequence $\boldsymbol{x}$. The forward algorithm computes the sum of path scores.

$$\begin{aligned} \mathcal{B}_{\text{forward}}(\mathcal{A}, \boldsymbol{x}) &= \sum_{\boldsymbol{p} \in \pi(\boldsymbol{x})} \mathcal{B}(\boldsymbol{\mathcal{A}}, \boldsymbol{p}) \\ &= \boldsymbol{\alpha}_0^T \cdot \left(\prod_{i=1}^{N} \boldsymbol{T}[x_i] \right) \cdot \boldsymbol{\alpha}_\infty \end{aligned} \tag{2}$$

The Viterbi algorithm computes the maximum of path scores.

$$\mathcal{B}_{\text{Viterbi}}(\mathcal{A}, \boldsymbol{x}) = \max_{\boldsymbol{p} \in \pi(\boldsymbol{x})} \mathcal{B}(\mathcal{A}, \boldsymbol{p}) \tag{3}$$

It can be computed by replacing matrix multiplication in Eqa.2 with the max-plus operator. For an FA $\mathcal{A}$, the forward score is exactly the number of paths in $\pi(\boldsymbol{x})$ while the Viterbi score indicates whether $\pi(\boldsymbol{x})$ is non-empty.

[2]Normally, we define that an FA has only one start state, but any FA with multiple start states can be converted into an FA with one start state by adding ϵ-transitions from a new start state to all the original start states.

3 Method

We show step-by-step how we can convert REs to a novel type of recurrent neural networks called FA-RNNs.

3.1 From REs to Recurrent Neural Networks

RE to FA As mentioned in Sec.2.3, we can convert an RE into an m-DFA. In order to obtain a concise FA with better interpretability and faster computation speed, we treat the wildcard '$' as a special word in the vocabulary and run the algorithms mentioned in Sec.2.3 to obtain a "pseudo" m-DFA $\mathcal{A}$.

FA as RNN As discussed in Sec.2.4, the FA $\mathcal{A}$ can be seen as a WFA with 0/1 weights which is parameterized by $\Theta = \langle \boldsymbol{\alpha}_0, \boldsymbol{T}, \boldsymbol{\alpha}_\infty \rangle$.

The computation of the WFA forward score (Eqa.2) can be rewritten into a recurrent form. Let $\boldsymbol{h}_t \in \mathbb{R}^K$ be the forward score vector after consuming t words in $\boldsymbol{x}$. $\boldsymbol{h}_t[i]$ can be interpreted as the number of paths starting from s_0 and reaching s_i at step t.

$$\begin{aligned} \boldsymbol{h_0} &= \boldsymbol{\alpha}_0^T \\ \boldsymbol{h}_t &= \boldsymbol{h}_{t-1} \cdot \boldsymbol{T}[x_t],\ 1 \le t \le N \\ \mathcal{B}_{\text{forward}}(\mathcal{A}, \boldsymbol{x}) &= \boldsymbol{h}_N \cdot \boldsymbol{\alpha}_\infty \end{aligned} \tag{4}$$

The computation of the WFA Viterbi score can be formulated in a similar way. Therefore, we can view a WFA as a form of recurrent neural networks (RNN) parameterized by Θ.

3.2 Decomposing the Parameter Tensor

Despite the equivalence to FAs and hence better interpretability, the RNNs proposed in Sec.3.1 has much more parameters than a traditional RNN because of the tensor $\boldsymbol{T} \in \mathbb{R}^{V \times K \times K}$. To reduce the parameter number, we propose to apply tensor rank decomposition (explained in the Appendix.A) and decompose $\boldsymbol{T}$ into three matrices $\boldsymbol{E}_\mathcal{R} \in \mathbb{R}^{V \times r}, \boldsymbol{D}_1 \in \mathbb{R}^{K \times r}, \boldsymbol{D}_2 \in \mathbb{R}^{K \times r}$, where r is a hyper-parameter. Note that if r is smaller than the rank of $\boldsymbol{T}$, then the decomposition is approximate. We empirically find that, for a 100-state FA converted from RE, we can obtain a small decomposition error ($\le 1\%$) if $r \ge 100$.

Now the RNN is parameterized by $\Theta_D = \langle \boldsymbol{\alpha_0}, \boldsymbol{\alpha_\infty}, \boldsymbol{E}_\mathcal{R}, \boldsymbol{D}_1, \boldsymbol{D}_2 \rangle$. $\boldsymbol{E}_\mathcal{R}$ has a dimension associated with vocabulary size V and can be viewed as a word embedding matrix containing RE information for each word. Let $\boldsymbol{v}_t \in \mathbb{R}^r$ be the embedding of word x_t contained in $\boldsymbol{E}_\mathcal{R}$. The recurrent

update in Eqa.4 becomes:

$$\begin{aligned} \boldsymbol{a} &= (\boldsymbol{h}_{t-1} \cdot \boldsymbol{D}_1) \circ \boldsymbol{v}_t \\ \boldsymbol{h}_t &= \boldsymbol{a} \cdot \boldsymbol{D}_2^T \end{aligned} \tag{5}$$

where $\circ$ denotes element-wise product. Eqa.5 produces the same result as Eqa.4 with sufficiently large r.

Note that the size of $\boldsymbol{h_t}$ is determined by the state number K of the m-DFA converted from RE. In some cases, K may be too small, resulting in limited representational power of the RNN. A simple method to solve this problem is to concatenate $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ with a $K' \times r$ zero matrix, hence increasing the hidden state size by K'. Subsequent training (to be introduced later) would update $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ so that these added dimensions can be utilized. This is equivalent to adding K' isolated states in the FA and relying on training to establish transitions between the old and new states.

3.3 Integrating Pretrained Word Embedding

Pretrained word embeddings have been found very useful in bringing external lexical knowledge into neural networks. Let $\boldsymbol{E}_w \in \mathbb{R}^{V \times D}$ be the word embedding matrix and $\boldsymbol{u}_t \in \mathbb{R}^D$ be the word embedding of x_t in $\boldsymbol{E}_w$. We introduce another matrix $\boldsymbol{G} \in \mathbb{R}^{D \times r}$ that can transform the D-dimensional word embedding $\boldsymbol{u}_t$ into r-dimension, which can then replace $\boldsymbol{v}_t$ in the recurrent update of Eqa.5. We initialize $\boldsymbol{G}$ by setting $\boldsymbol{G} = \boldsymbol{E}_w^\dagger \boldsymbol{E}_{\mathcal{R}}$, where $\boldsymbol{E}_w^\dagger$ is the pseudo-inverse of $\boldsymbol{E}_w$. In this way, we approximate $\boldsymbol{v}_t$ with $\boldsymbol{u}_t\boldsymbol{G}$ and hence the initialized RNN still tries to mimic the FA. After training, however, the RNN will be able to utilize the additional information contained in pretrained word embeddings and hence may outperform the original FA.

In practice, we find it beneficial to interpolate the two r-dimension embeddings $\boldsymbol{v}_t$ and $\boldsymbol{u}_t\boldsymbol{G}$ with a hyper-parameter $\beta \in [0, 1]$. When β is 1, we only use RE information. When β gets closer to 0, we integrate more external lexical information into the model. The recurrent update formula becomes:

$$\begin{aligned} \boldsymbol{z}_t &= \beta \boldsymbol{v}_t + (1-\beta)\boldsymbol{u}_t\boldsymbol{G} \\ \boldsymbol{a} &= (\boldsymbol{h_{t-1}} \cdot \boldsymbol{D}_1) \circ \boldsymbol{z}_t \\ \boldsymbol{h_t} &= \boldsymbol{a} \cdot \boldsymbol{D}_2^T \end{aligned} \tag{6}$$

We name this new form of RNNs as *FA-RNNs*, i.e., recurrent neural networks built from finite-state automata.

