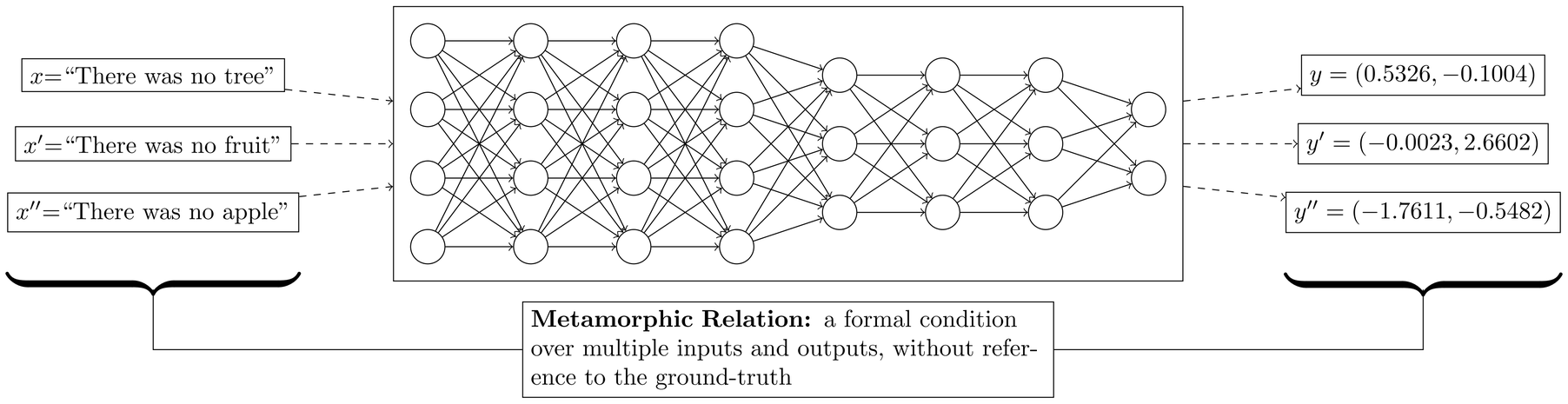
Generate **millions** of test cases from **few** **unlabelled** data with **metamorphic testing.**

Systematicity, Compositionality and Transitivity of Deep NLP Models: a Metamorphic Testing Perspective

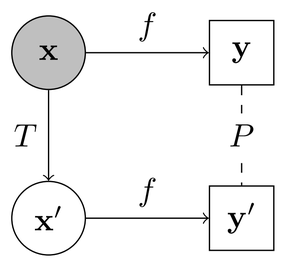
Edoardo Manino, Julia Rozanova, Danilo Carvalho, André Freitas, Lucas Cordeiro
University of Manchester (UK), Idiap Research Institute (CH), EnnCore project



Existing works: single-input metamorphic relations

Existing metamorphic testing for NLP:

- **One** input from test set
- Robustness-like relations
- $T \rightarrow$ typos, synonyms, etc.
- $P \rightarrow$ same output class



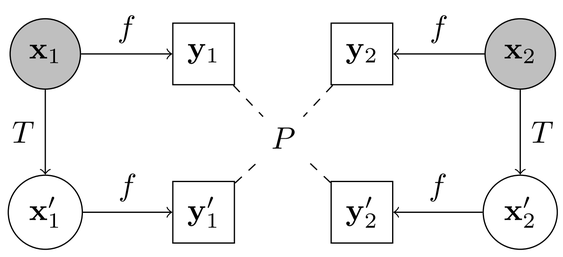
Is the output class preserved after replacing some input words with synonyms?

