Towards preserving word order importance through FORCED INVALIDATION

Hadeel Al-Negheimish[1], Pranava Madhyastha[2,1], Alessandra Russo[1]

[1]Imperial College London, [2]City, University of London

Summary

- **Order of words in a sentence is important** in natural languages such as English, but previous work has shown a **surprising lack of sensitivity to word order** in BERT-based Masked Language Models
- We propose a simple yet general approach to preserve the importance of word order and show that it is **effective on a variety of English NLU and QA tasks**.

Word order matters.. but not for the models

Given the following shuffled question:

red more, balls cubes Which do or we *shiny* have blue and?

..what was the intended question?

Which do we have more, red balls or *shiny* and blue cubes?

Which do we have more, red cubes or *shiny* and blue balls?

Which do we have more, *shiny* red or blue balls and cubes?

- Shuffling destroys syntax and semantics of a sentence
- However, recent work has showed that masked language models **suffer from a catastrophic lack of sensitivity to word order**.
- Models give the same predictions for well-ordered and shuffled input in Natural Language Understanding tasks

Our Solution: FORCED INVALIDATION

Main Idea: Instead of predicting the same value for a shuffled and well-ordered input, models should explicitly label shuffled input as "INVALID"

How?

1) Augmenting training data with {1,2,3}-gram permutation samples labelled with **invalid** as the additional label

2) Modifying models to account for the new label and training them in the standard setting with a combination of standard training examples and the augmented invalid samples

Advantages

- Simple and general approach
- Easily applied to multiple tasks, not just classification
- Requires minimal changes to model architecture
- We show that it drastically improves word order importance on multiple tasks, without compromising original performance

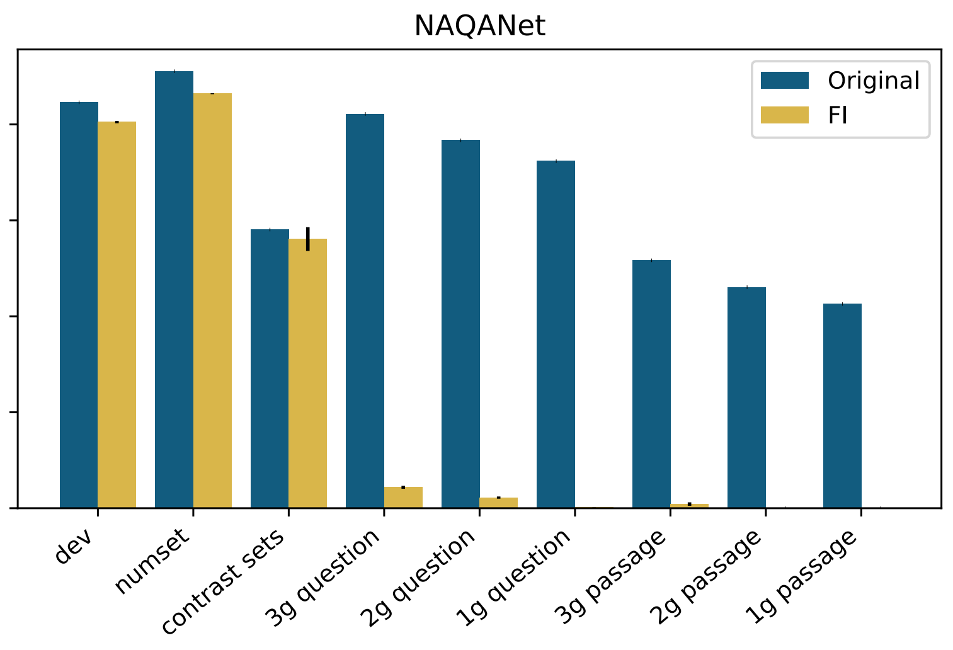
We evaluate FI on two classes of data: unperturbed dev set, and {1,2,3}-gram shuffles of that data