3.4 Extensions of FA-RNN

Gated Extension (FA-GRU) Inspired by the Gated Recurrent Unit (Chung et al., 2014), we sacrifice some interpretability and add an update gate $\boldsymbol{f}_t$ and a reset gate $\boldsymbol{r}_t$ into the FA-RNN. The update gate determines how much information from the past shall be retained. The reset gate determines whether to reset the previous score vector to $\boldsymbol{h}_0$, The recurrent update is as follows.

$$\begin{aligned} \boldsymbol{z}_t &= \beta \boldsymbol{v}_t + (1-\beta)\boldsymbol{u}_t\boldsymbol{G} \\ \boldsymbol{f}_t &= \boldsymbol{\sigma}(\boldsymbol{W}_f \boldsymbol{z}_t + \boldsymbol{U}_f \boldsymbol{h}_{t-1} + \boldsymbol{b}_f) \\ \boldsymbol{r}_t &= \boldsymbol{\sigma}(\boldsymbol{W}_r \boldsymbol{z}_t + \boldsymbol{U}_r \boldsymbol{h}_{t-1} + \boldsymbol{b}_r) \\ \hat{\boldsymbol{h}}_{t-1} &= (1 - \boldsymbol{r}_t) \circ \boldsymbol{h}_0 + \boldsymbol{r_t} \circ \boldsymbol{h_{t-1}} \\ \boldsymbol{a} &= (\hat{\boldsymbol{h}}_{\boldsymbol{t-1}} \cdot \boldsymbol{D}_1) \circ \boldsymbol{z}_t \\ \hat{\boldsymbol{h}}_t &= \boldsymbol{a} \cdot \boldsymbol{D}_2^T \\ \boldsymbol{h}_t &= (1 - \boldsymbol{f}_t) \circ \boldsymbol{h}_{t-1} + \boldsymbol{f}_t \circ \hat{\boldsymbol{h}}_t \end{aligned} \tag{7}$$

$\boldsymbol{\sigma}$ is the sigmoid activation function and $\boldsymbol{W}_f$, $\boldsymbol{W}_r$, $\boldsymbol{U}_f$, $\boldsymbol{U}_r$ are additional parameters for gates. Note that when $\boldsymbol{f}_t$ and $\boldsymbol{r}_t$ is close to $\mathbf{1}$, the FA-GRU degenerates to the FA-RNN. Therefore, we initialize $\boldsymbol{b}_f, \boldsymbol{b}_r$ to a large value and $\boldsymbol{W}_f, \boldsymbol{W}_r, \boldsymbol{U}_f, \boldsymbol{U}_r$ randomly using Xavier initialization (Glorot and Bengio, 2010) to ensure that the initialized FA-GRU is approximately equivalent to the FA-RNN and hence the original REs.

Bidirectional Extension (BiFA-RNN) Our networks can be easily extended to their bidirectional variants. For any RE, we can reverse it by simply reversing its word order (e.g., "*free $* (phone | phones) $**" can be reversed to "*$* (phone | phones) $* free*") and then convert the reversed RE into WFA $\overleftarrow{\mathcal{A}}$ and the corresponding FA-RNN. Score vector $\overleftarrow{\boldsymbol{h_N}}$ can be computed by applying Eqa.6 or Eqa.7 on the reversed input sentence $\overleftarrow{\boldsymbol{x}}$. Then we take the average of $\overleftarrow{\boldsymbol{h_N}}$ and the left-to-right score vector $\overrightarrow{\boldsymbol{h_N}}$ to obtain the final score vector $\boldsymbol{h}_N = (\overrightarrow{\boldsymbol{h_N}} + \overleftarrow{\boldsymbol{h_N}})/2$.

3.5 Aggregation Layer for Text Classification

As introduced in Sec.2.2, an RE system for text classification contains multiple REs that are aggregated to form a class label prediction. Here we describe how to convert such an RE system to an FA-RNN system for text classification (the bottom half of Figure.1).

For each RE r_i in the RE system, we convert it into a WFA $\mathcal{A}_i$ with K_i states, start weights $\boldsymbol{\alpha}_{0,i}$, and final weights $\boldsymbol{\alpha}_{\infty,i}$. We can view these WFAs

Logic	Soft Logic
$\neg A$	$1 - a$
$A \vee B$	$\min(1, a + b)$
$A \wedge B$	$\max(0, a + b - 1)$

Table 2: Soft logic. A, B are proposition symbols with soft truth values a, b.

as a single WFA $\mathcal{A}$ with a total number of $K = \sum_i K_i$ states and multiple start states. We then convert this WFA to an FA-RNN. After we run this FA-RNN on sentence $\boldsymbol{x}$, the last state vector $\boldsymbol{h}_N$ contains the matching information of all the REs.

To predict a class label from $\boldsymbol{h}_N$, we create a soft aggregation layer. First, we extract the forward or Viterbi score of each RE from $\boldsymbol{h}_N$. For forward scoring, we follow Eqa.4 and have:

$$\mathcal{B}_{\text{forward}}(\mathcal{A}_i, \boldsymbol{x}) = \boldsymbol{h}_N \cdot \bar{\boldsymbol{\alpha}}_{\infty,i} \tag{8}$$

where $\bar{\boldsymbol{\alpha}}_{\infty,i}$ expands $\boldsymbol{\alpha}_{\infty,i}$ by filling zeros for states not belonging to $\mathcal{A}_i$. For Viterbi scoring, we replace matrix multiplication with max-plus. The computed score for $\mathcal{A}_i$ can be seen as a soft matching result of RE r_i. Second, we rewrite the logical RE aggregation rules introduced in Sec.2.2 to soft logic expressions (Kimmig et al., 2012; Li and Srikumar, 2019) (Table 2). Instead of predicting a single label, the soft aggregation layer outputs the label logits $\boldsymbol{l} \in \mathbb{R}^k$. When all the elements in $\boldsymbol{h}_N$ are close to either 0 or 1, the output of the soft aggregation layer is approximately equivalent to that of the RE aggregation layer of Sec.2.2.

Since the logical RE aggregation rules can be expressed in the conjunctive normal form, we can implement the corresponding soft aggregation layer with a two-layer MLP with ReLU-like activation functions. This is similar to the MLP layer commonly used at the end of traditional neural networks to map the hidden representation to label logits. In practice, we find it sometimes beneficial to not use any activation function in the MLP.

3.6 Training with Labeled Data

So far we have introduced how to initialize an FA-RNN system that is approximately equivalent to an RE classification system. When there are labeled data, the FA-RNN can also be trained to improve its performance. We simply use the output logits $\boldsymbol{l}$ to compute the cross-entropy loss on the training data and use a gradient-based method such as Adam (Kingma and Ba, 2014) to optimize it.

	#Train	#Dev	#Test	$\lvert\mathcal{L}\rvert$	$\lvert\mathcal{R}\rvert$	K	%Acc
ATIS	3982	996	893	26	27	107	87.0
	*$ * flights | flight | ((go | get | fly) from $ * to $ *) $ * → FLIGHT*						
QC	4965	500	500	6	68	94	64.4
	*$ * what $? does $+ (stand? for) $* → ABBREVIATION*						
SMS	4502	500	500	2	36	52	93.2
	$ free $ * (phone | phones) $* → SPAM*						

Table 3: Dataset statistics and example REs. $\mathcal{L}$ is the label set. $\mathcal{R}$ is the RE set. K is the state number of the converted WFA. %Acc is the classification accuracy of the RE system. We provide an example RE and its targeting label for each dataset.

	ATIS	QC	SMS
RE system	87.01	64.40	93.20
FA-RNN	86.53	61.95	93.00
FA-GRU	86.81	62.90	93.20
BiFA-RNN	88.10	62.90	93.00
BiFA-GRU	88.63	62.90	93.20
BiGRU+*i*	1.34	18.75	11.90
BiGRU+*o*	30.74	27.50	30.40
BiGRU+*io*	38.69	25.70	73.25
BiGRU+*pr*	9.94	17.70	53.00
BiGRU+*kd*	9.94	17.70	53.00
BiGRU+*i*+*u*	86.42	64.85	92.75
BiGRU+*o*+*u*	83.03	64.95	93.05
BiGRU+*io*+*u*	86.14	64.75	92.70
BiGRU+*pr*+*u*	85.67	64.60	93.5
BiGRU+*kd*+*u*	87.37	63.70	93.55

Table 4: Accuracy of zero-shot classification. The RE system and baselines trained on RE-labeled data are included for reference.

