The core review

This paper is an empirical survey of recent works on data augmentation for NLP in the limited labeled data setting, and it covers various types of methods: token-level augmentations, sentence-level augmentations, adversarial augmentations, and hidden space augmentations. The main contributions/strengths are: 1) a comprehensive review of the methods with careful categorization; 2) a comparison on up to 11 NLP tasks by summarizing the datasets and experiments (i.e., Table 3).

Main weaknesses: I am not sure if a survey-type paper would be a good fit for the conference. In section 2, the summarized methods are very high-level, ignoring some model details, i.e., what neural network architectures, what language models, or word embedding. Such a way of description is also fit for non-deep learning methods. As the survey focuses on deep learning ones, it would be much better to stress how these methods are different from traditional non-deep learning ones. Moreover, limited in-depth analysis is made when comparing the methods in section 4.

Reasons to Accept

This is a comprehensive survey focusing on the problem of data augmentation for NLP tasks, while no other similar work has been done before. A good method categorization and dataset summarization and comparison. Besides, benchmarking dataset evaluation is provided.

Reasons to Reject

The way of categorizing methods is not so novel; only limited insights and discussion are provided in the comparison and future work.

Overall Recommendation: 3

Questions for the Author(s)

1. Section 2.2 Paraphrasing. How is this method different from non-traditional methods (i.e., WordNet might be able to provide replacing tokens), is it a special method for deep methods? Due to the paragraph, I found it difficult to understand in which perspective these models are better than the others.

2. The title of the paper is too broad, as “Limited Data Learning in NLP” might contain more languages rather than English and
various domains, it is better to have a better explanation in the introduction.

3. The equation in Line 470 is a good one but it may be possible to generalize “CE” as a distance function “D”, in this case, it is generalized to other possible functions, i.e., KL divergence or MSE.

4. Until Section 4 I could understand that the authors experimented with the datasets and models, it is better to make it more explicit in the introduction by saying “benchmark or empirical evaluation”. In line 94 “compare ...through experiments”, this could be misleading.

5. Table 2 is a nice table that evaluates and compares selected DA models on four datasets. Would these datasets be domain-specific? The models perform better in Table 2, are not always performing better in Table 3 (i.e., Synonym Replacement under supervised setting), what are the reasons that the performance is not stable? Maybe a more detailed analysis might be helpful.

Review #2

The core review
This paper presents a survey of data augmentation techniques in NLP and discusses potentially future directions of research. They empirically evaluate a few of the data augmentation techniques they have mentioned on a wide range of classification tasks including the GLUE benchmark. From their results, it seems that there is no one data augmentation technique that performs well in all settings. Also in some cases, data augmentation techniques worsen the performance of the model. While the survey seems exhaustive and well written, the empirical results seem lacking in certain aspects.

Reasons to Accept

- The paper is clearly written.
- The survey overall seems extremely thorough and the authors seem to have exhaustively covered all types of data augmentation methods.
- They have run their results over multiple seeds.

Reasons to Reject
This paper only compares the given data augmentation methods on text classification, for a more thorough evaluation other tasks like translation, question answering, etc. should also be considered.

The results tables seem too dense and too many results have been crammed into the tables. It's difficult to have any insights from it.

The authors have not thoroughly considered the data augmentation methods that they have defined in section 2. For example, they have not considered adversarial or interpolation-based techniques defined in sections 2.3 and 2.4. Also, they have considered random insertion, random deletion, and random swapping separately, while in practice these can be combined as well as they have mentioned in section 2.1.

Overall I think this paper lacks a thorough empirical evaluation. While it can be said that considering all the mentioned data augmentation techniques may be expensive, they should consider using at least one technique for each representative concept (interpolation-based, adversarial, token-level, etc).

Overall Recommendation: 3

Missing References
Interpolation-based techniques
- Augmenting NLP models using Latent Feature Interpolations
  COLING 2020

Review #3

The core review
This paper surveys methods for generating textual data by transforming existing training data in ways that preserve (or predictably transform) labels. Its empirical contributions are tables 2 and 3, showing performance of methods on a number of different tasks and data sizes. The best-performing methods look fairly randomly distributed.

Reasons to Accept
Survey papers can be very useful for the community.

Reasons to Reject
Surveys aren't about new methods, but I also don't see what meaningful takeaways the paper creates: there doesn't appear to be a pattern to the performance of the methods, which might lead to interesting investigation, but the confidence windows seem to show very low significance.

**Overall Recommendation:** 2.5
Rebuttal:

Expectations for survey paper.

