A Qualitative Evaluation Framework for Paraphrase Identification

Venelin Kovatchev[1,3], M. Antònia Martí[1,3], Maria Salamó[2,3], Javier Beltran[1,3]
[1]Facultat de Filología, Universitat de Barcelona
[2] Facultat de Matemàtiques i Informàtica, Universitat de Barcelona
[3]Universitat de Barcelona Institute of Complex Systems
Gran Vía de les Corts Catalanes, 585, 08007 Barcelona, Spain
{vkovatchev, amarti, maria.salamo, javier.beltran}@ub.edu

Abstract

In this paper, we present a new approach for the evaluation, error analysis, and interpretation of supervised and unsupervised Paraphrase Identification (PI) systems. Our evaluation framework makes use of a PI corpus annotated with linguistic phenomena to provide a better understanding and interpretation of the performance of various PI systems. Our approach allows for a qualitative evaluation and comparison of the PI models using human interpretable categories. It does not require modification of the training objective of the systems and does not place additional burden on the developers. We replicate several popular supervised and unsupervised PI systems. Using our evaluation framework we show that: 1) Each system performs differently with respect to a set of linguistic phenomena and makes qualitatively different kinds of errors; 2) Some linguistic phenomena are more challenging than others across all systems.

1 Introduction

In this paper we propose a new approach to evaluation, error analysis and interpretation in the task of Paraphrase Identification (PI). Typically, PI is defined as comparing two texts of arbitrary size in order to determine whether they have approximately the same meaning (Dolan et al., 2004). The two texts in 1a and 1b are considered paraphrases, while the two texts at 2a and 2b are non-paraphrases.[1] In 1a and 1b there is a change in the wording (*"magistrate" - "judge"*) and the syntactic structure (*"was ordered" - "ordered"*) but the meaning of the sentences is unchanged. In 2a and 2b there are significant differences in the quantities (*"5%" - "4.7%"* and *"$27.45" - "$27.54"*).

1a A federal magistrate in Fort Lauderdale ordered him held without bail.

1b He was ordered held without bail Wednesday by a federal judge in Fort Lauderdale, Fla.

2a Microsoft fell **5 percent** before the open to **$27.45** from Thursday's close of $28.91.

2b Shares in Microsoft slipped **4.7 percent** in after-hours trade to **$27.54** from a Nasdaq close of $28.91.

The task of PI can be framed as a binary classification problem. The performance of the different PI systems is reported using the Accuracy and F1 score measures. However this form of evaluation does not facilitate the interpretation and error analysis of the participating systems. Given the Deep Learning nature of most of the state-of-the-art systems and the complexity of the PI task, we argue that better means for evaluation, interpretation, and error analysis are needed. We propose a new evaluation methodology to address this gap in the field. We demonstrate our methodology on the ETPC corpus (Kovatchev et al., 2018a) - a recently published corpus, annotated with detailed linguistic phenomena involved in paraphrasing.

We replicate several popular state-of-the-art Supervised and Unsupervised PI Systems and demonstrate the advantages of our evaluation methodology by analyzing and comparing their performance. We show that while the systems obtain similar quantitative results (Accuracy and F1), they perform differently with respect to a set of human interpretable linguistic categories and make qualitatively different kinds of errors. We also show that some of the categories are more challenging than others across all evaluated systems.

2 Related Work

The systems that compete on PI range from using hand-crafted features and Machine Learning algorithms (Fernando and Stevenson, 2008; Madnani

[1]Examples are from the MRPC corpus (Dolan et al., 2004)

Proceedings of Recent Advances in Natural Language Processing, pages 568–577,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/978-954-452-056-4_067

et al., 2012; Ji and Eisenstein, 2013) to end-to-end Deep Learning models (He et al., 2015; He and Lin, 2016; Wang et al., 2016; Lan and Xu, 2018a; Kiros et al., 2015; Conneau et al., 2017). The PI systems are typically divided in two groups: Supervised PI systems and Unsupervised PI systems.

