Label Representation in Modeling Classification as Seq2Seq

Xinyi Chen
New York University
xc1121@nyu.edu

Jingxian Xu
New York University
jx880@nyu.edu

Abstract

In the settings of state-of-the-art transfer learning methods that model classification tasks as text generation, labels are represented as strings for the model to generate. We run T5-base with T5 default labels (Raffel et al., 2019) to obtain baseline. By implementing experiments on tasks with different label representations using T5 in rich and low-data regime, we find that for COPA (Roemmele et al., 2011) which contains indicative information in input, task-related labels perform the best, while rich data can help achieve better performance for other labels. However, for other tasks, labels do not have much influence on the performance. We provide our project source code on Github for reference[1].

1 Introduction

State-of-the-art transfer learning methods model classification tasks as text generation, e.g. autoregressive language modeling, such as GPT2 (Radford et al., 2019) or Seq2Seq, such as T5 (Raffel et al., 2019), and have led to improvements in almost every benchmark, contributing a lot to the field of natural language processing in the past few years. In this setting, pre-trained language models are fine-tuned on different downstream tasks, and the labels are represented as strings for the model to generate.

The outstanding performance of transfer learning methods has spurred on research about whether to attribute the performance to the pre-trained models or the fine-tuning (Talmor et al., 2019). They have shown that it actually depends. While T5 can exploit knowledge from both parts of the model (including those that are latent), the final layer of BERT (Devlin et al., 2018) has to learn from the beginning (Nogueira et al., 2020).

[1]https://github.com/silvia0v0/Label-Representation-in-Modeling-Classification-as-Seq2Seq

Despite the efforts, it remains unclear what roles label representations play in classification tasks modeled as text generation. Recent research on fine-tuning a particular task using Seq2Seq model (T5 [Raffel et al., 2019]) involving experiments on different types of labels has suggested that the linguistic property (linguistic relatedness, polarity scale, etc.) of the labels does matter to the overall performance (Nogueira et al., 2020). We therefore raise the question of the best string to represent the label in Seq2Seq models that are handled in the way of text generation, in this paper specifically, T5 (Raffel et al., 2019), or if the labels need to be represented with task-related strings at all.

Our experiments and results are summarized as follows: i) inspired by Nogueira et al. (2020), design a wide range of label representation types; ii) run T5-base model (Raffel et al., 2019) on CoLA (Warstadt et al., 2018), COPA (Roemmele et al., 2011), MRPC (Dolan and Brockett, 2005), SST-2 (Socher et al., 2013) using full datasets with label representations we design, and find that model performance differs on different labels only for COPA; iii) for SST-2 and COPA, run model on reduced datasets, which are obtained by randomly sampling from the original datasets, for each label we repeat three such trials and average the results since there is no significant differences observed; iv) find that for SST-2, reducing data does not make difference, and for COPA, it still does not impact performance of T5 default label and some task-related labels, but largely reduces accuracy for other labels. Complete experiment results and specific labels can be found in Appendix A.

2 Related Work

As Language Model pre-training and finetuning becomes increasingly popular, many studies have investigated in improving transfer learning. In the

setting of modeling classification task as a "text-to-text" problem, a lot of research focus on the impacts of input and output formats.

Target Word Probing Experiments Nogueira et al. (2020) probe the effects of target words on document ranking task with T5. They set baseline mapping as {Positive: "true", Negative: "false"}, and try reverse mapping, antonyms, related words, unrelated words and subwords. In low-data (2k) regime, they find that the baseline mapping produces accuracy significantly higher than other types of mappings. In rich-data regime, related words mapping is the most effective one, the differences in accuracy between different mappings are not as large as they are in low-data regime. But in both low-data and rich-data regime, the authors observe notable differences between reverse mapping and baseline mapping. Inspired by the target word probing experiments by Nogueira et al. (2020), we design more label representations and experiment on more classification tasks.

Cloze-style Statements Schick and Schütze (2020) introduce Patter Exploiting Training, where the input is reformulated in cloze-style phrases. For example, the input of a task identifying whether two sentences a and b contradict each other or agree with each other is reformulated to "a?_, b.", and the model is trained to generate "Yes" or "No" to fill in the blank. They claim that this semi-supervised training procedure significantly improves performance on several tasks.

Petroni et al. (2019) probe on knowledge presented by some state-of-the-art Language Models without fine-tuning. They find that Language Models learn a lot of factual knowledge in pre-training, and are surprising good at recalling knowledge when answering fact-related questions. We will further discuss the relevance of both research to our experiment results in Section 4.