Note that we typically fix $\boldsymbol{E}_{\mathcal{R}}$ during training because we find that updating $\boldsymbol{E}_{\mathcal{R}}$ is not helpful. Therefore, the number of trainable parameters in an FA-RNN is similar to (usually smaller than) that of an RNN. We compare the number of parameters of different models in Appendix.C.

4 Experiments

We use the forward score version of FA-RNNs by default in our experiments. We use GloVe (Pennington et al., 2014) as the word embedding and keep it fixed for our methods and all the baselines. We tune the learning rate, number of additional isolated states K', and interpolation coefficient β for our methods on the development set. We provide more details of hyper-parameter tuning for FA-RNNs and all the baselines in Appendix.D.

4.1 Datasets

We evaluate the performance of our methods on three text classification datasets that have been used in previous work of integrating REs and neural networks: ATIS (Hemphill et al., 1990), Question Classification (QC) (Li and Roth, 2002) and SMS (Alberto et al., 2015). ATIS is a popular dataset consisting of queries about airline information and services. QC contains questions that can be classified into general categories like *LOCATION*, *ENTITY*, etc. SMS is a spam-classification dataset. We write REs for ATIS and use a modified version of REs from Awasthi et al. (2020) for QC and SMS. We show dataset statistics and RE examples in Table 3.

4.2 Baselines

Basic Networks We compare FA-RNN with traditional recurrent neural networks including RNN (Elman, 1990), GRU (Chung et al., 2014), LSTM (Hochreiter and Schmidhuber, 1997), and their bidirectional variants. We also experiment with a 4-layer CNN (Kim, 2014) and a 4-layer DAN (Iyyer et al., 2015), which are also frequently used in text classification. We feed the hidden representation produced by these models into an MLP to obtain the label logits and use the cross-entropy loss as the objective function. We tune the learning rates and the number of hidden states in [50, 100, 150, 200] on the development set for each dataset.

RE-enhanced Basic Networks We also compare our method with the basic neural networks enhanced by existing methods of combining rules and neural networks. Luo et al. (2018) propose three ways to utilize RE matching results in a neural model: 1) use the results as additional input features; 2) use the results to guide attention; 3) use the results to directly tune the output logits. As our basic networks do not involve attention, we enhance them using 1), 3) or both, denoted as $+i$, $+o$ and $+io$ respectively. Another method of utilizing rules is the knowledge distillation framework (Hinton et al., 2015). It treats the RE system as the teacher and its label logits as the soft targets, and distills this knowledge into the basic networks. We denote this method as $+kd$. Hu et al. (2016) combines knowledge distillation with posterior regularization by iteratively projecting the student network into the rule-regularized space. We denote this method as $+pr$. Finally, in the zero-shot setting, we also enhance these baselines by training them using unlabeled data tagged by regular expressions. We denote this enhancement by $+u$.

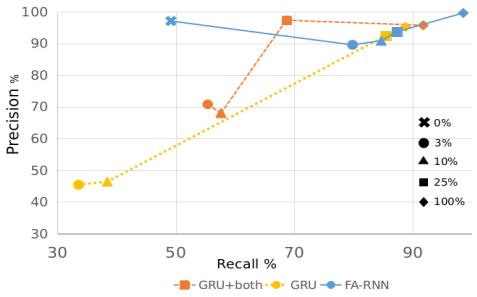


Figure 2: Precision and recall with different amounts of training data on SMS.

4.3 Zero-Shot Classification

We compare our methods with the RE system and RE-enhanced BiGRU in the zero-shot scenario, in which no training data (including the development set) is available. All the methods use or are initialized by exactly the same set of REs. For the $+u$ enhancement, we use the full training data with their labels removed as unlabeled data. We show the results in Table 4.

The results show that our methods are comparable to the RE system. The small differences in accuracy between the RE system and our methods are caused by approximation errors in decomposing the parameter tensor and integrating word embedding, as well as the introduction of gates in FA-GRUs. Our methods have much better performance than RE-enhanced BiGRUs, because RE-enhanced BiGRUs without training perform random guesses, except that in the cases of $+o$ and $+io$, RE matching results directly influence the outputs and hence improve the predictions. Baselines with the $+u$ enhancement can also match the accuracy of the RE system, but unlike our methods, they require training on sufficient RE-labeled data.

We also report the results of the other baselines in Appendix.E. Without any training, the basic networks not enhanced by RE just perform random guesses. The other RE-enhanced basic networks have similar behaviors to RE-enhanced BiGRUs.

4.4 Low-Resource and Full Training

We compare all the methods trained on 1%, 10%, and 100% of the training data. We use the original development set for the 10% and 100% experiments; but for the 1% experiment, we sample a smaller development set containing the same

	ATIS (26-class)			QC (6-class)			SMS (2-class)		
	1%	10%	100%	1%	10%	100%	1%	10%	100%
FA-RNN	90.43	90.79	96.52	**67.75**	79.6	91.3	93.1	96.75	98.8
FA-GRU	88.94	**90.85**	96.61	66.2	80.7	91.85	**94.25**	**96.8**	**99.2**
BiFA-RNN	89.31	**90.85**	96.72	57.65	81.5	91.55	91.7	96.7	99
BiFA-GRU	**90.62**	90.26	96.64	64.15	**82.8**	**92.4**	93.9	96.75	98.8
CNN	71.61	86.09	94.74	50.9	74.9	89.25	89.85	95.9	98.8
DAN	71.02	83.68	90.4	47.25	65.4	77.8	89.9	93.7	98.6
RNN	70.91	75.17	91.55	22.4	67.9	85	85.1	89.85	97.75
LSTM	69.37	78.14	95.72	40.45	75.75	90	86.2	95.75	97.85
GRU	70.72	88.52	96.3	42.35	79.75	91.2	86.15	95.55	98.05
BiRNN	70.72	79.98	93.39	49.35	75.95	87.35	86.75	94.9	97.8
BiLSTM	70.77	87.12	96.25	55.95	76.75	90.95	92.15	95.8	97.7
BiGRU	70.69	88.35	96.75	62.7	80.05	91.5	89.6	95.95	98.4
BiGRU +i	82.84	90.01	96.56	66.3	80.25	92	90.95	96.75	98.55
BiGRU +o	80.21	89.22	96.33	60.15	80.2	91.7	90.6	95.95	98.4
BiGRU +io	82.61	89.95	95.46	65.05	79.65	90.7	93.85	96.75	98.25
BiGRU +pr	72.4	88.89	96.5	61.6	80.45	91.85	90.9	96.05	98.45
BiGRU +kd	73.38	88.86	**96.75**	62.65	80.3	91.25	87.65	96	98.55

Table 5: Classification accuracy with different amounts of training data.

FA-RNN	ATIS	QC	SMS
-F	**96.52**	**91.30**	98.80
-V	95.66	88.20	97.85
-F-O	94.51	87.80	**99.20**
-F-Rand	92.16	80.60	95.40
-V-Rand	91.26	78.60	97.00
-F-Rand$\boldsymbol{E}_w$	94.17	84.40	97.00
-Train$\boldsymbol{E}_{\mathcal{R}}$	96.41	89.20	99.00

Table 6: Ablation Study. -F denotes the default method using forward scoring. -V denotes Viterbi scoring. -O denotes the undecomposed version described in Sec.3.1. -Rand denotes random initialization. -Rand$\boldsymbol{E}_w$ denotes using random word embedding. -Train$\boldsymbol{E}_{\mathcal{R}}$ denotes training $\boldsymbol{E}_{\mathcal{R}}$.

amount of data as 1% of the training data to simulate the low-resource setting. Table 5 shows the results. Because of space limit, for RE-enhanced networks, we only report the results of RE-enhanced BiGRUs, which perform the best among all the RE-enhanced networks. The complete results of all the methods with standard deviations can be found in Appendix.E.