"Supervised PI systems" (He et al., 2015; He and Lin, 2016; Wang et al., 2016; Lan and Xu, 2018a) are explicitly trained for the PI task on a PI corpora. "Unsupervised PI systems" in the PI field is a term used for systems that use a general purpose sentence representations such as Mikolov et al. (2013); Pennington et al. (2014); Kiros et al. (2015); Conneau et al. (2017). To predict the paraphrasing relation, they can compare the sentence representations of the candidate paraphrases directly (ex.: cosine of the angle), and use a PI corpus to learn a threshold. Alternatively they can use the representations as features in a classifier.

The complexity of paraphrasing has been emphasized by many researchers (Bhagat and Hovy, 2013; Vila et al., 2014; Benikova and Zesch, 2017). Similar observations have been made for Textual Entailment (Sammons et al., 2010; Cabrio and Magnini, 2014). Gold et al. (2019) study the interactions between paraphrasing and entailment.

Despite the complexity of the phenomena, the popular PI corpora (Dolan et al., 2004; Ganitkevitch et al., 2013; Iyer et al., 2017; Lan et al., 2017) are annotated in a binary manner. In part it is due to lack of annotation tools capable of fine-grained annotation of relations. WARP-Text (Kovatchev et al., 2018b) fills this gap in the NLP toolbox.

The simplified corpus format poses a problem with respect to the quality of the PI task and the ways it can be evaluated. The vast majority of the state-of-the-art systems in PI provide no or very little error analysis. This makes it difficult to interpret the actual capabilities of a system and its applicability to other corpora and tasks.

Some researchers have approached the problem of non-interpretability by evaluating the same architecture on multiple datasets and multiple tasks. Lan and Xu (2018b) apply this approach to Supervised PI systems, while Aldarmaki and Diab (2018) use it for evaluating Unsupervised PI systems and general sentence representation models.

Linzen et al. (2016) demonstrate how by modifying the task definition and the evaluation the capabilities of a Deep Learning system can be determined implicitly. The main advantage of such an approach is that it only requires modification and additional annotation of the corpus. It does not place any additional burden on the developers of the systems and can be applied to multiple systems without additional cost.

We follow a similar line of research and propose a new evaluation that uses ETPC (Kovatchev et al., 2018a): a PI corpus with a multi-layer annotation of various linguistic phenomena. Our methodology uses the corpus annotation to provide much more feedback to the competing systems and to evaluate and compare them qualitatively.

3 Qualitative Evaluation Framework

3.1 The ETPC Corpus

ETPC (Kovatchev et al., 2018a) is a re-annotated version of the MRPC corpus. It contains 5,801 text pairs. Each text pair in ETPC has two separate layers of annotation. The first layer contains the traditional binary label (paraphrase or non-paraphrase) of every text pair. The second layer contains the annotation of 27 *"atomic"* linguistic phenomena involved in paraphrasing, according to the authors of the corpus. All phenomena are linguistically motivated and humanly interpretable.

3a A federal **magistrate** in Fort Lauderdale ordered him held without bail.

3b He was ordered held without bail Wednesday by a federal **judge** in Fort Lauderdale, Fla.

We illustrate the annotation with examples 3a and 3b. At the binary level, this pair is annotated as "paraphrases". At the "atomic" level, ETPC contains the annotation of multiple phenomena, such as the *"same polarity substitution (habitual)"* of "magistrate" and "judge" (marked **bold**) or the *"diathesis alternation"* of *"...ordered him held"* and *"he was ordered by..."* (marked underline).

For the full set of phenomena, the linguistic reasoning behind them, their frequency in the corpus, real examples from the pairs, and the annotation guidelines, please refer to Kovatchev et al. (2018a).

3.2 Evaluation Methodology

We use the corpus to evaluate the capabilities of the different PI systems implicitly. That means, the training objective of the systems remains unchanged: they are required to correctly predict

the value of the binary label at the first annotation layer. However, when we analyze and evaluate the performance of the systems, we make use of both the binary and the atomic annotation layers. Our evaluation framework is created to address our main research question (RQ 1):

RQ 1 Does the performance of a PI system on each candidate-paraphrase pair depend on the different phenomena involved in that pair?