Prompt Experiments The effectiveness of prompts is proved to be important for Language Model performance on tasks. It is possible that we fail to retrieve information which LM knows due to the inefficiency of prompts. Jiang et al. (2019) propose several methods to generate more efficient prompts instead of just using manual ones, which might not be optimal because researchers create them without actually knowing which prompts are the best for utilizing what LM knows. One intriguing result is that, using "x who converted to y" instead of "x is affiliated with y religion" raises the accuracy by 60%. This shows that although for humans, some words or phrases have the same meaning, for LM they are probably not the same thing. Taking this into account, we test the performance of model on label representations as paraphrase of original labels, e.g. "equivalent" → "equal".

3 Methodology

3.1 Language Model

As we focus on models that model classification tasks as text generation, our project is based on a Seq2Seq model, which is T5 (Raffel et al., 2019) as we mentioned. To examine the impact of changing label representation, we constrain the variants and only utilize pre-trained T5-base as our model and finetune it on our tasks with re-assigned labels.

3.2 Dataset

To compare the performances of different label representations, we choose the following four tasks and datasets with finite possible targets:

- Sentence acceptability judgment (CoLA [Warstadt et al., 2018])
- Sentiment analysis (SST-2 [Socher et al., 2013])
- Paraphrasing/sentence similarity (MRPC [Dolan and Brockett, 2005])
- Sentence completion (COPA [Roemmele et al., 2011])

3.3 Methods

We first generate label representations for the tasks mentioned, then finetune the tasks using T5, and compare the results with our baselines.

3.4 Generate label representations

To compare different labels with the default labels presented by T5 (Raffel et al., 2019) that are listed in Appendix B, we take the CoLA task as an example, which has "unacceptable" and "acceptable" as original targets, and generate our label representations in the following ways:

Random Labels We want to test whether the labels need to make semantic meanings at all. So, we randomly generate labels with length ranging from 1 to 15, and map our targets to them. For example, we try random labels of similar length, e.g.

{“unacceptable”: “khhzjfofi”, “acceptable”: “douzzcmri”}, as well as random labels that differ largely in length, e.g.{“unacceptable”: “i”, “acceptable”: “eutelelwld”}.

Task-unrelated Labels We choose words as labels for the next step. These labels make sense semantically but are unrelated to the task or the original targets. Again, we pose the question of whether task-unrelated labels perform worse than task-related ones (e.g. original labels), and how differently they perform compared to labels that do not make sense linguistically at all. We generate them by following different rules:

- **irrelevant words:** We choose words that are not related to each other and map the targets to them, e.g. {“unacceptable”: “ice”, “acceptable”: “happy”}.
- **relevant words:** Then, we pick words as labels that are relevant but are not antonyms, and map the targets to them, e.g. {“unacceptable”: “potato”, “acceptable”: “tomato”}.
- **Antonyms:** We also choose antonyms that are semantically unrelated to our tasks, e.g. {“unacceptable”: “hot”, “acceptable”: “cold”}. For each pair of antonyms, we also try the reverse, e.g. {“unacceptable”: “cold”, “acceptable”: “hot”}.

Task-related Labels We further study if different task-related labels perform differently. We choose words that are related to our tasks, or generate labels based on the original ones, i.e. choosing synonyms, cognate words, etc.

- **Matched labels:** We choose words (usually antonyms) as labels that have the same polarity scale or meanings as the original targets, i.e. semantically, positive labels are mapped to positive results, e.g. {“unacceptable”: “reject”, “acceptable”: “accept”}.
- **Reversed labels:** In contrast, we also use labels that have the opposite polarity scale or meanings as the original targets as opposed to matched labels, e.g. {“unacceptable”: “accept”, “acceptable”: “reject”}.

4 Experiments and Results

Baseline We obtain baseline by running T5-base with default formatted inputs listed in Appendix B.

		MRPC	**SST-2**	**SST-2**	**CoLA**	**COPA**	**COPA**
	metrics	acc	acc	(sample)	MCC	acc	(sample)
label	size(train)	3,668	67,349	600	8,551	400	200
original **(baseline)**		88.9	94.9	95.5	58.765	70	72
reversed		89.461	94.266	94.5	58.748	69	68
random	short/short	89.951	94.151	95	58.485	64	46
	short/long	88.725	94.5	94	57.969	57	48
	long/short	89.9	94.036	94	57.814	64	46
	long/long	90.196	94.266	94	57.3	57	50
related	matched	89.335	94.728	94	58.677	71	72
	reversed	88.969	94.471	94	58.198	71	52
unrelated	antonyms	89.951	94.318	94.5	57.454	65	50
	reversed	88.725	94.295	95	57.335	66	54
	relevant	89.706	94.626	93.5	57.239	58	50
	irrelevant	89.615	94.658	93.5	57.047	60	36

note: “acc” refers to accuracy; “MCC” refers to Matthew’s correlation coefficient.