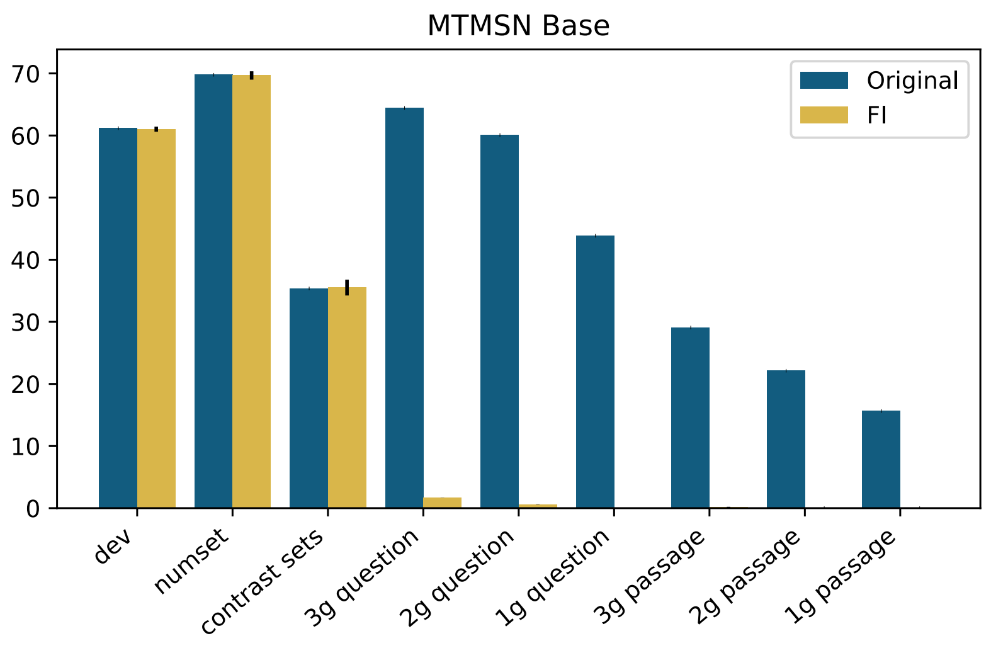
EACL 2023

Results

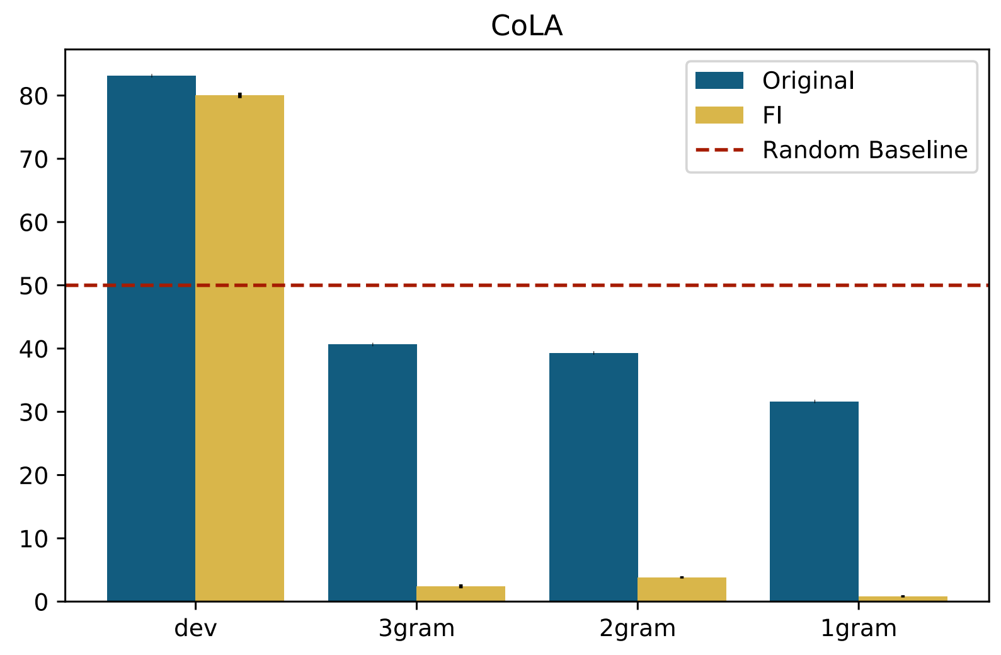
Unconstrained Question Answering

- We apply FI for two-module-based models for DROP, an MRC dataset that requires complex reasoning.
- FI makes models **more sensitive** to word order, while **retaining performance on well-ordered data**