We evaluate the performance of the systems in terms of their *"overall performance"* (Accuracy and F1) and *"phenomena performance"*.

"Phenomena performance" is a novelty of our approach and allows for qualitative analysis and comparison. To calculate *"phenomena performance"*, we create 27 subsets of the test set, one for each linguistic phenomenon. Each of the subsets consists of all text pairs that contain the corresponding phenomenon[2]. Then, we use each of the 27 subsets as a test set and we calculate the binary classification Accuracy (paraphrase or non-paraphrase) for each subset. This score indicates how well the system performs in cases that include one specific phenomenon. We compare the performance of the different phenomena and also compare them with the *"overall performance"*.

Prior to running the experiments we verified that: 1) the relative distribution of the phenomena in paraphrases and in non-paraphrases is very similar; and 2) there is no significant correlation (Pearson r <0.1) between the distributions of the individual phenomena. These findings show that the sub-tasks are non-trivial: 1) the binary labels of the pairs cannot be directly inferred by the presence or absence of phenomena; and 2) the different subsets of the test set are relatively independent and the performance on them cannot be trivially reduced to overlap and phenomena co-occurrence.

The *"overall performance"* and *"phenomena performance"* of a system compose its *"performance profile"*. With it we aim to address the rest of our research questions (RQs):

RQ 2 Which are the strong and weak sides of each individual system?

RQ 3 Are there any significant differences between the *"performance profiles"* of the systems?

RQ 4 Are there phenomena on which all systems perform well (or poorly)?

4 PI Systems

To demonstrate the advantages of our evaluation framework, we have replicated several popular Supervised and Unsupervised PI systems. We have selected the systems based on three criteria: popularity, architecture, and performance. The systems that we chose are popular and widely used not only in PI, but also in other tasks. The systems use a wide variety of different ML architectures and/or different features. Finally, the systems obtain comparable quantitative results on the PI task. They have also been reported to obtain good results on the MRPC corpus which is the same size as ETPC. The choice of system allows us to best demonstrate the limitations of the classical quantitative evaluation and the advantages of the proposed qualitative evaluation.

To ensure comparability, all systems have been trained and evaluated on the same computer and the same corpus. We have used the configurations recommended in the original papers where available. During the replication we did not do a full grid-search as we want to replicate and thereby contribute to generalizable research and systems. As such, the quantitative results that we obtain may differ from the performance reported in the original papers, especially for the Supervised systems. However, the results are sufficient for the objective of this paper: to demonstrate the advantages of the proposed evaluation framework.

We compare the performance of five Supervised and five Unsupervised systems on the PI task, including one Supervised and one Unsupervised baseline systems. We also include Google BERT (Devlin et al., 2018) for reference.

The **Supervised PI systems** include:

[S1] Machine translation evaluation metrics as hand-crafted features in a Random Forest classifier. Similar to Madnani et al. (2012) *(baseline)*

[S2] A replication of the convolutional network similarity model of He et al. (2015)

[S3] A replication of the lexical composition and decomposition system of Wang et al. (2016)

[2] i.e. The "diathesis alternation" subset contains all pairs that contain the "diathesis alternation" phenomenon (such as the example pair 3a–3b). Some of the pairs can also contain multiple phenomena: the example pair 3a–3b contains both *"same polarity substitution (habitual)"* and *"diathesis alternation"*. Therefore pair 3a–3b will be added both to the *"same polarity substitution (habitual)"* and to the *"diathesis alternation"* phenomena subsets. Consequentially, the sum of all subsets exceeds the size of the test set.

[S4] A replication of the pairwise word interaction modeling with deep neural network system by He and Lin (2016)

[S5] A character level neural network model by Lan and Xu (2018a)

The **Unsupervised PI systems** include:

[S6] A binary Bag-of-Word sentence representation (baseline)

[S7] Average over sentence of pre-trained Word2Vec word embeddings (Mikolov et al., 2013)

[S8] Average over sentence of pre-trained Glove word embeddings (Pennington et al., 2014)

[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).