Table 1: summary of results we get using full dataset and reduced dataset

Experiments with Label Representations We firstly experiment with different label representations listed in Section 3.4 with full datasets listed in Section 3.2. Table 1 shows that, no matter what label representations we use, the accuracies of MRPC, CoLA or SST-2 are always around baselines, while for COPA we observe notable differences on model performance using different labels. The accuracies when we use the original label, reversed original label and task-related labels are the best ones, all around 70%, while for random labels and task-unrelated labels the accuracies are a lot lower. Interestingly, table 2 shows that random label pair “i”/“c” results in much lower accuracy than “n”/“p”, which means that the model performance with very short random labels involves a degree of uncertainty. The fact that task-related labels all perform well implies that our language model contains both factual and linguistic knowledge, as it recognizes the connection between the original labels “choice1”/“choice2” and “first”/“second”. The accuracy we get using “first”/“second” as labels is 71%, while using “third”/“fourth” only gives 61%, meaning the model also recognizes that “third”/“fourth” are associated with different numerical values.

The differences between COPA and other three tasks could attribute to the fact that COPA input is “copa choice1:...choice2:...premise:...question:...”, which is the only task input involving the original labels. Inputs and outputs of other tasks provided by Raffel et al. (2019) is listed in Appendix B. The appearance of “choice1” and “choice2” in the input of COPA could be an indicative message which makes it easier for the model to directly generate one of them as output. Similarly, Nogueira et al. (2020) also embed indicative message in input, which is “Query: q Document: d Relevant:”,

and observe that different types of label representations result in notable differences in performance. As pointed out by Petroni et al. (2019), a language model is a container of factual and linguistic knowledge, so it is more sensible to generate "Query: q Document: d Relevant: True" than "Query: q Document: d Relevant: apple", so in this case "Relevant:" plays as an indicative message.

By generating labels as strings, the Seq2Seq model is essentially finishing a sentence completing task, and adding an indicative message (which are words related to labels) to the input actually imposes consistency, relevancy and logicality on the sentence to be completed. In this context, task-related words are more effective representations of labels.

Similarly, in Pattern Exploiting Training proposed by Schick and Schütze (2020), the input to model is reformulated in a cloze-style statement, which is considered as a pattern, and every task has a unique pattern of input. Though there might not be specific indicative words in reformulated input, its pattern has already been embedded with linguistic features and factual knowledge in natural language, which help the model understand the task better and generate task-related strings to fill in the blank. In the example we give in section 2, "Yes" and "No" are such strings.

label(type) size(train)	metrics label	acc 400	acc 200
random	i/c	64	46
	n/p	72	66
task-related	first/second	71	72
	second/first	71	52
task-unrelated	third/fourth	61	50

Table 2: some examples we run for COPA

Experiments with Dataset Size Noticing that the COPA task is special, inspired by Nogueira et al. (2020), who see significant differences between performances using different data sizes, we further compare the performances between COPA and SST-2 by running both tasks in a lower-data regime, which means that we sample part of the data to form reduced datasets, and re-run the model on reduced datasets. We repeat sampling and re-running for three times, but the results do not show a significant difference. The data shown in the tables are the average of results from three trials.

For SST-2, the full dataset size is 67,349. We use a reduced dataset of size 600, i.e. less than 1/100 of the original size. For COPA, the full dataset has a size of 400 training data, and we reduce it to half the original size. Our results in table 2 shows that the accuracies of sampled SST-2 do not vary significantly from that of the original dataset. However, for COPA, reducing data does not make a big difference for the original label and reversed original label. This is because pretraining on huge dataset has already enabled the model to learn the factual and linguistic knowledge and predict the original label strings, even if we reverse them. But for most of the other label representations, the accuracies we get by training in low-data regime are much lower than with full dataset.

Specifically, we observe that in low-data regime, when using the label pair "second"/"first", the model gives a significantly worse accuracy at about 52% comparing to 71% with full dataset, while using "first"/"second" gives us 72% accuracy, which is basically the same as with full dataset. On seeing that the gap between performances of different labels widens as we reduce the size of data and noticing that the full dataset size of COPA is still relatively small, we believe that dataset size accounts for the difference of performances between distinct label pairs.

Our experiments suggest that while the labels do have impact on the performance for tasks that involve directly indicative words in the input, this influence may be eliminated or improved by increasing dataset size.

5 Conclusion

In this research, we design a wide range of label representations, with which we run T5-base model on four tasks in rich and low data regimes. MRPC, SST-2 and CoLA are not sensitive to label or data size, while COPA is. We attribute this difference to the input format. If the task input format includes some direct indicative messages, then the label representations that are relevant to the indicative messages tend to result in better accuracy. However, if the input does not include any direct indicative message, different label representations do not make a big difference on the model performance.