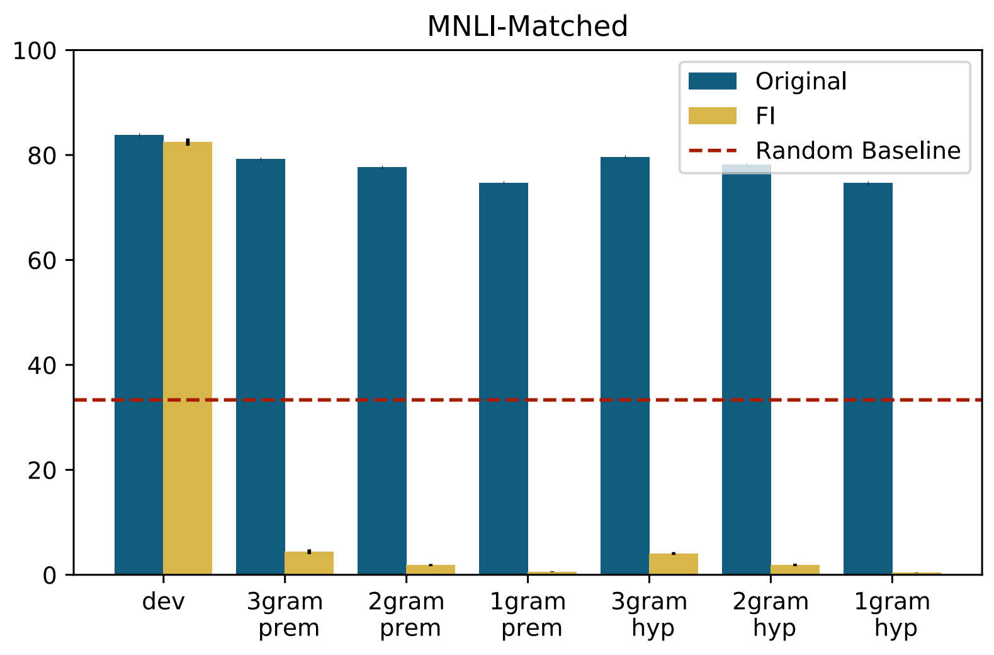
Grammatical Acceptability

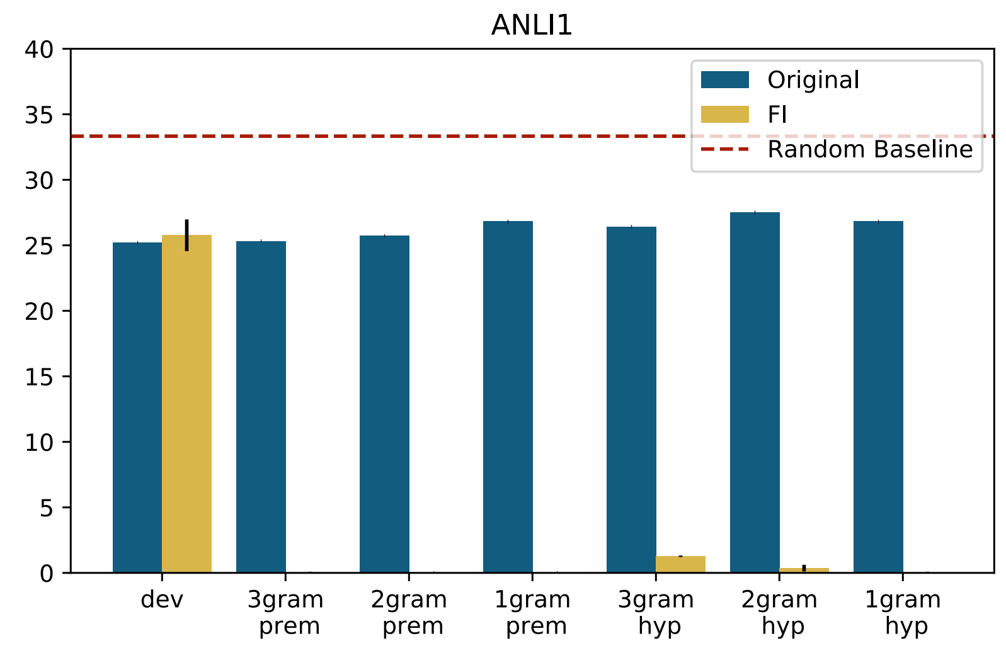


- Despite shuffled sentences being grammatically incorrect, the **original model predicts 'acceptable'** for many shuffles
- FI forces models to learn to flag shuffled sentences as invalid

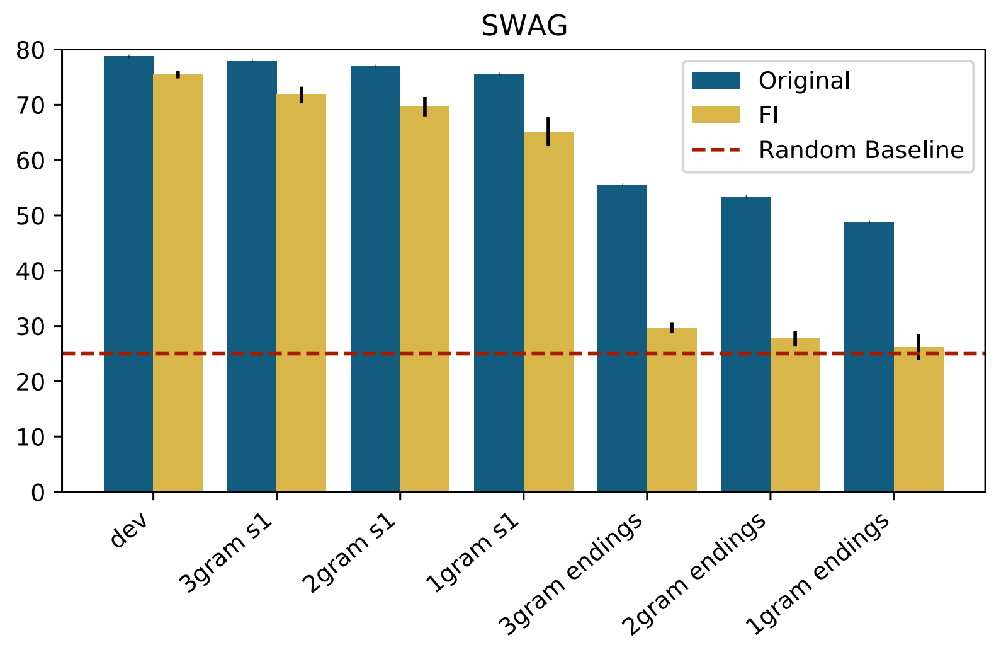
Natural Language Inference

- NLI has been one of the most important testbeds of previous work showing BERT insensitivity to word order
- FI **improves word-order sensitivity** on the dataset it has been trained on, and generalizes to similar unseen tasks, like ANLI
- We also find that FI makes models **more robust to shortcuts** in HANS





Multiple Choice Commonsense Reasoning



- Instead of changing the architecture of an MCQ model, **only the data is changed**
- In this case, we expect models to get random performance for shuffled input, which FI does for shuffled endings