

performs better when it has to deal with: 1) surface phenomena such as “*spelling changes*”, “*punctuation changes*”, and “*change of order*”; 2) dictionary related phenomena such as “*opposite polarity substitution (habitual)*”, “*converse substitution*”, and “*modal verb changes*”. **S3** performs worse when facing phenomena such as “*negation switching*”, “*ellipsis*”, “*opposite polarity substitution (contextual)*”, and “*addition/deletion*”.

**Table 6.3** Performance profiles of all systems

Phenomenon	Paraphrase Identification Systems										
	Supervised					Unsupervised					S11
	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	
OVERALL	.74	.75	.76	.76	.70	.68	.70	.72	.75	.73	.84
Inflectional	.77	.76	.79	.79	.75	.79	.75	.76	.78	.80	.84
Modal verb	.84	.89	.90	.89	.91	.92	.89	.84	.81	.89	.92
Derivational	.80	.83	.72	.73	.84	.80	.88	.86	.80	.77	.87
Spelling	.85	.83	.88	.90	.89	.85	.89	.88	.85	.89	.94
Same pol. (hab.)	.74	.77	.78	.76	.76	.76	.76	.75	.76	.76	.85
Same pol. (con.)	.74	.74	.75	.74	.70	.71	.71	.71	.73	.73	.81
Same pol. (NE)	.74	.72	.73	.75	.64	.67	.65	.70	.73	.66	.80
Change Format	.80	.79	.75	.84	.85	.82	.81	.80	.80	.71	.91
Opp. pol. (hab.)	1.0	1.0	1.0	.50	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Opp. pol. (con.)	.77	.84	.68	.84	.52	.84	.61	.77	.65	.52	.71
Synth./analytic	.73	.73	.77	.77	.74	.70	.72	.71	.73	.74	.83
Converse sub.	.93	.93	.92	.86	.93	.86	.79	.79	.93	.79	.86
Diathesis altern.	.77	.85	.83	.77	.83	.89	.85	.83	.84	.81	.85
Negation switc	1.0	.67	.33	.33	.33	.67	.33	.67	.33	.67	.33
Ellipsis	.77	.71	.64	.74	.80	.65	.81	.74	.61	.71	.81
Coordination	.92	.92	.77	.92	.77	.92	.85	.85	.92	.92	.92
Subord. & Nest.	.83	.84	.86	.84	.81	.81	.85	.86	.80	.85	.93
Punctuation	.88	.90	.87	.87	.86	.87	.89	.89	.89	.88	.93
Direct/indirect	.84	.84	.76	.80	.76	.80	.80	.84	.80	.80	.92
Syntax/Disc.	.80	.83	.83	.81	.78	.81	.80	.80	.76	.78	.82
Add./Del.	.69	.68	.70	.72	.67	.64	.65	.66	.70	.67	.82
Change of order	.82	.83	.81	.81	.77	.82	.82	.82	.83	.84	.89
Contains neg.	.78	.74	.78	.79	.78	.72	.74	.78	.75	.76	.85
Semantic (Inf.)	.80	.89	.80	.81	.88	.90	.90	.92	.76	.79	.90
Identity	.74	.75	.77	.77	.73	.72	.73	.73	.76	.74	.85
Non-Paraphrase	.76	.77	.81	.75	.71	.55	.67	.68	.77	.79	.88
Entailment	.80	.80	.76	.76	.88	.80	.84	.88	.92	.88	.76

### 6.5.3 Comparing Performance Profiles

Table 6.3 shows the full performance profiles of all systems. The systems are identified by their IDs, as shown in Table 6.1. In addition to providing a better error analysis for every individual system, the “*performance profiles*” of the different systems can be used to compare them qualitatively. This comparison is much more informative than the “*overall performance*” comparison shown in Table 6.1. Using the “*performance profile*”, we can quickly compare the strong and weak sides of the different systems.

When looking at the “*overall performance*”, we already pointed out that **S3** [Wang et al., 2016] and **S4** [He and Lin, 2016] have almost identical quantitative results: 0.76 accuracy, 0.833 F1 for **S3** against 0.76 accuracy, 0.827 F1 for **S4**. However, when we compare their “*phenomena performance*” it is evident that, while these systems make approximately the same number of correct and incorrect predictions, the actual predictions and errors can vary.

Looking at the accuracy, we can see that **S3** performs better on phenomena such as “*Converse substitution*”, “*Diathesis alternation*”, and “*Non-Paraphrase*”, while **S4** performs better on “*Change of format*”, “*Opposite polarity substitution (contextual)*”, and “*Ellipsis*”.

We performed McNemar paired test comparing the errors of the two systems for each phenomena. Table 6.4 shows some of the more interesting results. Four of the phenomena with largest difference in accuracy show significant difference with  $p < 0.1$ . These differences in performance are substantial, considering that the two systems have nearly identical quantitative performance.

**Table 6.4** Difference in phenomena performance between S3 [Wang et al., 2016] and S4 [He and Lin, 2016]

Phenomenon	#3	#4	p
Format	.75	.84	.09
Opp. Pol. Sub (con.)	.68	.84	.06
Ellipsis	.64	.74	.08
Non-Paraphrase	.81	.75	.07

We performed the same test on systems with a larger quantitative difference. Table 6.5 shows the comparison between **S3** and **S5** [Lan and Xu, 2018b]. Ten of the phenomena show significant difference with  $p < 0.1$  and seven with  $p < 0.05$ . These results answer our **RQ 3**: we show that there are significant differences between the “*performance profiles*” of the different systems.

**Table 6.5** Difference in phenomena performance: S3 [Wang et al., 2016] and S5 [Lan and Xu, 2018b]

Phenomenon	#3	#5	p
Derivational	.72	.84	.03
Same Pol. Sub (con.)	.75	.70	.02
Same Pol. Sub (NE)	.73	.64	.01
Format	.75	.85	.03
Opp. Pol. Sub (con.)	.68	.52	.10
Ellipsis	.64	.80	.10
Addition/Deletion	.70	.67	.02
Identity	.77	.73	.01
Non-Paraphrase	.81	.71	.01
Entailment	.76	.88	.08

#### 6.5.4 Comparing Performance by Phenomena

The “phenomena performance” of the individual systems clearly differ among them, but they also show noticeable tendencies. Looking at the performance by phenomena, it is evident that certain phenomena consistently obtain lower than average accuracy across multiple systems while other phenomena consistently obtain higher than average accuracy.

In order to quantify these observations and to confirm that there is a statistical significance we performed Friedman-Nemenyi test [Demšar, 2006]. For each system, we ranked the performance by phenomena from 1 to 27, accounting for ties. We calculated the significance of the difference in ranking between the phenomena using the Friedman test [Friedman, 1940] and obtained a Chi-Square value of 198, which rejects the null hypothesis with  $p < 0.01$ . Once we had checked for the non-randomness of our results, we computed the Nemenyi test [Nemenyi, 1963] to find out which phenomena were significantly different. In our case, we compute the two-tailed Nemenyi test for  $k = 27$  phenomena and  $N = 11$  systems. The Critical Difference (CD) for these values is 12.5 at  $p < 0.05$ .

Figure 6.1 shows the Nemenyi test with the CD value. Each phenomenon is plotted with its average rank across the 11 evaluated systems. The horizontal lines connect phenomena which rank is within CD of each other. Phenomena which are not connected by a horizontal line have significantly different ranking. We can observe that each phenomenon is significantly different from at least half of the other phenomena.

We can observe that some phenomena, such as “opposite polarity substitution (habitual)”, “punctuation changes”, “spelling”, “modal verb changes”, and