With the input which takes advantage of the knowledge learned by model during pretraining, it is possible for future studies to find more informative label representations and improve model performance.

Collaboration Statement

We developed the plan for the project together and both edited the code. Xinyi ran experiments on CoLA and COPA, and Jingxian ran experiments of MRPC and SST-2. For the presentation, Jingxian made most of the slides, and Xinyi created the tables. We both prepared and implemented the presentation. For the paper, Jingxian found related papers and wrote Sections 2, 5, and the first two parts of Section 4, Xinyi wrote the rest. We both edited it.

References

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

William B Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2019. How can we know what language models know?

Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. 2020. Document Ranking with a Pretrained Sequence-to-Sequence Model. *arXiv e-prints*, page arXiv:2003.06713.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2019. Language models as knowledge bases?

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv e-prints*, page arXiv:1910.10683.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Timo Schick and Hinrich Schütze. 2020. Exploiting cloze questions for few shot text classification and natural language inference.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2019. oLMpics – On what Language Model Pre-training Captures. *arXiv e-prints*, page arXiv:1912.13283.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2018. Neural Network Acceptability Judgments. *arXiv e-prints*, page arXiv:1805.12471.

A Tables

Label Type	Task Metrics Size(train)	**MRPC** Accuracy 3,668	**CoLA** Matthew's 67,349	**SST-2** Accuracy 8,551	**SST-2(sample)** Accuracy 600	**COPA** Accuracy 400	**COPA(sample)** Accuracy 200
original (**baseline**)		not_equivalent/equivalent	unacceptable/acceptable	negative/positive		choice1/choice2	
		88.9	58.765	94.9	95.5	70	72
reversed original		equivalent/not_equivalent	acceptable/unacceptable	positive/negative		choice2/choice1	
		89.461	58.748	94.266	94.5	69	68
random	short/short	i/c					
		89.951	58.485	94.151	95	64	46
		n/p					
		89.856	58.485	94.375	94.5	72	66
	short/long	i/eutelelwld					
		88.725	57.969	94.5	94	57	48
	long/short	eutelelwld/i					
		89.9	57.814	94.036	94	64	46
	long/long	khhzjfofi/douzzcmri					
		90.196	57.3	94.266	94	57	50
task-related	matched	unequal/equal	accept/reject	no/yes		first/second	
		89.335	58.677	94.728	94	71	72
	reversed	equal/unequal	reject/accept	yes/no		second/first	
		88.969	58.198	94.471	94	71	52
task-unrelated	antonyms	cold/hot					
		89.951	57.454	94.318	94.5	65	50
	reversed antonyms	hot/cold					
		88.725	57.335	94.295	95	66	54
	relevant	apple/orange					
		89.706	57.239	94.626	93.5	58	50
						third/fourth	
						61	50
	irrelevant	ice/happy					
		89.615	57.047	94.658	93.5	60	36

Table 3: All data with specific label strings noted

B Preprocessed examples

In this section, we provide examples of our preprocessing (same as T5 (Raffel et al., 2019)) for each of the datasets we use.

- **CoLA**
  Original input:
  `Sentence: John made Bill master of himself.`
  Original target: `1`
  Processed input:
  `cola sentence: John made Bill master of himself.`
  Processed target: `acceptable`

- **MRPC**
  Original input:
  `Sentence 1: We acted because we saw the existing evidence in a new light , through the prism of our experience on 11 September , " Rumsfeld said . Sentence 2: Rather , the US acted because the administration saw " existing evidence in a new light , through the prism of our experience on September 11 " .`
  Original target: `1`
  Processed input:
  `mrpc sentence1: We acted because we saw the existing evidence in a new light , through the prism of our experience on 11 September , " Rumsfeld said . sentence2: Rather , the US acted because the administration saw " existing evidence in a new light , through the prism of our experience on September 11 " .`
  Processed target: `equivalent`

- **SST-2**
  Original input:
  `Sentence: it confirms fincher 's status as a film maker who artfully bends technical know-how to the service of psychological insight .`
  Original target: `1`
  Processed input:
  `sst2 sentence: it confirms fincher 's status as a film maker who artfully bends technical know-how to the service of psychological insight .`
  Processed target: `positive`

- **COPA**
  Original input:
  `Question: effect Premise: Political violence broke out in the nation. Choice 1: Many citizens relocated to the capitol. Choice 2: Many citizens took refuge in other territories.`
  Original target: `1`
  Processed input:
  `copa choice1: Many citizens relocated to the capitol. choice2: Many citizens took refuge in other territories. premise: Political violence broke out in the nation. question: effect`
  Processed target: `choice2`