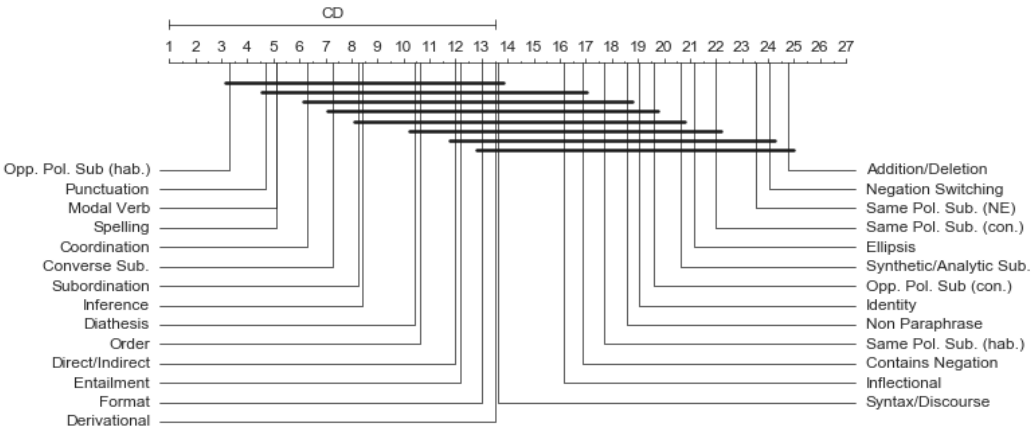

Figure 6.1: Critical Difference diagram of the average ranks by phenomena

"coordination changes" are statistically much easier according to our evaluation, as they are consistently among the best performing phenomena across all systems. Other phenomena, such as *"negation switching", "addition/deletion", "same polarity substitution (named entity)", "opposite polarity substitution (contextual)",* and *"ellipsis"* are statistically much harder, as they are consistently among the worst performing phenomena across all systems. With the exception of *"negation switching"* and *"opposite polarity substitution (habitual)"*, these phenomena occur in the corpus with sufficient frequency. These results answer our **RQ 4**: we show that there are phenomena which are easier or harder for the majority of the evaluated systems.

6.6 Discussion

In Section 6.3.2 we described our evaluation methodology and posed four research questions. The experiments that we performed and the analysis of the results answered all four of them. We briefly discuss the implications of the findings.

By addressing **RQ 1**, we showed that the performance of a system can differ significantly based on the phenomena involved in each candidate-paraphrase pair. By addressing **RQ 4**, we showed that some phenomena are consistently easier or harder across the majority of the systems. These findings empirically prove the complexity of paraphrasing and the task of PI. The results justify the distinction between the qualitatively different linguistic phenomena involved in paraphrasing and demonstrate that framing PI as a binary classification problem is an oversimplification.

By addressing **RQ 2**, we showed that each system has strong and weak sides, which can be identified and interpreted via its *"performance profile"*. This information can be very valuable when analyzing the errors made by the system or when reusing it on another task. Given the Deep architecture of the systems, such a detailed interpretation is hard to obtain via other means and metrics. By addressing **RQ 3**, we showed that two systems can differ significantly in their performance on candidate-paraphrase pairs involving particular phenomenon. These differences can be seen even in systems that have almost identical quantitative (Acc and F1) performance on the full test set. These findings justify the need for a qualitative evaluation framework for PI. The traditional binary evaluation metrics do not account for the difference in phenomena performance. They do not provide enough information for the analysis or for the comparison of different PI systems. Our proposed framework shows promising results.

Our findings demonstrate the limitations of the traditional PI task definition and datasets and the way PI systems are typically interpreted and evaluated. We show the advantages of a qualitative evaluation framework and emphasize the need to further research and improve the PI task. The *"performance profile"* also enables the direct empirical comparison of related phenomena such as *"same polarity substitution (habitual)"* and *"(contextual)"* or *"contains negation"* and *"negation switching"*. These comparisons, however, fall outside of the scope of this paper.

Our evaluation framework is not specific to the ETPC corpus or the typology behind it. The framework can be applied to other corpora and tasks, provided they have a similar format. While ETPC is the largest corpus annotated with paraphrase types to date, it has its limitations as some interesting paraphrase types (ex.: *"negation switching"*) do not appear with a sufficient frequency. We release the code for the creation and analysis of the *"performance profile"* [4].

6.7 Conclusions and Future Work

We present a new methodology for evaluation, interpretation, and comparison of different Paraphrase Identification systems. The methodology only requires at evaluation time a corpus annotated with detailed semantic relations. The training corpus does not need any additional annotation. The evaluation also does not require any additional effort from the systems' developers. Our methodology has clear advantages over using simple quantitative measures (Accuracy and F1 Score): 1) It allows for a better interpretation and error analysis on the individual systems; 2) It allows for a better qualitative comparison between the different

[4] https://github.com/JavierBJ/paraphrase_eval

systems; and 3) It identifies phenomena which are easy/hard to solve for multiple systems and may require further research.

We demonstrate the methodology by evaluating and comparing several of the state-of-the-art systems in PI. The results show that there is a statistically significant relationship between the phenomena involved in each candidate-paraphrase pair and the performance of the different systems. We show the strong and weak sides of each system using human-interpretable categories and we also identify phenomena which are statistically easier or harder across all systems.

As a future work, we intend to study phenomena that are hard for the majority of the systems and proposing ways to improve the performance on those phenomena. We also plan to apply the evaluation methodology to more tasks and systems that require a detailed semantic evaluation, and further test it with transfer learning experiments.

Acknowledgements

We would like to thank Darina Gold, Tobias Horsmann, Michael Wojatzki, and Torsten Zesch for their support and suggestions, and the anonymous reviewers for their feedback and comments.

This work has been funded by Spanish Ministery of Science, Innovation, and Universities Project PGC2018-096212-B-C33, by the CLiC research group (2017 SGR 341), and by the APIF grant of the first author.