Models	MSRA	AS	PKU	CTB	CKIP	CITYU	NCC	SXU	Avg.
w/o state s_i	95.68	95.02	95.71	**97.60**	94.73	94.84	92.34	96.58	95.44
w/o input embeddings e_x	96.28	94.82	95.98	97.18	94.91	95.14	**93.46**	96.54	95.54
w/o task embedding e_m	93.36	93.49	92.90	92.65	93.81	91.61	89.25	93.02	92.51
w/o e_x & e_m	93.26	93.50	92.60	93.31	94.23	91.74	88.74	93.10	92.56
w/o s_i & e_m	93.39	91.97	91.66	92.51	92.09	91.22	88.43	92.33	91.70
Randomly switch in training	92.64	92.90	92.84	93.51	94.20	91.63	89.45	93.05	92.53
Randomly switch in testing	85.56	93.63	90.16	97.16	93.53	93.22	85.56	91.49	91.29
All (Full model)	**96.71**	**95.03**	**96.22**	97.55	**95.57**	**95.38**	93.37	**96.91**	**95.84**

Table 5: Results of the proposed model with different switch configurations on test sets of eight CWS datasets for multi-criteria learning. Only F-values are reported. Term "w/o" denotes "without". The maximum F values are highlighted for each dataset.

# of instances	Models	MSRA	AS	PKU	CTB	CKIP	CITYU	NCC	SXU	Avg.
100	Single	75.71	68.51	78.87	79.65	65.59	77.18	71.75	81.29	74.82
	Transfer	89.08	91.93	91.50	92.08	94.02	93.62	88.97	93.34	91.82
300	Single	81.50	72.77	83.95	88.46	73.76	82.12	78.20	85.97	80.84
	Transfer	88.99	92.05	91.97	91.92	94.52	93.43	89.94	94.03	92.11
500	Single	83.52	75.29	86.02	90.33	75.12	86.43	79.89	87.12	82.97
	Transfer	89.17	92.23	91.90	94.94	94.61	93.72	89.98	**94.19**	92.59
700	Single	83.55	76.29	87.70	91.59	77.76	87.97	81.72	88.33	75.25
	Transfer	**89.41**	92.10	91.74	**95.49**	94.49	**93.87**	90.12	93.67	92.61
1000	Single	85.75	78.78	88.75	91.82	78.51	88.70	82.52	90.09	76.30
	Transfer	89.04	**92.35**	**92.51**	95.39	**94.64**	93.82	**90.45**	93.85	**92.76**

Table 6: Results of transfer learning of the proposed model on test sets of eight CWS datasets for single-criteria learning. Only F-values are reported. "# of instances" denotes how many instances involved for training. Single model is the conventional Bi-LSTM model. "Transfer" denotes the proposed Switch-LSTMs by fixing all the parameters learned from other 7 datasets except the new involved task embedding.

plies that a normal switch is crucial to our model, and switch mechanism contributes a lot in boosting performance.

Knowledge Transfer

Switch-LSTMs could also be easily transferred to other new datasets. To evaluate the transfer capability of our model, we leave one dataset out and train Switch-LSTMs on other 7 datasets. Then, we fixed all the parameters except the newly introduced task embedding when training on instances of the leave out dataset. As shown in Table 6, Switch-LSTMs could obtain excellent performance when only 100, 300. 500, 700, 1000 training instances are available. The single model (conventional LSTM) cannot learn from such few instances. For instance, when we train with 1000 training examples, the single model only obtains 76.30 in average F-value, whereas Switch-LSTMs could obtain 92.76 (boosts 16.46 averagely). It shows that Switch-LSTMs could adapt to a new criterion by only learning a new task embedding, and the newly learned task embedding leads to a new switch strategy for the new criterion.

Related Work

It is a common practice for utilizing annotated data from different but similar domain to boost the performance for each task. Many efforts have been made to better utilizing the homogeneous factor in various tasks to help improve multiple tasks especially those barren tasks with few examples.

Recently, some efforts have been made to transfer knowledge between NLP Tasks. Zoph and Knight; Johnson et al. (2016; 2017) have been jointly training translation models from and to different languages, it is achieved simply by jointly train encoder or both encoder and decoder. (Jiang, Huang, and Liu 2009; Sun and Wan 2012; Qiu, Zhao, and Huang 2013; Li et al. 2015; 2016; Chen, Zhang, and Liu 2016) adopted the stack-based model to take advantage of annotated data from multiple sources, and show that tasks can indeed help improve each other.

Chen, Zhang, and Liu (2016) adopted two neural models based on stacking framework and multi-view framework respectively, which boosts POS-tagging performance by utilizing corpora in heterogeneous annotations. Chen et al. (2017) have proposed a multi-criteria learning framework for CWS. Using a similar framework as in Caruana (1997), there are private layers for each task to extract criteria-specific features, and a shared layer for the purpose of transferring information between tasks, to avoid negative transfer, they pose an adversarial loss on the shared layer to impose source indistinguishability thus make it criteria-invariant.

Conclusions

In this paper, we propose a flexible model, called Switch-LSTMs, for multi-criteria CWS, which can improve the performance of every single criterion by fully exploiting the underlying shared sub-criteria across multiple heterogeneous