From the results we can see that our methods outperform all the other methods in the low-resource settings, especially on 1% training data. With 100% training data, overall our methods are much better than RNN, DAN and CNN, and are either slightly better than or comparable to BiLSTM, BiGRU, and RE-enhanced BiGRUs. RE-enhanced BiGRUs are indeed better than non-enhanced BiGRU in general, and *+pr* and *+kd* seem to be more data-hungry than *+i*, *+o* and *+io*.

SMS is a binary classification task of spam detection, so we regard [*spam*] as the positive label and calculate the precisions and recalls of FA-RNN and two baselines GRU and GRU+*io* given different amounts of training data (Fig.2). With no training data, FA-RNN is almost equivalent to REs and has high precision but moderate recall; but with just 3% data, its recall is greatly improved and its precision drops only moderately; and with additional data, its precision and recall are both improved. For the baselines, the precision and recall of GRU always increase with more data, while the changes of the precision and recall of GRU+*io* seem less stable.

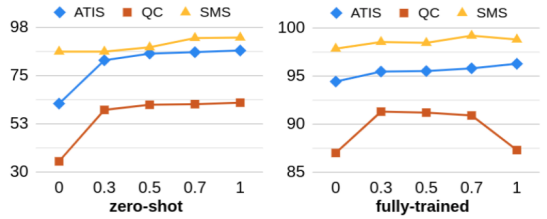


Figure 3: Performance of FA-RNN with different β.

5 Analysis

Impact of β β from Eqa.6 controls the influence of pretrained word embedding. Fig.3 shows how β impacts the performance of FA-RNN on the zero-shot and fully-trained scenarios. It can be seen that using pretrained word embedding does not help in the zero-shot scenario but can be helpful in the fully-trained scenario. One possible explanation is that word similarities encoded in the pretrained word embedding may not be compatible with a classification task, but training with data could adapt the model (in particular, by updating $\boldsymbol{G}$) to better utilize the information contained in the pretrained word embedding.

Ablation Study Table 6 shows the results of variants of FA-RNN when trained on the full datasets. From the results we can conclude that: 1) forward scoring outperforms Viterbi scoring; 2) tensor decomposition described in Sec.3.2 results in not only fewer parameters but also better overall performance; 3) RE initialization is helpful because it is much better than random initialization; 4) integrating pretrained word embedding is beneficial because the performance drops by a large margin

if using random word embedding; 5) training $\boldsymbol{E}_{\mathcal{R}}$ does not result in better performance on average, possibly because it introduces too many trainable parameters.

6 Interpretability

We regard the approximate equivalence between RE/WFA and our RE-initialized model as indication of good interpretability based on the following two reasons: 1) for people who are familiar with REs and automata, our model is interpretable once converted back into a WFA; 2) for non-experts who are unfamiliar with automata and REs, we may run a RE/WFA of a specific classification label on an input sentence and show which part of the sentence contributes to the (best) matching of the RE/WFA with the sentence, which can be easily understood by non-experts.

Note that not only can an FA-RNN be easily converted back into a RE/WFA at initialization, but the conversion can also be done after training. We can use the trained parameters of the FA-RNN $\Theta_{\text{RE}} = \left\langle \hat{\boldsymbol{E}}_{\mathcal{R}}, \hat{\boldsymbol{D}}_1, \hat{\boldsymbol{D}}_2, \hat{\boldsymbol{G}} \right\rangle$ and word embedding matrix $\boldsymbol{E}_w$ to reconstruct the WFA tensor $\boldsymbol{T}$.

$$\begin{aligned} \hat{\boldsymbol{E}}_{w\mathcal{R}} &= \beta \cdot \hat{\boldsymbol{E}}_{\mathcal{R}} + (1-\beta) \cdot \boldsymbol{E}_w \hat{\boldsymbol{G}} \\ \hat{\boldsymbol{T}}_{(1)} &= (\hat{\boldsymbol{D}}_2 \odot \hat{\boldsymbol{D}}_1) \hat{\boldsymbol{E}}_{w\mathcal{R}}^T \end{aligned} \tag{9}$$

where $\hat{\boldsymbol{T}}_{(1)}$ denotes the mode-1 unfolding of the reconstructed tensor $\hat{\boldsymbol{T}}$ and $\odot$ denotes the Khatri-Rao product. Further, we can use a thresholding function $f(x) = \mathbb{1}\{x \geq \gamma\}$ to convert the weights into $\{0, 1\}$ to recover an FA, where γ is a fixed scalar. Similarly, we can round the weights in the soft aggregation layer to reconstruct the logical aggregation layer. In this way, we can convert a trained FA-RNN back into an RE system.

In our experiments, we find that although the reconstructed RE systems underperform the corresponding trained FA-RNNs because of thresholding and rounding during reconstruction, they often outperform the original REs. The reconstructed RE systems achieve 73.6% accuracy for QC ($+9.2\%$ compared with the original REs) and 87.45% for ATIS ($+0.45\%$ compared with the original REs). For SMS, the reconstructed REs underperform the original ones (-1.2%) probably because the original REs are already good enough. We show an example in Fig.4, in which our model can be seen to learn interesting new patterns such as 'jet' and '737'. We show another example in Appendix.F.

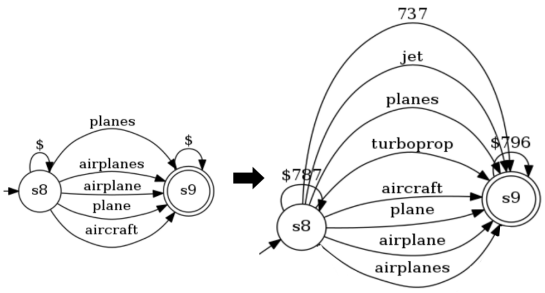


Figure 4: Part of an original RE and a reconstructed RE corresponding to label [*aircraft*]. On the right, '$787' means 787 words can activate the transition. Similar for '$796'. They are stricter than wildcard '$' that allows all possible words in the vocabulary.

Good interpretability of our models opens the possibility of fine-grained manipulation of the model, e.g., adding new REs without retraining the model. To inject a new set of REs, we convert them to a new FA-RNN with parameters $\Theta_{\text{new}} = \langle \boldsymbol{E}_{\mathcal{R}}, \boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{G} \rangle$ and merge it into the original trained FA-RNN with parameters Θ_{RE} by concatenating the parameter matrices:

$$\left\langle [\hat{\boldsymbol{E}}_{\mathcal{R}}\ \boldsymbol{E}_{\mathcal{R}}], \begin{bmatrix} \hat{\boldsymbol{D}}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D}_1 \end{bmatrix}, \begin{bmatrix} \hat{\boldsymbol{D}}_2 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D}_2 \end{bmatrix}, [\hat{\boldsymbol{G}}\ \boldsymbol{G}] \right\rangle \tag{10}$$

To add new logical aggregation rules, we can update the aggregation layer parameters similarly by concatenation. To disable an RE in an FA-RNN, we reconstruct the WFA, remove all the states of the RE from the WFA except those that can be reached from states of other REs, and finally convert the WFA back to an FA-RNN.

7 Related Work

Neural Networks Enhanced by Rules Hu et al. (2016); Li and Rush (2020) use rules to constrain neural networks by knowledge distillation and posterior regularization. Awasthi et al. (2020) inject rule knowledge into neural networks using multi-task learning. Lin et al. (2020) train a trigger matching network using additional annotation and use the output of trigger matching results as the attention of a sequence labeler. Rocktäschel et al. (2015); Xu et al. (2018); Hsu et al. (2018) use parsed rule results to regularize neural network predictions by additional loss terms. Li and Srikumar (2019); Luo et al. (2018) inject declarative knowledge in the form of parsed RE results or first-order expressions into neural networks by hacking the prediction logits or the attention scores. Hu et al. (2016); Hsu et al. (2018) use rules as additional input features.

