Models		MSRA	AS	PKU	CTB	CKIP	CITYU	NCC	SXU	Avg.
Single-Criterion Learning										
LSTM	P	95.13	93.66	93.96	95.36	91.85	94.01	91.45	95.02	93.81
	R	95.55	94.71	92.65	85.52	93.34	94.00	92.22	95.05	92.88
	F	95.34	94.18	93.30	95.44	92.59	94.00	91.83	95.04	93.97
	OOV	63.60	69.83	66.34	76.34	68.67	65.48	56.28	69.46	67.00
Bi-LSTMs (Chen et al. 2017)	P	95.70	93.64	93.67	95.19	92.44	94.00	91.86	95.11	93.95
	R	95.99	94.77	92.93	95.42	93.69	94.15	92.47	95.23	94.33
	F	95.84	94.20	93.30	95.30	93.06	94.07	92.17	95.17	94.14
	OOV	66.28	70.07	66.09	76.47	72.12	65.79	59.11	71.27	68.40
Stacked Bi-LSTM (Chen et al. 2017)	P	95.69	93.89	94.10	95.20	92.40	94.13	91.81	94.99	94.03
	R	95.81	94.54	92.66	95.40	93.39	93.99	92.62	95.37	94.22
	F	95.75	94.22	93.37	95.30	92.89	94.06	92.21	95.18	94.12
	OOV	65.55	71.50	67.92	75.44	70.50	66.35	57.39	69.69	68.04
Switch-LSTMs This Work	P	96.07	93.83	95.92	97.13	92.02	93.69	91.81	95.02	94.44
	R	96.86	95.21	95.56	97.05	93.76	93.73	92.43	96.13	95.09
	F	96.46	94.51	95.74	97.09	92.88	93.71	92.12	95.57	94.76
	OOV	69.90	77.80	72.70	81.80	71.60	59.80	55.50	67.30	69.55
Multi-Criteria Learning										
Bi-LSTMs	P	94.64	93.54	93.24	92.87	93.26	91.41	89.30	92.61	92.61
	R	93.20	94.06	91.94	91.75	93.41	90.64	88.04	92.42	91.93
	F	93.91	93.80	92.59	92.31	93.33	91.02	88.66	92.51	92.27
	OOV	65.60	89.20	64.90	75.40	80.00	74.80	64.00	68.50	72.80
Multi-Task Framework (Chen et al. 2017)	P	95.76	93.99	94.95	95.85	93.50	95.56	92.17	96.10	94.74
	R	95.89	95.07	93.48	96.11	94.58	95.62	92.96	96.13	94.98
	F	95.82	94.53	94.21	95.98	94.04	95.59	92.57	96.12	94.86
	OOV	70.72	72.59	73.12	81.21	76.56	82.14	60.83	77.56	74.34
Switch-LSTMs This Work	P	97.69	94.42	96.24	97.09	94.53	95.85	94.07	96.88	95.85
	R	97.87	96.03	96.05	97.43	95.45	96.59	94.17	97.62	96.40
	F	**97.78**	**95.22**	**96.15**	**97.26**	**94.99**	**96.22**	**94.12**	**97.25**	**96.12**
	OOV	64.20	77.33	69.88	83.89	77.69	73.58	69.76	78.69	74.38

Table 3: Results of the proposed model on the test sets of eight CWS datasets. There are two blocks. The first block consists of single-criterion learning models. LSTM and Bi-LSTMs are baselines and the results on them are reported in Chen et al. (2017). The second block consists of the multi-criteria learning model. Multi-task framework for multi-criterion Chinese word segmentation is proposed by Chen et al. (2017). Here, P, R, F, OOV indicate the precision, recall, F value and OOV recall rate respectively. The maximum F values are highlighted for each dataset.

(single-criterion learning). By concatenating all datasets, Bi-LSTMs performs poorly in multi-criteria learning scenario (the worst). Experimental results show that Switch-LSTMs outperform both Bi-LSTMs and multi-task learning framework on all the corpora. In average, Switch-LSTMs boost about +1% (96.12 in F-value) compared to multi-task learning framework (94.86 in F-value), and boosts +3.85% compared to Bi-LSTMs model (92.27 in F-value).

We could also observe that the performance benefits from multi-criteria learning, since, in this case, the model could learn extra helpful information from other corpora. Concretely, in average F-value, Switch-LSTMs for multi-criteria learning boosts +1.36% (96.12 in F-value) compared to Switch-LSTMs for single-criterion learning (94.76 in F-value).

Model Selection

Figure 4 shows the relationship between switch number and performance in the multi-criteria learning scenario. As we can see, models with more than 2 switches are better than 1-switch-LSTM with a considerable margin, and the case with 4-way switches is slightly better than other settings. So we employ 4-way Switch-LSTMs for the following experiments. 1-way Switch-LSTMs are the traditional LSTM. So, LSTM could be viewed as a special case of the proposed Switch-LSTMs.

Scale of Parameter Set

Table 4 gives the results of multi-task framework and Switch-LSTMs on the test sets of eight datasets for multi-criteria learning. For 8 datasets, the multi-task framework contains 8 private Bi-LSTMs and 1 shared Bi-LSTMs, whereas Switch-LSTMs do not have any private parameters, consisting of K LSTM cells associated with one switch for control. As we could observe, the parameter set size of multi-task framework is 25K, while the parameter set size of Switch-LSTMs ranges from 4K to 36K with respect to various number of switches. However, as mentioned, Switch-LSTMs perform great when we have more than 2 switches. Concretely, 2-way Switch-LSTMs obtain 95.53 on F-value averagely, outperforming the multi-task framework (94.86 on F-value). But the parameter set size of 2-way Switch-LSTMs is only 7K. Therefore, Switch-LSTMs could outperform multi-task framework with fewer parameters.