TASK	POS TAGGING/DEPENDENCY PARSING						SENTIMENT ANALYSIS			
DOMAIN	**A**	**EM**	**N**	**R**	**WB**	**WSJ**	**B**	**D**	**K**	**E**
LABELED	3.5K	4.9K	2.4K	3.8K	2.0K	3.0K	2K	2K	2K	2K
UNLABELED	27K	1,194K	1,000K	1,965K	525K	30K	4.5K	3.6K	5.7K	5.9K

Table 1: Statistics of all datasets used in our experiments, with the number presenting labeled or unlabeled samples in each domain. The domain abbreviations in different tasks are explained as follows. A:Answer, EM:Email, N:News, R:Reviews, WB:Weblogs, WSJ:Wall Street Journal, and B:Book, D:DVD, K:Kitchen, E:Electronics.

3 Experiment

To evaluate our approach, we conduct experiments on three representative NLP tasks: POS tagging, dependency parsing, and sentiment analysis. Details about the experiments are described as follows.

3.1 Datasets

Two popular datasets are used in our experiments. For POS tagging and dependency parsing, we use the dataset from the SANCL 2012 shared task (Petrov and McDonald, 2012), with six different domains. For sentiment analysis, we use the product review dataset from (Blitzer et al., 2007b), with four domains. Note that for all datasets, there exists both labeled and unlabeled samples in each domain. The statistics and the domains for the aforementioned datasets are reported in Table 1.

3.2 Settings

A major difference between our approach and other data selection methods is that the threshold (number of instances to be selected), n, is not fixed in our approach. Instead, it chooses the most effective ones automatically. For fair comparison, we record the resulted n from our approach in different tasks and use it in other methods to guide their selection. In all experiments, we use a multi-source domain setting where the source domain includes all labeled data from the dataset except that for the target domain, i.e., we take turns selecting a domain as the target domain, and use the union of the rest as the source domain. The number of bags, N, is set separately for each dataset to ensure a uniform bag size of 1K samples. For the guidance set, we follow Ruder and Plank (2017) and randomly select half of the instances from all the test data in the target domain discarding their labels.

Consider that the starting reward needs to be calculated from a reliable feature extractor, we adopt a "soft starting" before the regular training, were we pre-train the predictor on all source data for 2 epochs, then initialize parameters of SDG with Gaussian variables. Afterwards the predictor and SDG follow ordinary learning paradigm in each training epoch. In all experiments, we use Adam (Kingma and Ba, 2014) as the optimizer, and set γ to 0.99 following Fan et al. (2017) and n_J to 3.

	A	**EM**	**N**	**R**	**WB**	**WSJ**
JS-E	93.16	93.77	94.29	93.32	94.92	94.08
JS-D	92.25	93.43	93.54	92.84	94.45	93.32
T-S	93.59	94.65	94.76	93.92	95.32	94.44
TO-S	93.36	94.65	94.43	94.65	94.03	94.22
T+TO-S	94.33	92.55	93.96	93.94	94.51	94.98
T-S+D	93.64	94.21	93.57	93.86	95.33	93.84
TO-S+D	94.02	94.33	94.62	94.19	94.93	94.67
RANDOM	92.76	93.43	93.75	92.62	93.53	92.68
ALL	95.16	95.90	95.90	95.03	95.79	95.64
SDG (JS)	95.37	95.45	96.23	95.64	96.19	95.74
SDG (MMD)	**95.75**	96.23	96.40	95.51	**96.95**	96.12
SDG (RÉNYI)	95.52	**96.31**	**96.62**	**95.97**	96.75	**96.35**
SDG (LOSS)	95.46	95.77	95.92	95.50	96.03	95.82

Table 2: POS tagging results (accuracy %).

3.3 POS tagging

The Predictor We use the Bi-LSTM tagger proposed in Plank et al. (2016) as the predictor.

Baselines Following Ruder and Plank (2017), we compare our approach to five baselines: 1) **JS-E**: top instances selected according to Jensen-Shannon divergence. 2) **JS-D**: top instances selected from the most similar source domain, where the similarity between domains are determined by Jensen-Shannon divergence. 3) Bayesian optimization (Brochu et al., 2010) with the following settings: **T-S**, term distribution similarity; **TO-S**, topic distribution similarity; **T+TO-S**, joint term and topic distribution similarity; **T-S+D**, term distribution similarity and diversity; **TO-S+D**, topic distribution similarity and diversity. 5) **RANDOM**: a random selection model that selects the same number of instances with the n given by our approach. 6) **ALL**: The predictor is trained on all source data.

Results POS tagging results are reported in Table 2. Overall, our approach with different distri-