

[S9] InferSent sentence embeddings [Conneau et al., 2017]

[S10] Skip-Thought sentence embeddings [Kiros et al., 2015]

In the unsupervised setup we first represent each of the two sentences under the corresponding model. Then we obtain a feature vector by concatenating the absolute distance and the element-wise multiplication of the two representations. The feature vector is then fed into a logistic regression classifier to predict the textual relation. This setup has been used in multiple PI papers, more recently by Aldarmaki and Diab [2018]. While the vector representations of BERT are unsupervised, they are fine-tuned on the dataset. Therefore we put them in a separate category (System #11).

6.5 Results

6.5.1 Overall Performance

Table 6.1 Overall Performance of the Evaluated Systems

ID	System Description	Acc	F1
SUPERVISED SYSTEMS			
1	MTE features (baseline)	.74	.819
2	He et al. [2015]	.75	.826
3	Wang et al. [2016]	.76	.833
4	He and Lin [2016]	.76	.827
5	Lan and Xu [2018b]	.70	.800
UNSUPERVISED SYSTEMS			
6	Bag-of-Words (baseline)	.68	.790
7	Word2Vec (average)	.70	.805
8	GLOVE (average)	.72	.808
9	InferSent	.75	.826
10	Skip-Thought	.73	.816
11	Google BERT	.84	.889

Table 6.1 shows the “*overall performance*” of the systems on the 1725 text pairs in the test set. Looking at the table, we can observe several regularities. First, the deep systems outperform the baselines. Second, the baselines that we choose are competitive and obtain high results. Since both baselines make their predictions based on lexical similarity and overlap, we can conclude that the dataset is

biased towards those phenomena. Third, the supervised systems generally outperform the unsupervised ones, but without running a full grid-search the difference is relatively small. And finally, we can identify the best performing systems: **S3** [Wang et al., 2016] for the supervised and **S9** [Conneau et al., 2017] for the unsupervised. BERT largely outperforms all other systems.

The “*overall performance*” provides a good overview of the task and allows for a quantitative comparison of the different systems. However, it also has several limitations. It does not provide much insight into the workings of the systems and does not facilitate error analysis. In order to study and improve the performance of a system, a developer has to look at every correct and incorrect predictions and search for custom defined patterns. The “*overall performance*” is also not very informative for a comparison between the systems. For example **S3** [Wang et al., 2016] and **S4** [He and Lin, 2016] obtain the same Accuracy score and only differ by 0.06 F1 score. With only looking at the quantitative evaluation it is unclear which of these systems would generalize better on a new dataset.

6.5.2 Full Performance Profile

Table 6.2 shows the full “*performance profile*” of **S3** [Wang et al., 2016], the supervised system that performed best in terms of “*overall performance*”. Table 6.2 shows a large variation of the performance of **S3** on the different phenomena. The accuracy ranges from .33 to 1.0. We also report the statistical significance of the difference between the correct and incorrect predictions for each phenomena and the correct and incorrect predictions for the full test set, using the Mann–Whitney U-test³ [Mann and Whitney, 1947].

Ten of the phenomena show significant difference from the overall performance at $p < 0.1$. Note that eight of them are also significant at $p < 0.05$. The statistical significance of “*Opposite polarity substitution (habitual)*”, and “*Negation Switching*” cannot be verified due to the relatively low frequency of the phenomena in the test set.

The demonstrated variance in phenomena performance and its statistical significance address **RQ 1**: we show that the performance of a PI system on each candidate-paraphrase pair depends on the different phenomena involved in that pair or at least there is a strong observable relation between the performance and the phenomena.

The individual “*performance profile*” also addresses **RQ 2**. The profile is humanly interpretable, and we can clearly see how the system performs on various sub-tasks at different linguistic levels. The qualitative evaluation shows that **S3**

³The Mann–Whitney U-test is a non-parametric equivalence of T-test. The U-Test does not assume normal distribution of the data and is better suited for small samples.

Table 6.2 Performance profile of Wang et al. [2016]

OVERALL PERFORMANCE		
Overall Accuracy	.76	
Overall F1	.833	
PHENOMENA PERFORMANCE		
Phenomenon	Acc	p
Morphology-based changes		
Inflectional changes	.79	.21
Modal verb changes	.90	.01
Derivational changes	.72	.22
Lexicon-based changes		
Spelling changes	.88	.01
Same polarity sub. (habitual)	.78	.18
Same polarity sub. (contextual)	.75	.37
Same polarity sub. (named ent.)	.73	.14
Change of format	.75	.44
Lexico-syntactic based changes		
Opp. polarity sub. (habitual)	1.0	na
Opp. polarity sub. (context.)	.68	.14
Synthetic/analytic substitution	.77	.39
Converse substitution	.92	.07
Syntax-based changes		
Diathesis alternation	.83	.12
Negation switching	.33	na
Ellipsis	.64	.07
Coordination changes	.77	.47
Subordination and nesting	.86	.01
Discourse-based changes		
Punctuation changes	.87	.01
Direct/indirect style	.76	.5
Syntax/discourse structure	.83	.05
Other changes		
Addition/Deletion	.70	.05
Change of order	.81	.04
Contains negation	.78	.32
Semantic (General Inferences)	.80	.21
Extremes		
Identity	.77	.29
Non-Paraphrase	.81	.04
Entailment	.76	.5