Similarity, Transformation and the Newly Found Invariance of Influence Functions

Andrew Liu, Gerald Penn
University of Toronto, Department of Computer Science

UNIVERSITY OF TORONTO

https://github.com/Andrwyl/SimilarityTransformationsInfluence

Summary

- Influence functions have re-emerged as a feasible way to assign blame to datapoints. **We find influence scores can be something more.**
- Influence functions exhibit more properties that we would expect from a *true* encoding of semantics.

Methods

Sentence-BERT Generates vector representations of semantics.
Influence Functions Score the influence of a sentence on generating another sentence.

Experimental Setup

Grammatical Transformations

Baseline	Alexander conquered Persia.
Passivization	Persia was conquered by Alexander.
Clefting	It was Persia that Alexander conquered.
Topicalization	Persia, Alexander Conquered.
VP-Topicalization	Conquered Persia, Alexander did.

Grammatical Transformation Dataset

50 hand-crafted factual SVO sentences and their transformations (250 total).

Made-Up Entity Dataset

A smaller dataset that contains the same SVO sentences where the subjects are made-up nonsense names.

Problem Description

Motivation

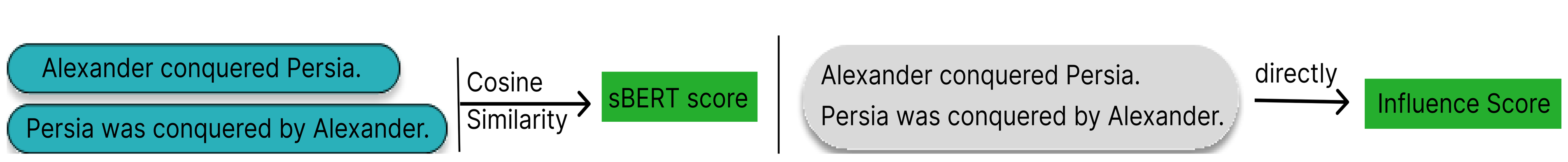
For a true semantic model, its behavior on a semantic task should be **indistinguishable between any of the four transformations.**

For the subsequent figures, a turquoise box with text represents the sBERT embedding for that text
A grey box with text represents the influence of the bottom text on the top text

Sentence Similarity

We examine the similarity between the baseline and each of its transformations.

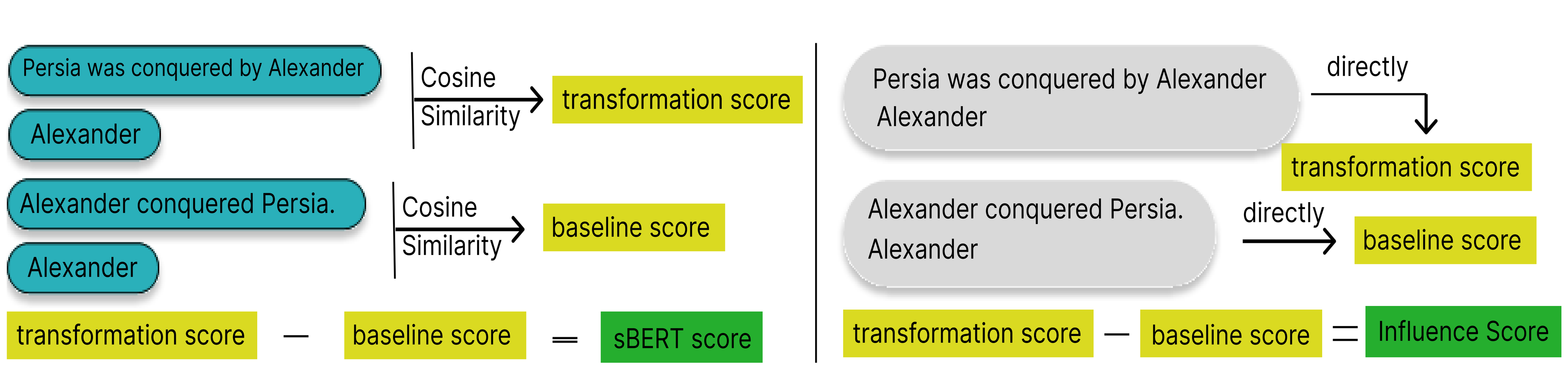
- How similar is *Alexander conquered Persia* to *Persia was conquered by Alexander*?



Entity Invariance

We examine the *similarity* between the subject and the base sentence relative to the *similarity* between the deep subject and the transformed sentence.

- *Alexander* relative to *Alexander conquered Persia* versus *Alexander* relative to *Persia was conquered by Alexander*



Findings

Influence Captures Semantics

Influence and s-BERT scores are tightly correlated. **Pearson** correlation is **0.9326** with a p-value of **2.62×10^{-178}**

Significance of Grammatical Transformations

Friedman tests indicate that

- In sentence similarity, both metrics are **biased** towards some transformation
- In entity invariance, influence exhibits **no significance**

In the entity invariance task, **behaviour of influence functions is indistinguishable between the different transformations**

	Sentence Similarity	Entity Invariance
sBERT	2.32×10^{-15}	3.97×10^{-7}
Influence	1.69×10^{-15}	0.983

	Sentence Similarity	Entity Invariance
sBERT	71.23 / 1.1936	32.57 / 0.807
Influence	71.88 / 1.199	0.17 / *0.058

Trends within the Transformations

	Passivization	Clefting	Topicalization	VP-Topic
Influence on Sentence Similarity	1.570795571	1.570795892	1.570796019	1.570796057
sBERT on Sentence Similarity	0.937302351	0.9078437984	0.8857396245	0.8987811208
sBERT on Entity Invariance	-0.05444133282	-0.08711430431	-0.05307358504	-0.07616019249

Above are medians of the scores of the different transformations for statistically significant configs. For each task and model, darker squares correspond to a transformation that is *more* similar/*more* variant.

Made-Up Dataset Results

	p-values	Test Statistics	Effect Size
Sentence Similarity	3.79×10^{-14}	65.568	1.145
Entity Invariance	0.008	11.712	0.484

Repeating the influence functions experiments on the made-up dataset, entity invariance is no longer indifferent across the different transformations.

Discussion

- Influence functions are good for more than just assigning blame, at minimum they correlate tightly with sBERT embeddings
- Under a higher-order semantic task, influence functions exhibit a robustness towards transformations.
- Using made-up subjects alters the behaviour of influence functions on the invariance task

We conclude that influence functions may be a step towards handling semantic meaning rather than just surface level aspects of their syntactic realizations

SCiL SOCIETY FOR COMPUTATION IN LINGUISTICS