Single-input metamorphic test	
Input:	$\mathbf{x}$ = The cat sat on the mat.
	$\mathbf{x}'$ = The pet stood onto the mat.
T:	*replace any word of the input with a synonym.*
P:	$\mathbf{y} = f(\mathbf{x}) \wedge \exists i \, \forall j \neq i \, (y_i > y_j) \wedge (y'_i > y'_j)$

Contribution 1: pairwise systematicity relations

Test internal consistency of model:

- **Two** inputs from test set
- Read their output relation
- Is it preserved after applying T?
- 112M+ tests from 11K+ data
- RoBERTa sentiment: 5-10% errors



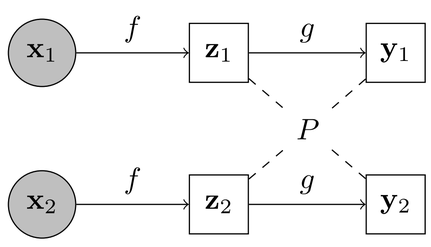
Is the polarity between the two sentences preserved after concatenation of the fragment?

Pairwise systematicity metamorphi	
Input:	$\mathbf{x}_1$ = Light, cute and forgettable.
	$\mathbf{x}_2$ = A masterpiece four years in the making.
	$\mathbf{x}'_1$ = Thank you. Light, cute and forgettable.
	$\mathbf{x}'_2$ = Thank you. A masterpiece four years in the making.
T:	*concatenate the text* Thank you. *at the beginning of the input.*
P:	$s_{pos}(f(\mathbf{x}_1)) > s_{pos}(f(\mathbf{x}_2)) \iff s_{pos}(f(\mathbf{x}'_1)) > s_{pos}(f(\mathbf{x}'_2))$

Contribution 2: pairwise compositionality relations

Metamorphic version of probing:

- **Two** inputs from test set
- Probe hidden reps. after f
- Does it correlate with output?
- 9M+ tests from fewer than 1K data
- RoBERTa entailment: 25-70% errors



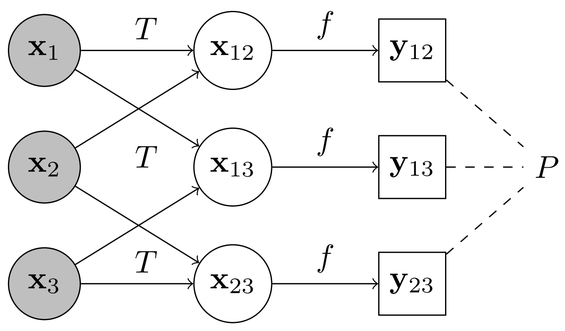
Does the polarity between the two embeddings correspond to the polarity of the output?

Pairwise compositionality metamorp	
Input:	$\mathbf{x}_1$ = There was no tree. There was no cherry tree.
	$\mathbf{x}_2$ = There was no fruit. There was no apple.
Hidden:	$f(\mathbf{x}_1)$ = *contextual embeddings of the tokens (* tree. cherry tree. *)*
	$f(\mathbf{x}_2)$ = *contextual embeddings of the tokens (* fruit. apple. *)*
P:	$s_{hyp}(f(\mathbf{x}_1)) > s_{hyp}(f(\mathbf{x}_2)) \iff s_{ent}(g(f(\mathbf{x}_1))) > s_{ent}(g(f(\mathbf{x}_2)))$

Contribution 3: three-way transitivity relations

Are mistakes transitive too?

- **Three** inputs from test set
- If two pairs are predicted true…
- …the third *must* be true too!
- Cubic number of test cases
- RoBERTa lex. rel.: 60-80% errors



When the model classifies two input pairs as positive, does it also classify the third as positive?

Three-way transitivity metamorphi	
Input:	$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ = arrangement symmetrical together
	$\mathbf{x}_{12}$ = (arrangement symmetrical)
	$\mathbf{x}_{23}$ = (symmetrical together)
	$\mathbf{x}_{13}$ = (arrangement together)
T:	*choose two words from the source triplet* $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$
P_{syn}:	$v_{syn}(f(\mathbf{x}_{12})) \wedge v_{syn}(f(\mathbf{x}_{23})) \implies v_{syn}(f(\mathbf{x}_{13}))$
P_{hyp}:	$v_{hyp}(f(\mathbf{x}_{12})) \wedge v_{hyp}(f(\mathbf{x}_{23})) \implies v_{hyp}(f(\mathbf{x}_{13}))$

Scan QR code to get the full paper

EnnCore