All these previous methods use matching results or truth values of rules to enhance existing neural models. In contrast, we directly turn REs into a novel type of trainable networks.

Relating Neural Networks and WFA Schwartz et al. (2018) propose a type of neural networks for learning soft surface patterns (a subset of REs), which is inspired by WFAs but cannot be converted from WFAs or surface patterns. In contrast, our FA-RNN can be initialized from REs and converted back to REs. Peng et al. (2018); Dodge et al. (2019) formulate the update of each hidden dimension of various RNN architectures as a small WFA (2-4 states). Weiss et al. (2018); Merrill (2019) provide theoretical analysis of various neural networks and their accepting languages. Our work differs from these more theoretical studies in that we aim for a practical text classification approach. Omlin et al. (1998); Giles et al. (1999) show the equivalence between WFA and second-order RNN. The main differences between our model and theirs include the following. First, compared with the undecomposed version of our FA-RNN, their RNN model involves nonlinear activation functions which complicate the model. Second, our FA-RNN further decomposes the tensor parameter, integrate word embeddings, and propose the gated and bidirectional extensions. Third, while their work is mostly theoretical, we empirically show the usefulness of our model in text classification.

8 Conclusion and Future Work

We propose a type of recurrent neural networks called FA-RNN. It can be initialized from REs and can also learn from data, hence applicable to various scenarios including zero-shot, cold-start, low-resource and rich-resource scenarios. It is also interpretable and can be converted back into REs. Our experiments on text classification show that it outperforms previous neural approaches in both zero-shot and low-resource scenarios and is very competitive in rich-resource scenarios. In the future, we plan to apply FA-RNN to other tasks and explore other variants of FA-RNN. We release our data, RE rules and code at https://github.com/jeffchy/RE2RNN.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (61976139). We thank Dongwu Lin for his support in implementing a tool for creating automata and rules.

References

Túlio C Alberto, Johannes V Lochter, and Tiago A Almeida. 2015. Tubespam: Comment spam filtering on youtube. In *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, pages 138–143. IEEE.

Abhijeet Awasthi, Sabyasachi Ghosh, Rasna Goyal, and Sunita Sarawagi. 2020. Learning from rules generalizing labeled exemplars. *ICLR*.

Leonard E. Baum and Ted Petrie. 1966. Statistical inference for probabilistic functions of finite state markov chains. *Ann. Math. Statist.*, 37(6):1554–1563.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Jesse Dodge, Roy Schwartz, Hao Peng, and Noah A. Smith. 2019. RNN architecture learning with sparse regularization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1179–1184, Hong Kong, China. Association for Computational Linguistics.

Jeffrey L Elman. 1990. Finding structure in time. *Cognitive science*, 14(2):179–211.

C. L. Giles, C. W. Omlin, and K. K. Thornber. 1999. Equivalence in knowledge representation: automata, recurrent neural networks, and dynamical fuzzy systems. *Proceedings of the IEEE*, 87(9):1623–1640.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256.

Alex Graves, Greg Wayne, and Ivo Danihelka. 2014. Neural turing machines. *arXiv preprint arXiv:1410.5401*.

Charles T Hemphill, John J Godfrey, and George R Doddington. 1990. The atis spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

John Hopcroft. 1971. An n log n algorithm for minimizing states in a finite automaton. In *Theory of machines and computations*, pages 189–196. Elsevier.

John E Hopcroft, Rajeev Motwani, and Jeffrey D Ullman. 2001. Introduction to automata theory, languages, and computation. *Acm Sigact News*, 32(1):60–65.

Haruo Hosoya and Benjamin Pierce. 2001. Regular expression pattern matching for xml. In *Proceedings of the 28th ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*, POPL '01, page 67–80, New York, NY, USA. Association for Computing Machinery.

Wan-Ting Hsu, Chieh-Kai Lin, Ming-Ying Lee, Kerui Min, Jing Tang, and Min Sun. 2018. A unified model for extractive and abstractive summarization using inconsistency loss. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 132–141, Melbourne, Australia. Association for Computational Linguistics.

Zhiting Hu, Xuezhe Ma, Zhengzhong Liu, Eduard Hovy, and Eric Xing. 2016. Harnessing deep neural networks with logic rules. *ACL*.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1681–1691.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Angelika Kimmig, Stephen H. Bach, Matthias Broecheler, Bert Huang, and Lise Getoor. 2012. A short introduction to probabilistic soft logic. In *NIPS 2012*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Tao Li and Vivek Srikumar. 2019. Augmenting neural networks with first-order logic. *arXiv preprint arXiv:1906.06298*.

Xiang Lisa Li and Alexander Rush. 2020. Posterior control of blackbox generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2731–2743, Online. Association for Computational Linguistics.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics.

Bill Yuchen Lin, Dong-Ho Lee, Ming Shen, Ryan Moreno, Xiao Huang, Prashant Shiralkar, and Xiang Ren. 2020. TriggerNER: Learning with entity triggers as explanations for named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8503–8511, Online. Association for Computational Linguistics.

Chu-Cheng Lin, Hao Zhu, Matthew R Gormley, and Jason Eisner. 2019. Neural finite-state transducers: Beyond rational relations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 272–283.

Bingfeng Luo, Yansong Feng, Zheng Wang, Songfang Huang, Rui Yan, and Dongyan Zhao. 2018. Marrying up regular expressions with neural networks: A case study for spoken language understanding. *ACL*.

William Merrill. 2019. Sequential neural networks as automata. *CoRR*, abs/1906.01615.

Christian W Omlin, Karvel K Thornber, and C Lee Giles. 1998. Fuzzy finite-state automata can be deterministically encoded into recurrent neural networks. *IEEE Transactions on Fuzzy Systems*, 6(1):76–89.

Hao Peng, Roy Schwartz, Sam Thomson, and Noah A Smith. 2018. Rational recurrences. *EMNLP*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Michael O Rabin and Dana Scott. 1959. Finite automata and their decision problems. *IBM journal of research and development*, 3(2):114–125.

Tim Rocktäschel, Sameer Singh, and Sebastian Riedel. 2015. Injecting logical background knowledge into embeddings for relation extraction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1119–1129, Denver, Colorado. Association for Computational Linguistics.

Roy Schwartz, Sam Thomson, and Noah A Smith. 2018. Sopa: Bridging cnns, rnns, and weighted finite-state machines. *ACL*.

Ken Thompson. 1968. Programming techniques: Regular expression search algorithm. *Communications of the ACM*, 11(6):419–422.

A. Viterbi. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269.

Gail Weiss, Yoav Goldberg, and Eran Yahav. 2018. On the practical computational power of finite precision rnns for language recognition. *arXiv preprint arXiv:1805.04908*.

Jingyi Xu, Zilu Zhang, Tal Friedman, Yitao Liang, and Guy Van Den Broeck. 2018. A semantic loss function for deep learning with symbolic knowledge. *35th International Conference on Machine Learning, ICML 2018*.

Shanshan Zhang, Lihong He, Slobodan Vucetic, and Eduard Dragut. 2018. Regular expression guided entity mention mining from noisy web data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1991–2000, Brussels, Belgium. Association for Computational Linguistics.

