

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.

6.3 Qualitative Evaluation Framework

6.3.1 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].

6.3.2 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 phenomenon². 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)?

6.4 PI Systems

To demonstrate the advantages of our evaluation framework, we have replicated several popular Supervised and Unsupervised PI systems. We have selected the

²i.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.

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., 2019] 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]
- [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 [2018b]

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., 2013b]
- [S8] Average over sentence of pre-trained Glove word embeddings [Pennington et al., 2014]