Evaluating Whether Personal Goals Are SMART

Neel Rao

Stanford University Department of Computer Science

Problem and Background

Goal: Identify whether a given written goal is Specific, Measurable, and/or Time-bound.

Motivation: Findings from Education research show that the practice of writing learning goals is beneficial for cognitive and noncognitive skill development, and that these effects are increased when the goals follow the SMART framework, meaning they are Specific, Measurable, Attainable, Relevant, and Time-bound (Lawlor, 2012; Moeller et al., 2012). Because attainability and relevancy vary based on the writer, I focus on the other three facets here.

Previous Work: This is a previously unpublished-on area of research, and no public datasets exist for this task. However, much work exists on the methodologies utilized (e.g., multitask learning, fine-tuning BERT Transformers)

Data

Task: Given a written personal goal, the task is to evaluate whether it is specific, relevant, and/or time-bound.

Ex. 1) "I want my model to achieve 90% accuracy on the test set in the next week." [1,1,1]

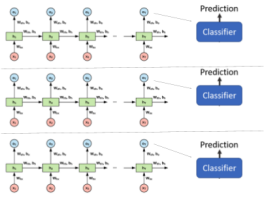
Ex. 2) "I want to get better at programming" [0,0,0]

Generation: The dataset was hand collected, cleaned, labeled and preprocessed. It was collected by scraping the internet for goal examples and crowdsourcing from various Stanford students.These examples were then labeled as specific, measurable, and/or time-bound by a single rater. While a lack of multiple raters is suboptimal, SMART goal evaluation has been found to have high interrater reliability, so it is believed to be sufficient (Lawlor, 2012). The dataset consists of 1000 goals, each mapped to a triple. 694 are labeled as specific, 473 are measurable, and 357 are time-bound. Lastly, care was taken to make sure that the correlations between each label was not too high. Goals that exhibited multiple SMART characteristics were sometimes chopped up such that they only exhibited one of the characteristics.

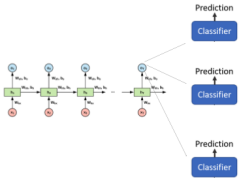
Approach and Methods

I experiment with four systems:

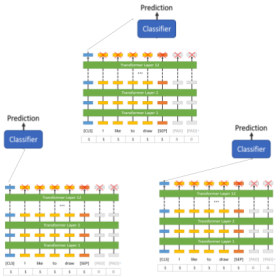
System 1: Three separate vanilla RNNs with final linear output layer, one model for each classification task.



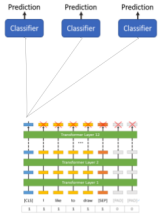
System 2: A single vanilla RNN branching into three task-specific output layers, utilizing multitask learning with hard parameter sharing



System 3: Three BERT models with linear final layer for binary classification, each fine-tuned on a separate task.



System 4: A single fine-tuned BERT model with three task-specific output layers, utilizing multitask learning with hard parameter sharing.



Multitask learning was experimented with for performance and deployment considerations. It can help produce more robust representations, prevent overfitting to the training set, and reduce number of parameters necessary

Results

F1 Scores

System	Specificity Task	Measurability Task	Time-boundedness Task
System 1 (Three RNNs)	.827	.606	.538
System 2 (Single Multitask RNN)	.77	.636	.61
System 3 (Three BERT Transformers)	**.933**	.890	**.938**
System 4 (Single Multitask BERT Transformer)	.923	**.905**	.915

Accuracies

System	Specificity Task	Measurability Task	Time-boundedness Task
System 1 (Three RNNs)	70%	62%	72%
System 2 (Single Multitask RNN)	66%	60%	69%
System 3 (Three BERT Transformers)	**90**%	89%	**95**%
System 4 (Single Multitask BERT Transformer)	89%	**90**%	93%

Results for the 4 systems on each of the 3 tasks are shown above.

– **The BERT Transformer models significantly outperform** the RNN models, as expected.

– **The two BERT systems perform almost identically,** and both achieve fairly promising results, with F1 scores around .9. The similarity between the two is encouraging because it allows for a decrease in the amount of parameters necessary to get these results.

Conclusions and Future Work

I believe this project offers two contributions:

– First, it creates a **dataset of 1000 written personal goals**, labeled for whether they are Specific, Measurable, and/or Time-bound. Hopefully this dataset can be used for other projects similar to this one, to develop AI coaching.

– Secondly, this project offers a**n NLP system to evaluate written personal goals for specificity, measurability, and time-boundedness.** This model consists of a single fine-tuned BERT base, that then branches into three task-specific linear output layers. The model achieves good (though not spectacular) results, with F1 scores and accuracies in the low .9 range across the 3 tasks.

– The results are not yet high enough for deployment, but this project hopefully provides a **proof of concept to further explore Transformer-based multitask learning for these tasks**.

– I am hopeful that with more data, and more refinement, it can be a **component of an AI coach** in the future.