A Tensor Rank Decomposition (CPD)

A 3-way tensor $\boldsymbol{T} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ can be approximated using r rank-1 tensors.

$$\begin{aligned} \boldsymbol{T} \approx \hat{\boldsymbol{T}} &= \sum_{i=1}^{r} \boldsymbol{a_i} \otimes \boldsymbol{b_i} \otimes \boldsymbol{c_i} \\ \hat{\boldsymbol{T}}_{(1)} &= (\boldsymbol{C} \odot \boldsymbol{B})\boldsymbol{A}^T \\ \boldsymbol{A} &= [\boldsymbol{a_1} \cdots \boldsymbol{a_r}], \boldsymbol{A} \in \mathbb{R}^{d_1 \times r}, \\ \boldsymbol{B} &= [\boldsymbol{b_1} \cdots \boldsymbol{b_r}], \boldsymbol{B} \in \mathbb{R}^{d_2 \times r}, \\ \boldsymbol{C} &= [\boldsymbol{c_1} \cdots \boldsymbol{c_r}], \boldsymbol{C} \in \mathbb{R}^{d_3 \times r}, \end{aligned} \tag{11}$$

$\hat{\boldsymbol{T}}_{(1)}$ denotes the mode-1 unfolding of tensor $\hat{\boldsymbol{T}}$. $\odot$ denotes the Khatri-Rao product while $\otimes$ denotes the outer product. When the rank of $\boldsymbol{T}$ is less than or equal to r, then the decomposition can be made exact.

B Tricks for CPD

Speeding up CPD Decomposing the WFA tensor $\boldsymbol{T}^{V \times K \times K}$ is hard when the vocabulary size V is large. However, if we neglect the wildcard '$', other words appear in RE usually form a small subset Σ' of the whole vocabulary. Denote V_1 the size of Σ', we can use a wildcard matrix $\boldsymbol{W} \in \mathbb{R}^{K \times K}$ and a much smaller tensor $\boldsymbol{T}_1^{V_1 \times K \times K}$ to represent $\boldsymbol{T}$. $\boldsymbol{W}[i, j] = 1$ when the WFA transits from s_i to s_j in respond to '$', otherwize 0. Similarly, $\boldsymbol{T}_1[\sigma, i, j] = 1$ when the WFA transits from s_i to s_j in respond to Σ'_σ, otherwize 0. By this construction, if $\Sigma_{\sigma_1} = \Sigma'_{\sigma_2}$, $\boldsymbol{T}[\boldsymbol{\sigma_1}] = \boldsymbol{W} + \boldsymbol{T}_1[\sigma_2]$.

Because Σ' is small, it is much easier to decompose T_1 to get $\boldsymbol{E}'_{\mathcal{R}} \in \mathbb{R}^{V_1 \times r}, \boldsymbol{D}'_1 \in \mathbb{R}^{K \times r}, \boldsymbol{D}'_2 \in \mathbb{R}^{K \times r}$. After obtaining these matrices, we pad the $\boldsymbol{E}'_{\mathcal{R}}$ back into a matrix $\boldsymbol{E}''_{\mathcal{R}}$ sized $V \times r$ with 0s, such that $\boldsymbol{E}''_{\mathcal{R}}[\sigma] = \boldsymbol{0}$ if $\Sigma_\sigma \notin \Sigma'$, and $\boldsymbol{E}''_{\mathcal{R}}[\sigma] = \boldsymbol{E}'_{\mathcal{R}}[\sigma_1]$ if $\Sigma_\sigma = \Sigma'_{\sigma_1} \in \Sigma'$.

Let $\boldsymbol{v}_t \in \mathbb{R}^r$ be the embedding of word x_t contained in $\boldsymbol{E}''_{\mathcal{R}}$, the recurrent update of FA-RNN now becomes:

$$\begin{aligned} \boldsymbol{a} &= (\boldsymbol{h}_{t-1} \cdot \boldsymbol{D}_1) \circ \boldsymbol{v}_t \\ \boldsymbol{h_t} &= \boldsymbol{a} \cdot \boldsymbol{D}_2^T + \boldsymbol{h}_{t-1}\boldsymbol{W} \end{aligned} \tag{12}$$

We do not train the wildcard matrix $\boldsymbol{W}$ by default. The new recurrent update will get exactly same result as the one without this trick.

Normalizing $\boldsymbol{E}_{\mathcal{R}}, \boldsymbol{D}_1$ and $\boldsymbol{D}_2$ We find normalizing $\boldsymbol{E}_{\mathcal{R}}, \boldsymbol{D}_1$ and $\boldsymbol{D}_2$ to ensure they have similar average Frobenius norm results in better performance of our methods. The average Frobenius norm is the Frobenius norm divided by the number of matrix elements. Denote a, b and c the average Frobenius norms for $\boldsymbol{E}_{\mathcal{R}}, \boldsymbol{D}_1$ and $\boldsymbol{D}_2$ respectively, and $y = (abc)^{1/3}$. We can normalize $\boldsymbol{E}_{\mathcal{R}}, \boldsymbol{D}_1$ and $\boldsymbol{D}_2$ by multiplying them with the factors $y/a, y/b$ and y/c respectively. The tensor reconstructed by the normalized matrices is the same as the tensor reconstructed by the original ones.

Model	Parameters
FA-RNN	$2Kr + Dr$
FA-GRU	$2(Kr + KK + K) + 2Kr + Dr$
RNN	$DH + HH$
GRU	$3(DH + HH + H)$
LSTM	$4(DH + HH + H)$

Table 7: Formulas of parameter numbers.

C Number of Parameters

We show calculation of model parameters of our FA-RNN and traditional recurrent neural networks in Table 7. K is the number of WFA states, r is the tensor decomposition rank, D is the word embedding dimension, and H is the hidden dimension in recurrent neural networks. In most cases, K, r are smaller than or comparable to H.

Table 8 shows the numbers of trainable parameters. The model sizes are tuned and selected using the development set. The parameters associated with the aggregation layer or the MLP layer are also included. The result shows that our methods usually have fewer trainable model parameters than baselines.

	ATIS	QC	SMS
FA-RNN	56100	57600	30900
FA-GRU	102312	113060	61500
BiFA-RNN	94200	121200	52524
BiFA-GRU	204624	244240	104236
GRU	185226	113406	112802
BiGRU	232826	181206	225602
CNN	283126	281106	451052

Table 8: Numbers of model parameters after tuning on different datasets.

D Hyper-parameters

We report the ranges of each hyper-parameters. For all methods and baselines, we select learning rates (lr) from $[0.01, 0.005, 0.001, 0.0005, 0.0001]$, For FA-RNN and its variants, we select ranks r from $[150, 200]$, additional hidden states from $[0, 30]$, and β from $[0.3, 0.5, 0.7, 1.0]$. For traditional neural networks, we select the hidden dimensions from $[50, 100, 150, 200]$. For *+i, +o, +io*, we select the RE tag dimension from $[20, 50]$, for *+pr, +kd*, we select the α from $[0.3, 0.5, 0.7]$, it is used for balancing between imitating the teacher and predicting the true hard labels. We select the best hyper-parameters for each methods based on the averaged development set accuracy.

E Full Experimental Results

Table 9, 10, 11 show the full experimental results with standard deviations. We run each model under each setting for four times with different random seeds. The standard deviations are large on low-resource scenarios because we also randomly choose the training data.

F Additional Interpretability Example

We present a more complicated example of original and reconstructed REs from the ATIS dataset in Fig.5. The trained RE contains a more sophisticated pattern with more transitions and a slightly different structure.

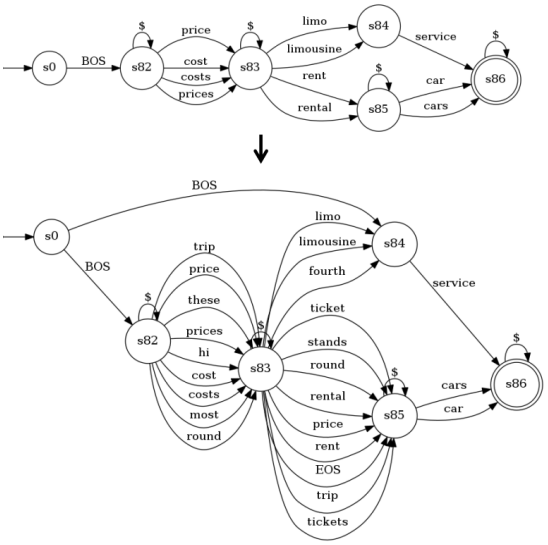


Figure 5: An untrained and trained RE corresponding to the label [*ground_fare*], which covers questions about ground service costs in airports.