General Response to Reviewers:

We thank all the reviewers for their constructive feedback. We are glad that reviewers think our survey is well written (R2), comprehensive (R1, R2) with careful categorization and experiments (R1) and will be useful for the community (R3)

The contributions in our work are similar to previous surveys at NLP conferences like [1, 2] which summarize and categorize previous methods and discuss challenges and future directions. Additionally, we perform experiments on a wide range of text classification datasets to show how they work on different tasks.

The major takeaway is that unlike many previous works [3, 4] which assume specific augmentations work better based on limited tasks; our empirical results suggest that there are no augmentation methods that work best for all tasks. So more work is needed to better augment text data. Also, our results table could provide guidelines for which augmentation strategy to use for given tasks. Future work should experiment with various tasks to demonstrate the effectiveness. There are also several patterns we observe, for example, for supervised learning with limited data, word replacement works the best for simple tasks such as new or topic classification, but performs the worst for more difficult tasks such as MNLI. Also roundtrip translations work very well overall for semi-supervised learning across various tasks.

We address the reviewers’ comments below and will incorporate all the feedback in our revised version.


Response to Review #1:

Thanks for the detailed review!

1. Model details in section 2: We did not go over the specific architectures since most of the augmentation methods are model-agnostic and performed in the input space so that they can be applied to any neural architectures/language models/embeddings. For experiments, we only perform augmentations for neural models (BERT), since they are the current de facto model for the problems we consider.

2. Limited In Depth Analysis: We do observe certain patterns like the best augmentation methods differ for supervised and semi-supervised settings. But there are not clear trends across all the tasks, suggesting that the best augmentation method depends on the dataset and setup. See the General Response for more details.

3. Response to Questions
   ○ Q1: The paraphrasing we are referring to in section 2.2 is more from a sentence-level perspective, i.e. modifying the entire sentence (including sentence structures, word choices and etc.), rather than only replacing certain tokens like the token-level augmentations (like WordNet replacement) mentioned in section 2.1. It is hard to say which way is definitely better than the other based on our experiment results (synonym replacement vs round-trip translation). The superiority is dependent on tasks/datasets.
   ○ We will fix Q2, Q3, and Q4 in the revised version.
   ○ Q5: Table 2 is more about news/topic classification while Table 3 provides a broader range of tasks like predicting relations between multiple sentences and sentiment classifications. The performance differences actually shows that the optimal augmentation method varies across different tasks, which differs from previous works and suggests there is not one method that will work best. We speculate synonym replacement works well for topic classification since the texts are simple, and not all the words are necessary to classify the input, while for tasks like MNLI, each token is more important for the task, and synonym replacements might not always make sense in context.

Response to Review #2:

Thanks for the constructive feedback! We make several clarifications here:

1. More tasks: In this work, our experiments mainly focus on classification tasks (like GLUE) to examine how different augmentations help natural language understanding, though
we do review augmentation methods for different tasks as shown in Table 1. We leave exploration of data augmentation on a wide range of other tasks (machine translation, question answering, etc.) for future work.

2. **Table:** Due to the space limit, the table looks a bit dense. We will fix the table and make it easier to interpret. The insight is that no augmentation methods work best for every task. Future work can actually refer to our results for suggestions to select augmentation strategies for their own tasks, e.g., using Cutoff for sentiment classification based on the results for SST-2.

3. **Ensembling:** As an empirical survey paper, our focus is on comparing different individual augmentation methods, not ensembling them to achieve state-of-the-art performance since there is an exponential number of possible combinations. But we could add results for ensembling all the augmentation methods in the final version.

4. **Other Augmentation Methods:** In our experiments, we mainly examine the most commonly used methods which can be applied to both supervised and semi-supervised settings, including the token-level, sentence-level, and hidden-space augmentations. We will include interpolation / adversarial-based augmentations for supervised settings in the final version’s extra page.

**Response to Review #3:**

Thank you for the feedback! Here we would like to clarify the contributions in our survey. More takeaways can be found in the General Response.

1. We thoroughly summarize and categorize recent augmentation methods to help people get a better understanding of data augmentations in the NLP fields. Also, we summarize the semi-supervised framework (consistency training) to apply augmentation methods to unlabeled data.

2. We perform empirical studies on the same datasets for a fair comparison showing that different tasks might require different augmentation strategies. Our results can provide simple guidelines for people to select augmentation methods for their own tasks. This is similar to other empirical survey-style papers [1]

3. We also discuss current challenges and future directions of data augmentation which might inspire future work in augmentation for NLP.
