Tailor: Generating and Perturbing Text with Semantic Controls

Alexis Ross*, Sherry Tongshuang Wu*, Hao Peng, Matthew E. Peters, Matt Gardner
{alexisr,matthewp}@allenai.org, {wtshuang,hapeng}@cs.washington.edu, mattgardner@microsoft.com

TL;DR: fine-grained, compositional perturbs. with LMs

Motivation and Problem

Controllable generations through semantic perturbations are useful, but hard to get:

- Rely on pre-defined control signals;
- Task-specific data & training is expensive;
- Perturbations are underspecified & not compositional.

Contribution

Tailor =

- Controlled generator: follow *task-agnostic* semantic control codes.
- + Operations: Combine control codes into *task-specific* perturbations.

Constructs lexically diverse & less biased contrast sets.

Support counterfactual data augmentation

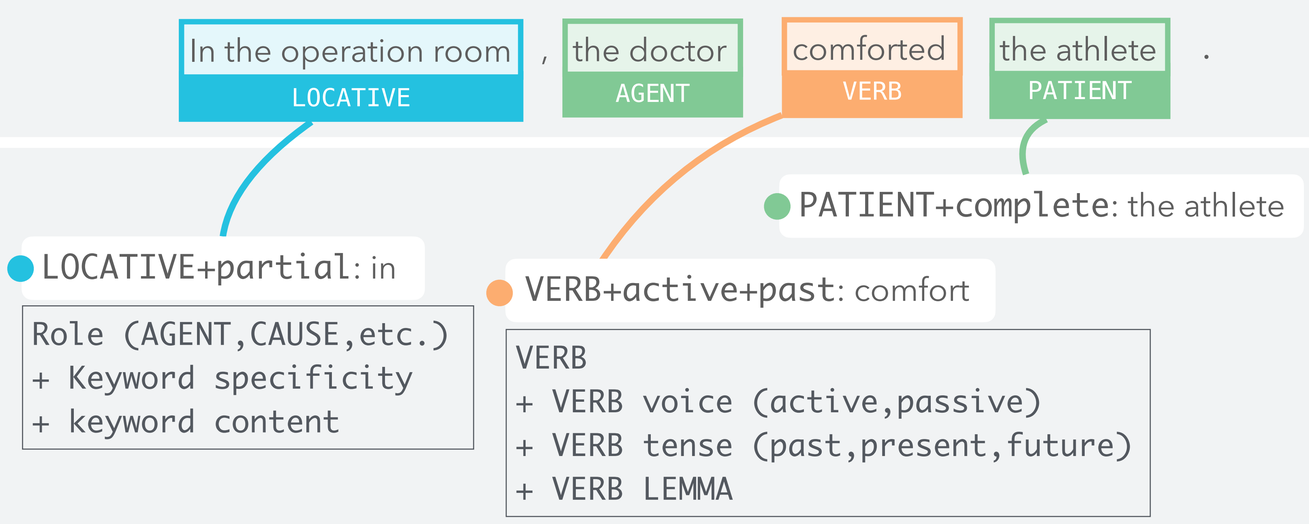
Walkthrough example: designing & using Tailor

Original sentence: In the operation room , the doctor comforted the athlete .

