Training	Testing	Cross-Domain								
		CRF			CRF+R			L-CRF		
		$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}_1$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}_1$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}_1$
−Computer	Computer	86.6	51.4	64.5	23.2	90.4	37.0	82.2	62.7	71.1
−Camera	Camera	84.3	48.3	61.4	21.8	86.8	34.9	81.9	60.6	69.6
−Router	Router	86.3	48.3	61.9	24.8	92.6	39.2	82.8	60.8	70.1
−Phone	Phone	72.5	50.6	59.6	20.8	81.2	33.1	70.1	59.5	64.4
−Speaker	Speaker	87.3	60.6	71.6	22.4	91.2	35.9	84.5	71.5	77.4
−DVDplayer	DVDplayer	72.7	63.2	67.6	16.4	90.7	27.7	69.7	71.5	70.6
−Mp3player	Mp3player	87.5	49.4	63.2	20.6	91.9	33.7	84.1	60.7	70.5
	Average	82.5	53.1	64.3	21.4	89.3	34.5	79.3	63.9	70.5
	In-Domain									
−Computer	−Computer	84.0	71.4	77.2	23.2	93.9	37.3	81.6	75.8	78.6
−Camera	−Camera	83.7	70.3	76.4	20.8	93.7	34.1	80.7	75.4	77.9
−Router	−Router	85.3	71.8	78.0	22.8	93.9	36.8	82.6	76.2	79.3
−Phone	−Phone	85.0	71.1	77.5	25.1	93.7	39.6	82.9	74.7	78.6
−Speaker	−Speaker	83.8	70.3	76.5	20.1	94.3	33.2	80.1	75.8	77.9
−DVDplayer	−DVDplayer	85.0	72.2	78.1	20.9	94.2	34.3	81.6	76.7	79.1
−Mp3player	−Mp3player	83.2	72.6	77.5	20.4	94.5	33.5	79.8	77.7	78.7
	Average	84.3	71.4	77.3	21.9	94.0	35.5	81.3	76.0	78.6

Table 3: Aspect extraction results in precision, recall and F_1 score: Cross-Domain and In-Domain (−X means all except domain X)

5.3 Experiment Setting

To compare the systems using the same training and test data, for each dataset we use 200 sentences for training and 200 sentences for testing to avoid bias towards any dataset or domain because we will combine multiple domain datasets for CRF training. We conducted both cross-domain and in-domain tests. Our problem setting is cross-domain. In-domain is used for completeness. In both cases, we assume that extraction has been done for the 50 domains.

Cross-domain experiments: We combine 6 labeled domain datasets for training (1200 sentences) and test on the 7th domain (not used in training). This gives us 7 *cross-domain* results. This set of tests is particularly interesting as it is desirable to have the trained model used in cross-domain situations to save manual labeling effort.

In-domain experiments: We train and test on the same 6 domains (1200 sentences for training and 1200 sentences for testing). This also gives us 7 *in-domain* results.

Evaluating Measures: We use the popular precision $\mathcal{P}$, recall $\mathcal{R}$, and $\mathcal{F}_1$-score.

5.4 Results and Analysis

All the experiment results are given in Table 3.

Cross-domain: Each −X in column 1 means that domain X is not used in training. X in column 2 means that domain X is used in testing. We can see that L-CRF is markedly better than CRF and CRF+R in $\mathcal{F}_1$. CRF+R is very poor due to poor precisions, which shows treating the reliable aspects set K as a dictionary isn't a good idea.

In-domain: −X in training and test columns means that the other 6 domains are used in both training and testing (thus in-domain). We again see that L-CRF is consistently better than CRF and CRF+R in $\mathcal{F}_1$. The amount of gain is smaller. This is expected because most aspects appeared in training probably also appear in the test data as they are reviews from the same 6 products.

6 Conclusion

This paper proposed a lifelong learning method to enable CRF to leverage the knowledge gained from extraction results of previous domains (unlabeled) to improve its extraction. Experimental results showed the effectiveness of L-CRF. The current approach does not change the CRF model itself. In our future work, we plan to modify CRF so that it can consider previous extraction results as well as the knowledge in previous CRF models.

Acknowledgments

This work was supported in part by grants from National Science Foundation (NSF) under grant no. IIS-1407927 and IIS-1650900.