5 Results

5.1 Overall Performance

Table 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.

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 (2018a)	.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 1: Overall Performance of the Evaluated Systems

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.

5.2 Full Performance Profile

Table 2 shows the full *"performance profile"* of **S3** (Wang et al., 2016), the supervised system that performed best in terms of *"overall performance"*. Table 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-test[3] (Mann and Whitney, 1947).

Ten of the phenomena show significant difference from the overall performance at $p < 0.1$. Note

[3]The 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.

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

Table 2: Performance profile of Wang et al. (2016)

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** 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"*.

5.3 Comparing Performance Profiles

Table 3 shows the full performance profiles of all systems. The systems are identified by their IDs, as shown in Table 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 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

PHENOMENON	PARAPHRASE IDENTIFICATION SYSTEMS									
	SUPERVISED					UNSUPERVISED				
	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11
OVERALL ACC.	.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. sub. (hab.)	.74	.77	.78	.76	.76	.76	.76	.75	.76	.76	.85
Same pol. sub. (con.)	.74	.74	.75	.74	.70	.71	.71	.71	.73	.73	.81
Same pol. sub. (NE)	.74	.72	.73	.75	.64	.67	.65	.70	.73	.66	.80
Change of format	.80	.79	.75	.84	.85	.82	.81	.80	.80	.71	.91
Opp. pol. sub. (hab.)	1.0	1.0	1.0	.50	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Opp. pol. sub. (con.)	.77	.84	.68	.84	.52	.84	.61	.77	.65	.52	.71
Synthetic/analytic sub.	.73	.73	.77	.77	.74	.70	.72	.71	.73	.74	.83
Converse substitution	.93	.93	.92	.86	.93	.86	.79	.79	.93	.79	.86
Diathesis alternation	.77	.85	.83	.77	.83	.89	.85	.83	.84	.81	.85
Negation switching	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. & nesting	.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 style	.84	.84	.76	.80	.76	.80	.80	.84	.80	.80	.92
Syntax/disc. struct.	.80	.83	.83	.81	.78	.81	.80	.80	.76	.78	.82
Addition/Deletion	.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 negation	.78	.74	.78	.79	.78	.72	.74	.78	.75	.76	.85
Semantic (Inferences)	.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

Table 3: Performance profiles of all systems

the errors of the two systems for each phenomena. Table 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.

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

Table 4: Difference in phenomena performance between S3 (Wang et al., 2016) and S4 (He and Lin, 2016)

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

Table 5: Difference in phenomena performance: S3 (Wang et al., 2016) and S5 (Lan and Xu, 2018a)

We performed the same test on systems with

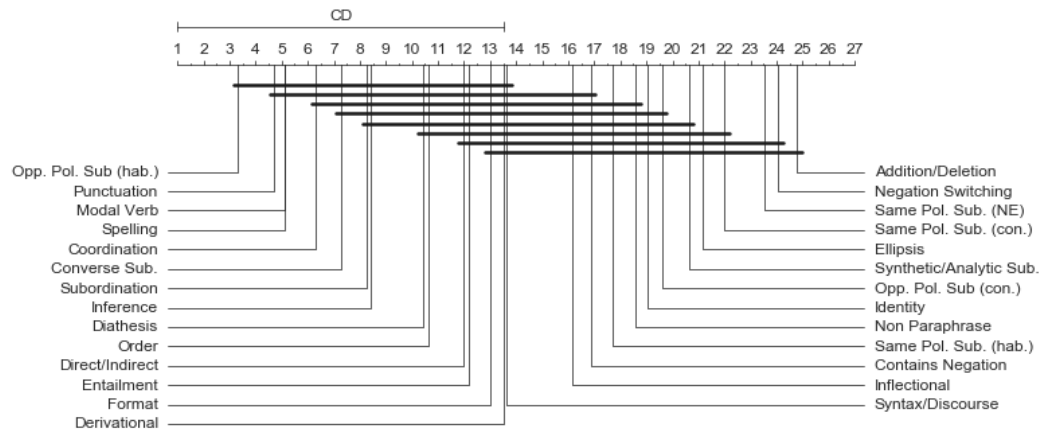

Figure 1: Critical Difference diagram of the average ranks by phenomena

a larger quantitative difference. Table 5 shows the comparison between **S3** and **S5** (Lan and Xu, 2018a). 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.