Propbank semantic formalism

SPAN-LEVEL CONTROLS

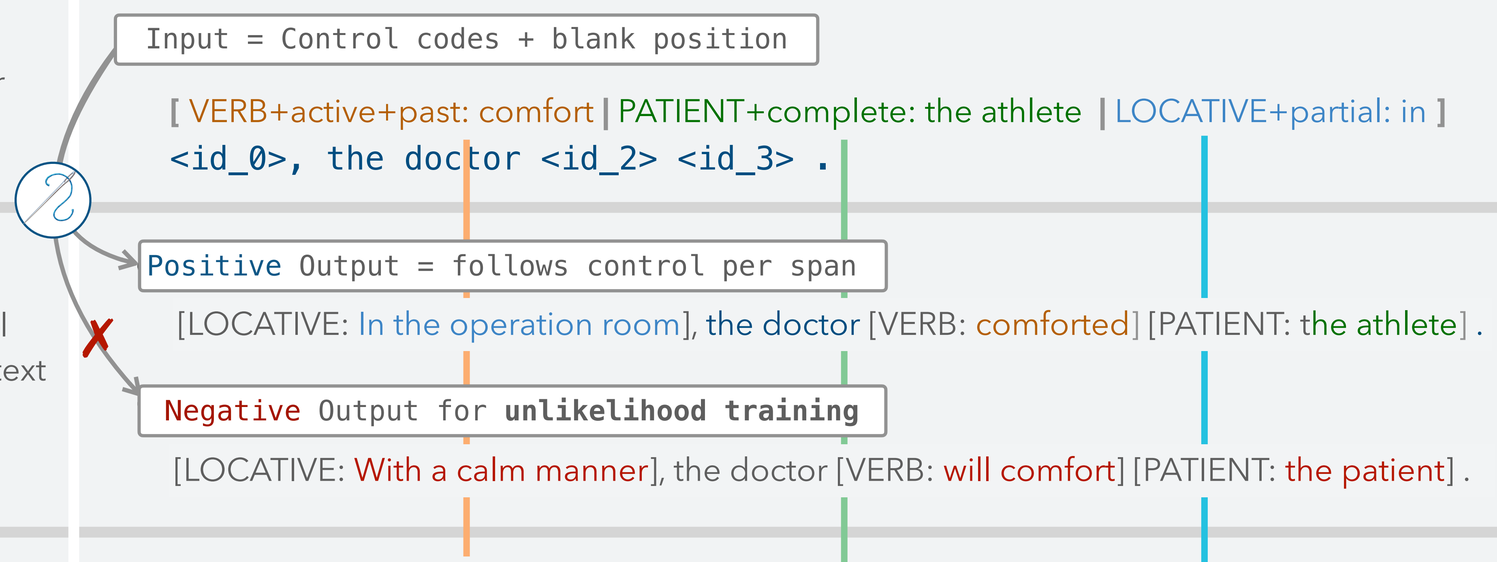
Fine-grained controls, specifying *thematic roles, verb form, keyword content & specificity*