	ATIS-0%		ATIS-1%		ATIS-10%		ATIS-100%	
	acc	std	acc	std	acc	std	acc	std
FA-RNN-F	86.53	0.06	90.43	0.46	90.79	0.19	96.52	0.28
BiFA-RNN-F	88.10	2.75	89.31	1.53	**90.85**	0.45	96.72	0.26
FA-GRU-F	86.81	0.14	88.94	0.94	90.85	0.54	96.61	0.31
BiFA-GRU-F	**88.63**	1.90	**90.62**	0.06	90.26	0.64	96.64	0.09
RNN+i	0.50	0.61	85.02	1.16	87.71	0.87	93.39	0.13
RNN+o	56.27	36.02	81.72	5.75	79.84	4.93	92.08	0.84
RNN+io	6.77	2.73	82.31	2.56	88.63	0.82	93.20	0.19
RNN+pr	1.60	1.48	70.41	0.73	77.02	0.92	92.25	0.58
RNN+kd	1.60	1.48	70.91	0.53	75.42	0.32	91.66	1.23
RNN	1.60	1.48	70.91	0.53	75.17	0.38	91.55	0.82
LSTM+i	1.62	2.37	84.80	2.12	88.72	1.56	96.33	0.23
LSTM+o	28.02	34.90	75.00	6.78	76.20	6.64	96.14	0.35
LSTM+io	9.29	5.75	84.71	1.33	89.19	2.12	96.53	0.55
LSTM+pr	0.53	0.66	70.77	0.00	78.16	8.58	95.94	0.33
LSTM+kd	1.40	0.72	69.37	2.80	78.86	9.42	96.02	0.37
LSTM	1.40	0.72	69.37	2.80	78.14	8.58	95.72	0.57
GRU+i	17.78	30.79	84.07	2.44	89.95	0.52	96.42	0.24
GRU+o	60.61	37.57	78.19	1.95	89.28	1.18	96.75	0.18
GRU+io	55.46	33.67	82.81	3.37	89.98	0.88	96.47	0.06
GRU+pr	0.87	0.51	70.66	0.16	89.03	1.97	96.47	0.21
GRU+kd	0.87	0.51	70.74	0.06	89.03	2.01	96.19	0.51
GRU	0.87	0.51	70.72	0.11	88.52	1.65	96.30	0.48
CNN+i	1.79	2.99	75.53	5.52	89.67	0.48	95.44	0.51
CNN+o	29.23	33.14	76.79	3.31	86.00	1.58	95.16	0.76
CNN+io	28.02	30.76	76.37	3.46	89.64	0.53	95.30	0.33
CNN+pr	0.87	0.69	72.93	1.09	86.11	1.21	94.79	0.45
CNN+kd	0.87	0.69	72.84	1.09	86.39	1.24	94.85	0.32
CNN	0.87	0.69	71.61	0.64	86.09	0.70	94.74	0.64
DAN+i	3.14	1.53	82.00	3.02	89.25	1.12	93.20	0.28
DAN+o	42.75	36.87	73.07	12.02	81.35	3.47	91.94	0.35
DAN+io	23.57	31.48	82.03	2.97	89.17	1.11	92.83	0.35
DAN+pr	0.36	0.45	68.48	3.57	83.65	1.72	90.51	0.82
DAN+kd	0.36	0.45	72.17	2.65	83.87	2.24	90.76	0.21
DAN	0.36	0.45	71.02	3.10	83.68	1.99	90.40	0.70
BiRNN+i	0.20	0.17	75.78	1.97	86.00	0.91	94.06	0.53
BiRNN+o	25.78	36.00	78.67	6.48	81.16	3.31	92.81	0.52
BiRNN+io	55.63	30.62	75.11	2.97	87.60	1.40	94.23	0.41
BiRNN+pr	14.73	18.39	70.49	0.56	83.17	1.03	92.97	0.51
BiRNN+kd	14.73	18.39	69.82	1.90	82.42	1.37	93.37	0.55
BiRNN	14.73	18.39	70.72	0.11	79.98	0.23	93.39	0.40
BiLSTM+i	1.26	1.40	84.66	1.94	89.73	1.32	96.25	0.35
BiLSTM+o	42.55	37.90	76.18	8.87	85.55	1.72	95.97	0.29
BiLSTM+io	10.92	1.91	83.40	3.87	89.11	1.32	96.72	0.19
BiLSTM+pr	0.73	0.79	70.94	0.34	87.88	1.75	96.58	0.35
BiLSTM+kd	0.73	0.79	71.84	2.13	87.49	1.95	96.25	0.53
BiLSTM	0.73	0.79	70.77	0.00	87.12	1.41	96.25	0.67
BiGRU+i	1.34	1.30	82.84	1.41	90.01	1.51	96.56	0.57
BiGRU+o	30.74	35.07	80.21	4.01	89.22	1.35	96.33	0.25
BiGRU+io	38.69	37.49	82.61	3.02	89.95	1.53	95.46	0.34
BiGRU+pr	9.94	18.54	72.40	1.39	88.89	2.26	96.50	0.19
BiGRU+kd	9.94	18.54	73.38	1.27	88.86	2.23	**96.75**	0.51
BiGRU	9.94	18.54	70.69	0.17	88.35	1.58	**96.75**	0.33

Table 9: Full results on ATIS dataset.