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 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 *"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 Discussion

In Section 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 perfor-

mance 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].

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 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.

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

References

Hanan Aldarmaki and Mona Diab. 2018. Evaluation of unsupervised compositional representations. In *Proceedings of COLING 2018*.

Darina Benikova and Torsten Zesch. 2017. Same same, but different: Compositionality of paraphrase granularity levels. In *Proceedings of RANLP 2017*.

Rahul Bhagat and Eduard H. Hovy. 2013. What is a paraphrase? *Computational Linguistics*, 39(3):463–472.

Elena Cabrio and Bernardo Magnini. 2014. Decomposing semantic inferences.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. *CoRR*, abs/1705.02364.

Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, 7:1–30.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of Coling 2004*, pages 350–356, Geneva, Switzerland. COLING.

Samuel Fernando and Mark Stevenson. 2008. A semantic similarity approach to paraphrase detection. *Computational Linguistics UK (CLUK 2008) 11th Annual Research Colloqium*.

M. Friedman. 1940. A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics*, 11(1):86–92.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. Ppdb: The paraphrase database. In *HLT-NAACL*, pages 758–764. The Association for Computational Linguistics.

Darina Gold, Venelin Kovatchev, and Torsten Zesch. 2019. Annotating and analyzing the interactions between meaning relations. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 26–36, Florence, Italy. Association for Computational Linguistics.

Hua He, Kevin Gimpel, and Jimmy Lin. 2015. Multi-perspective sentence similarity modeling with convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1576–1586. Association for Computational Linguistics.

Hua He and Jimmy Lin. 2016. Pairwise word interaction modeling with deep neural networks for semantic similarity measurement. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Shankar Iyer, Nikhil Dandekar, and Kornl Csernai. 2017. First quora dataset release: Question pairs.

Yangfeng Ji and Jacob Eisenstein. 2013. Discriminative improvements to distributional sentence similarity. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 891–896.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3294–3302. Curran Associates, Inc.

Venelin Kovatchev, M. Antònia Martí, and Maria Salamó. 2018a. Etpc - a paraphrase identification corpus annotated with extended paraphrase typology and negation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Venelin Kovatchev, M. Antònia Martí, and Maria Salamó. 2018b. WARP-text: a web-based tool for annotating relationships between pairs of texts. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 132–136, Santa Fe, New Mexico. Association for Computational Linguistics.

Wuwei Lan, Siyu Qiu, Hua He, and Wei Xu. 2017. A continuously growing dataset of sentential paraphrases. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 1224–1234.

Wuwei Lan and Wei Xu. 2018a. Character-based neural networks for sentence pair modeling. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Wuwei Lan and Wei Xu. 2018b. Neural network models for paraphrase identification, semantic textual similarity, natural language inference, and question answering. In *Proceedings of COLING 2018*.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Nitin Madnani, Joel Tetreault, and Martin Chodorow. 2012. Re-examining machine translation metrics for paraphrase identification. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 182–190, Stroudsburg, PA, USA. Association for Computational Linguistics.

H. B. Mann and D. R. Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Statist.*, 18(1):50–60.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, USA. Curran Associates Inc.

P.B. Nemenyi. 1963. *Distribution-free Multiple Comparisons*. Ph.D. thesis, Princeton University.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *In EMNLP*.

Mark Sammons, V. G. Vinod Vydiswaran, and Dan Roth. 2010. "ask not what textual entailment can do for you...". In *ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden*, pages 1199–1208.

M. Vila, M. A. Martí, and H. Rodríguez. 2014. "is this a paraphrase? what kind? paraphrase boundaries and typology. ". pages 205–218.

Zhiguo Wang, Haitao Mi, and Abraham Ittycheriah. 2016. Sentence similarity learning by lexical decomposition and composition. *CoRR*, abs/1602.07019.