Blank tokens, to control *span order*

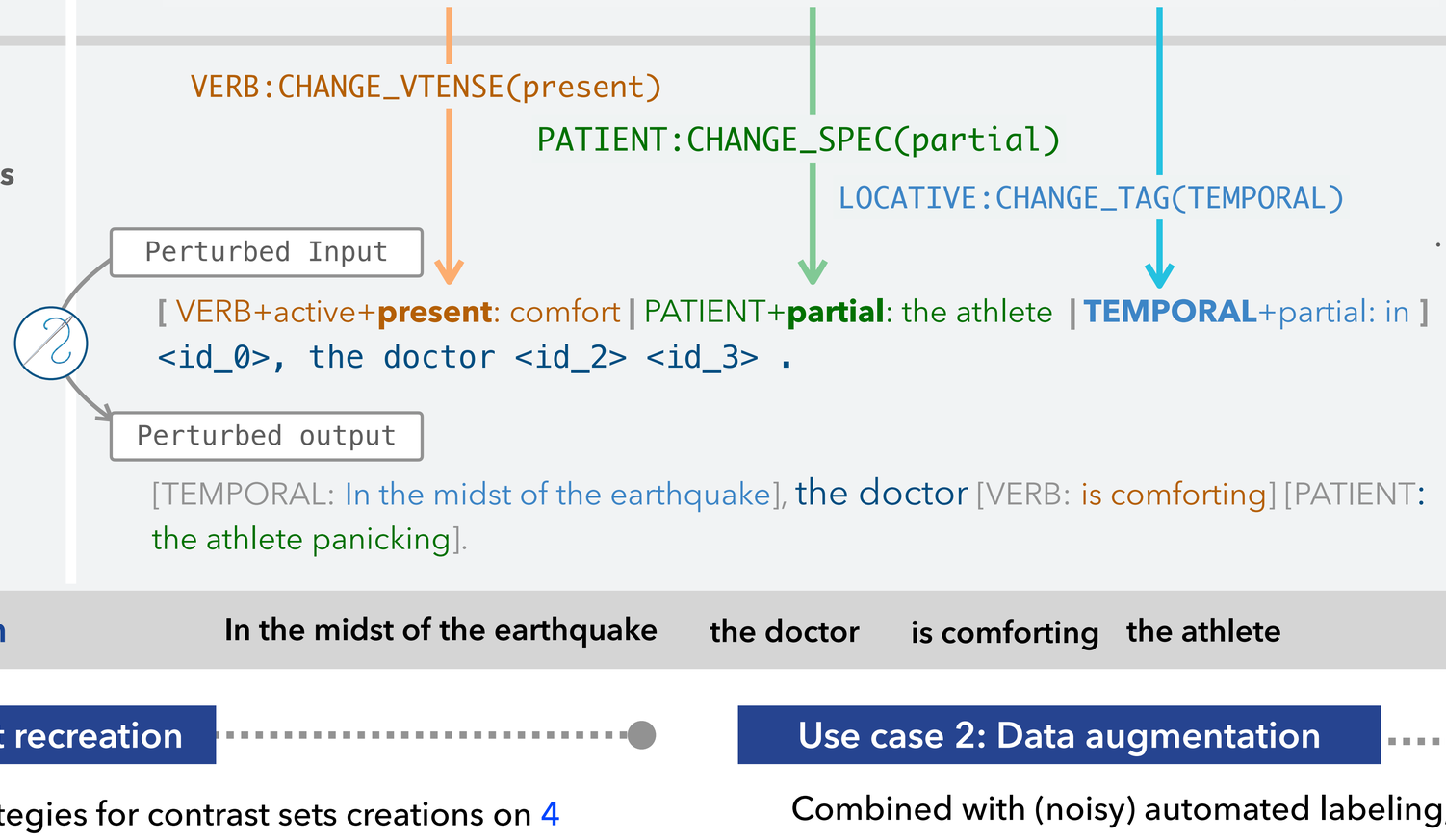
TRAINING

Train **seq2seq** model to translate controls to text



PERTURB

Perturbation operations for easily expanding to *compositional* perturbations



Cleaned perturbation: In the midst of the earthquake the doctor is comforting the athlete

Training Tailor

Generator

Fine-tune T5-base

Data

Inputs/outputs automatically created from OntoNotes 5.0

~220k positive examples
~540k negative examples

Result

Fluent perturbations

Follows controls

Unlikelihood improves controllability

Checkout our paper!

Checkout our GitHub repo!

Use case 1: Contrast set recreation

Recreate perturbation strategies for contrast sets creations on 4 diverse tasks, to reduce annotation effort.

Example: *Change preposition attachment in UD dependency parsing*

Original: It has a diverse range of food at all prices and styles.

Perturbed: It has a diverse range of food at every turn.

Dataset & task	Validity
UD English (dependency parsing)	63%
BoolQ (boolean question answering)	82%
MATRES (temporal relation extraction)	71%
SQuAD (extractive question answering)	81%

The recreated sets have high validity (need little or no manual fix.)
They also have high lexical diversity and reduced dataset bias.

Use case 2: Data augmentation

Combined with (noisy) automated labeling, Tailor improves model generalization via data augmentation.

Experiment: *Recreate HANS (McCoy et al., 2019) perturbation strategies, auto-label following Min et al. (2020)*

Premise: The judge behind the manager saw the doctors.

Lexical overlap, non-entailment

Hypothesis: The doctors saw the manager.

Training Data	SNLI	HANS Subset All	HANS Subset Entail.	HANS Subset Non-entail.
SNLI Training set	**91.1**	64.7	**99.0**	30.5
+ Syntactic Perturbations	91.0	67.5	95.8	39.2
+ **Tailor Perturbations**	**91.1**	**70.5**	81.3	**59.7**

2% additional perturbation mitigates over-reliance on syntactic heuristics.