	QC-0%		QC-1%		QC-10%		QC-100%	
	acc	std	acc	std	acc	std	acc	std
FA-RNN-F	61.95	0.19	**67.75**	1.12	79.60	2.20	91.30	0.82
BiFA-RNN-F	62.90	0.62	57.65	6.71	**81.50**	2.38	91.55	1.50
FA-GRU-F	62.90	1.67	66.20	2.43	80.70	2.62	91.85	1.20
BiFA-GRU-F	**63.75**	1.71	64.15	1.76	82.80	1.91	**92.40**	0.52
RNN+i	17.25	6.19	61.80	1.60	73.30	1.87	87.40	2.28
RNN+o	37.45	10.46	43.80	6.65	68.30	3.92	87.40	2.18
RNN+io	26.55	2.41	60.80	3.41	72.95	3.02	88.00	0.59
RNN+pr	11.65	7.11	26.20	10.58	66.85	2.67	88.10	0.48
RNN+kd	11.65	7.11	28.80	14.00	67.00	3.88	87.15	0.89
RNN	11.65	7.11	22.40	10.93	67.90	3.66	85.00	2.21
LSTM+i	28.40	21.66	63.95	5.21	78.85	1.67	89.90	1.41
LSTM+o	33.90	7.33	48.60	6.44	76.55	2.46	90.30	0.96
LSTM+io	34.60	16.17	63.35	4.93	76.95	0.97	89.55	0.50
LSTM+pr	20.35	4.99	36.80	16.04	76.85	2.46	89.80	0.33
LSTM+kd	20.35	4.99	36.45	6.13	76.45	1.56	89.45	1.50
LSTM	20.35	4.99	40.45	4.40	75.75	2.36	90.00	0.40
GRU+i	18.65	13.55	65.20	1.12	77.45	1.81	90.55	1.00
GRU+o	41.90	19.06	40.40	6.16	79.45	2.44	90.45	1.10
GRU+io	30.15	15.72	69.35	1.65	79.65	1.84	90.70	0.68
GRU+pr	15.30	10.82	40.75	3.72	78.90	1.29	91.60	0.49
GRU+kd	15.30	10.82	41.05	4.29	79.60	2.62	90.95	0.70
GRU	15.30	10.82	42.35	1.15	79.75	1.72	91.20	1.38
CNN+i	17.25	4.73	56.35	2.03	79.20	1.23	89.55	0.75
CNN+o	44.55	6.68	56.20	10.02	75.30	3.24	90.35	1.20
CNN+io	30.15	10.31	59.50	8.79	77.80	3.63	89.55	0.38
CNN+pr	11.65	6.63	52.00	2.73	75.05	3.10	90.80	0.65
CNN+kd	11.65	6.63	50.40	3.30	73.30	3.70	89.65	1.02
CNN	15.00	4.92	50.90	4.26	74.90	3.89	89.25	0.57
DAN+i	17.20	10.03	60.10	7.60	76.15	0.53	82.20	1.77
DAN+o	31.50	8.42	43.15	5.21	66.65	2.33	81.85	0.60
DAN+io	28.05	12.42	61.90	10.79	76.70	1.65	81.80	0.71
DAN+pr	16.75	10.80	46.70	5.24	66.90	2.05	77.95	1.09
DAN+kd	16.75	10.80	49.00	4.99	67.10	1.43	77.45	0.96
DAN	16.75	10.80	47.25	4.70	65.40	2.63	77.80	1.23
BiRNN+i	19.75	7.70	56.70	2.93	75.65	1.89	88.10	1.60
BiRNN+o	29.85	23.67	51.10	9.58	73.30	3.71	87.25	0.93
BiRNN+io	34.65	10.96	58.70	5.57	75.35	2.21	87.25	0.91
BiRNN+pr	18.85	6.45	47.85	6.08	75.55	2.57	88.00	1.39
BiRNN+kd	18.85	6.45	50.75	2.82	74.95	2.97	86.25	1.02
BiRNN	18.85	6.45	49.35	7.24	75.95	3.18	87.35	1.12
BiLSTM+i	13.85	3.19	64.75	3.37	77.95	0.57	91.35	0.57
BiLSTM+o	41.40	8.17	53.75	12.84	75.45	1.60	91.65	0.34
BiLSTM+io	28.00	10.54	64.70	2.05	76.35	1.85	89.85	2.95
BiLSTM+pr	18.60	6.30	59.00	2.27	80.10	2.89	91.45	0.57
BiLSTM+kd	18.60	6.30	58.95	3.85	76.45	1.56	91.25	1.89
BiLSTM	18.60	6.30	55.95	2.67	76.75	2.70	90.95	0.81
BiGRU+i	18.75	9.86	66.30	3.96	80.25	0.85	92.00	0.63
BiGRU+o	27.50	17.40	60.15	2.18	80.20	0.78	91.70	0.77
BiGRU+io	25.70	13.07	65.05	1.22	79.65	1.84	90.70	0.68
BiGRU+pr	17.70	7.85	61.60	4.06	80.45	2.01	91.85	0.66
BiGRU+kd	17.70	7.85	62.65	3.18	80.30	2.56	91.25	0.91
BiGRU	17.70	7.85	62.70	0.95	80.05	1.65	91.50	1.57

Table 10: Full results on QC dataset.

	SMS-0%		SMS-1%		SMS-10%		SMS-100%	
	acc	std	acc	std	acc	std	acc	std
FA-RNN-F	93.00	0.00	93.10	3.19	96.75	0.82	98.80	0.19
BiFA-RNN-F	93.00	0.00	91.70	1.06	96.70	1.16	99.00	0.20
FA-GRU-F	**93.20**	0.00	**94.25**	0.90	**96.80**	0.43	**99.20**	0.26
BiFA-GRU-F	**93.20**	0.00	93.90	1.23	96.75	1.10	98.80	0.10
RNN+i	30.35	41.51	87.15	3.83	95.80	0.59	98.00	0.23
RNN+o	48.10	47.78	87.45	3.81	95.60	0.37	97.55	0.44
RNN+io	30.05	42.21	87.00	4.41	89.85	3.87	97.85	0.62
RNN+pr	67.95	35.70	85.40	0.99	95.80	0.82	98.10	0.53
RNN+kd	67.95	35.70	84.45	2.76	96.00	1.10	98.20	0.37
RNN	67.95	35.70	85.10	1.52	89.85	3.87	97.75	0.25
LSTM+i	50.00	46.23	93.85	1.30	96.60	1.07	98.50	0.12
LSTM+o	48.35	48.05	87.95	2.70	96.25	0.72	97.90	0.38
LSTM+io	88.25	3.30	93.50	0.77	96.70	0.81	98.35	0.25
LSTM+pr	49.75	41.97	86.00	0.82	96.10	1.00	98.35	0.10
LSTM+kd	49.75	41.97	86.95	2.11	96.45	0.55	98.45	0.10
LSTM	49.75	41.97	86.20	0.28	95.75	0.34	97.85	0.34
GRU+i	87.55	3.18	93.55	0.57	96.55	0.62	98.15	0.47
GRU+o	89.85	3.75	94.00	1.12	96.25	0.72	97.75	0.70
GRU+io	13.45	0.10	93.10	0.66	96.13	0.49	98.40	0.43
GRU+pr	68.05	36.30	89.80	3.49	96.25	0.30	98.40	0.28
GRU+kd	68.05	36.30	89.45	3.32	95.70	0.50	98.40	0.33
GRU	68.05	36.30	86.15	0.34	95.55	0.50	98.05	0.50
CNN+i	38.75	33.34	91.90	3.55	96.80	0.23	98.70	0.42
CNN+o	53.85	45.45	90.80	2.81	96.45	0.34	98.70	0.26
CNN+io	50.00	46.23	93.70	0.53	96.65	0.41	98.55	0.30
CNN+pr	49.70	35.94	85.75	2.07	96.00	0.33	98.35	0.50
CNN+kd	49.70	35.94	83.75	2.68	96.50	0.48	98.35	0.10
CNN	49.70	35.94	89.85	0.44	95.90	0.26	98.80	0.23
DAN+i	28.35	38.95	91.70	3.06	95.65	1.64	98.40	0.28
DAN+o	69.95	37.83	88.30	2.07	93.55	4.64	98.35	0.34
DAN+io	51.65	48.05	90.40	3.25	95.80	1.75	98.45	0.19
DAN+pr	50.00	42.26	88.50	2.20	91.80	2.72	98.65	0.30
DAN+kd	50.00	42.26	89.05	2.92	91.40	5.34	98.25	0.66
DAN	50.00	42.26	89.90	3.99	93.70	4.49	98.60	0.28
BiRNN+i	33.10	32.88	87.30	4.22	96.40	0.33	98.15	0.25
BiRNN+o	29.15	38.39	86.40	0.28	95.15	0.44	97.55	0.89
BiRNN+io	48.90	47.42	88.70	3.12	95.95	0.47	98.00	0.43
BiRNN+pr	65.70	35.02	84.95	2.90	95.20	0.28	98.30	0.35
BiRNN+kd	65.70	35.02	86.10	1.00	95.40	0.57	98.10	0.12
BiRNN	65.70	35.02	86.75	0.75	94.90	0.20	97.80	0.28
BiLSTM+i	61.55	35.24	93.60	0.98	96.75	0.25	98.30	0.38
BiLSTM+o	51.65	44.25	90.70	3.18	95.70	0.42	97.90	0.74
BiLSTM+io	71.60	43.20	92.75	1.86	96.70	0.53	98.35	0.34
BiLSTM+pr	44.40	33.39	87.10	1.99	95.65	0.34	98.35	0.10
BiLSTM+kd	44.40	33.39	86.50	5.01	96.05	0.74	98.50	0.38
BiLSTM	44.40	33.39	92.15	2.32	95.80	0.52	97.70	0.26
BiGRU+i	11.90	1.72	90.95	3.29	96.75	0.10	98.55	0.19
BiGRU+o	30.40	41.83	90.60	4.63	95.95	0.55	98.40	0.16
BiGRU+io	73.25	39.90	93.85	1.73	96.75	0.50	98.25	0.57
BiGRU+pr	53.00	30.96	90.90	2.95	96.05	0.87	98.45	0.19
BiGRU+kd	53.00	30.96	87.65	0.50	96.00	0.67	98.55	0.38
BiGRU	53.00	30.96	89.60	2.42	95.95	0.62	98.40	0.59

Table 11: Full results on SMS dataset.