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)

This work is funded by the EPSRC grant EP/T026995/1 entitled "EnnCore: End-to-End Conceptual Guarding of Neural Architectures" under Security for all in an AI enabled society









Motivation

Trend towards learning from unlabelled data

- Unsupervised, semi-supervised, self-supervised
- No need for costly dataset annotation

Testing without ground-truth?

- Current paradigms need ground-truth annotations
- In-distribution testing: train-validate-test split
- More recent: out-of-distribution testing, probing

Metamorphic testing!

- Formal definition of input-output behaviour
- Checks whether the NLP model satisfies it
- ▶ Less reliance on ground-truth ⇒ large number of test cases



Existing metamorphic works for NLP

Table: Example of robustness relations from the literature [Li 2017]. Robustness relations belong to the class of single-input relations.

They all focus on the same simple structure

- Pick a single input x from the test set
- Apply transformation x' = T(x): e.g. typos, synonyms
- \triangleright Check that x, x' satisfy P: e.g. same class (robustness)

Contribution 1: pairwise systematicity

```
Pairwise systematicity metamorphic relations
x_1 = \begin{array}{c} \text{Light, cute and forgettable.} \\ \text{Input:} & x_2 = \begin{array}{c} \text{A masterpiece four years in the making.} \\ x_1' = \begin{array}{c} \text{Thank you.} \\ \text{A masterpiece four years in the making.} \\ \end{array}
x_2' = \begin{array}{c} \text{Thank you.} \\ \text{A masterpiece four years in the making.} \\ \text{A masterpiece four years in the making.} \\ \text{T:} & concatenate the text} \begin{array}{c} \text{Thank you.} \\ \text{Thank you.} \end{array} \text{ at the beginning of the input.} \\ \text{P:} & s_{pos}(f(x_1)) > s_{pos}(f(x_2)) \iff s_{pos}(f(x_1')) > s_{pos}(f(x_2')) \\ \end{array}
```

Table: Example of pairwise systematicity relations for sentiment analysis.

Let's test the internal consistency of an NLP model

- ightharpoonup Pick **two** unrelated inputs x_1, x_2 from the test set
- Read the relation between their outputs y₁, y₂
- Check whether it still holds after transforming both inputs



Contribution 2: pairwise compositionality

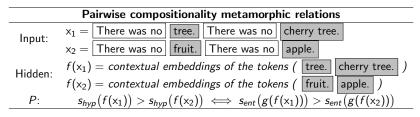


Table: Example of pairwise compositionality relations for NLI. Pairwise compositionality relations do not have a transformation T.

A metamorphic version of probing intermediate layers

- ▶ Think of the neural network as the composition of f and g
- Pick **two** unrelated inputs x_1, x_2 from the test set
- ▶ Read the relation between their embeddings $f(x_1), f(x_2)$
- Check whether the relation carries to the outputs y₁, y₂



Contribution 3: three-way transitivity

Table: Example of three-way transitivity relations for the lexical relations of synonymy and hypernymy.

Do NI P models make transitive errors?

- ▶ Pick **three** unrelated inputs x_1, x_2, x_3 from the test set
- ▶ Create all input pairs $x_{ij} = (x_i, x_j)$ with boolean output $v(y_{ij})$
- ▶ Check whether $v(y_{12}) \land v(y_{23}) = \top$ always implies $v(y_{13}) = \top$

Empirical results

Number of metamorphic test cases we can generate

- ▶ Pair. system.: quadratic (112M+ from 11K+ unlabelled set)
- ▶ Pair. compos.: quadratic (9M+ from less than 1K set)
- 3-way transitivity: cubic (we had to subsample them)

Empirical results on state-of-the-art RoBERTa model

- ▶ Pairwise systematicity: from 5% to 10% violations
- Pairwise compositionality: from 25% to 70% violations
- ► Three-way transitivity: from 60% to 80% violations

Final remarks

- ▶ Metamorphic testing does **not** replace traditional testing
- ▶ It complements it by checking the internal consistency

