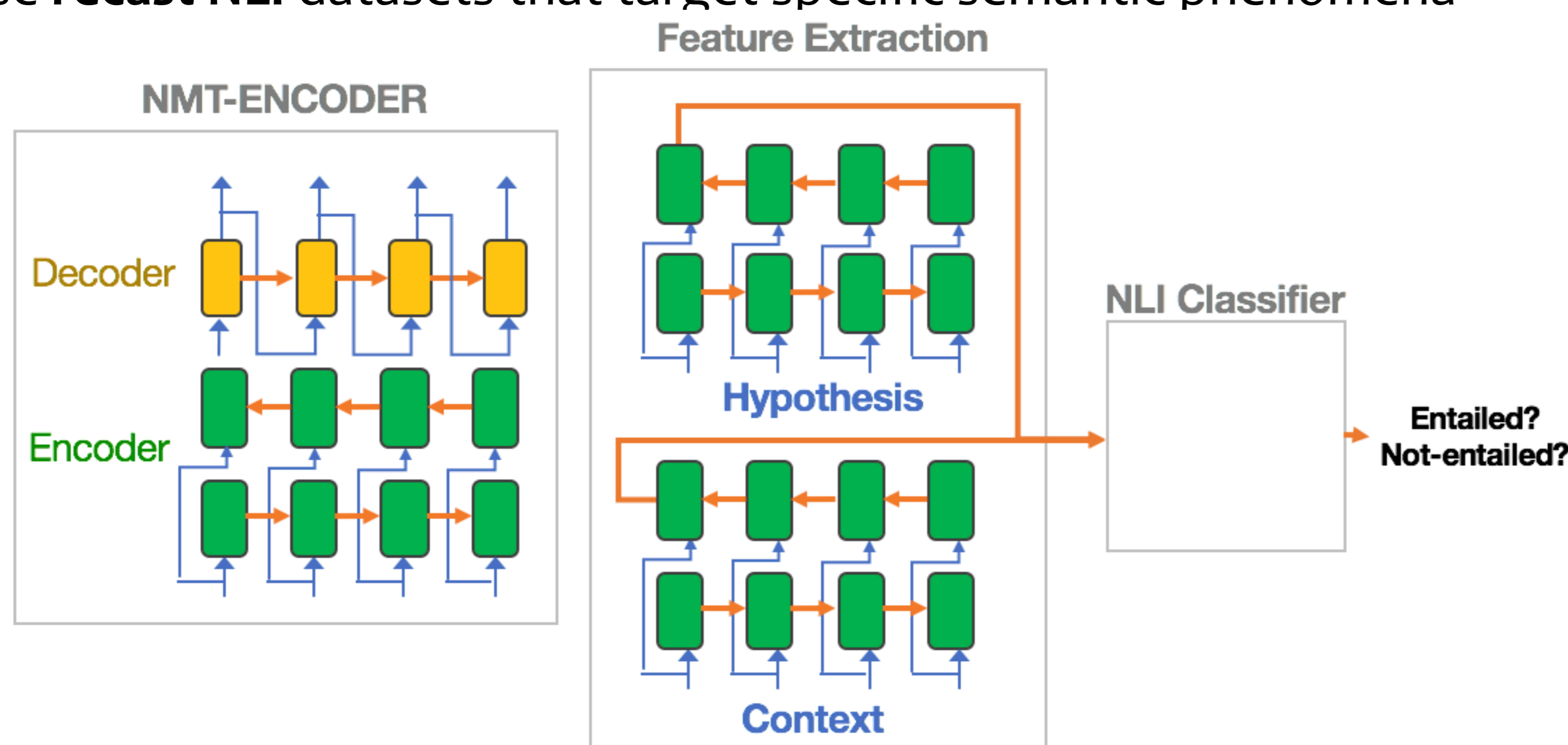


Adam Poliak, Yonatan Belinkov, James Glass, Benjamin Van Durme

How to evaluate semantics in NMT models?

Method:

- Pre-train NMT models:
 - English -> {Arabic, Spanish, Chinese, German}
- Extract sentence representations from pre-trained NMT encoders
- Use **recast NLI** datasets that target specific semantic phenomena



Recasting: convert semantic annotations into labeled NLI

Paraphrastic Inference:

Premise: Iran possesses five research reactors
 Hypothesis: Iran has five research reactors
entailed

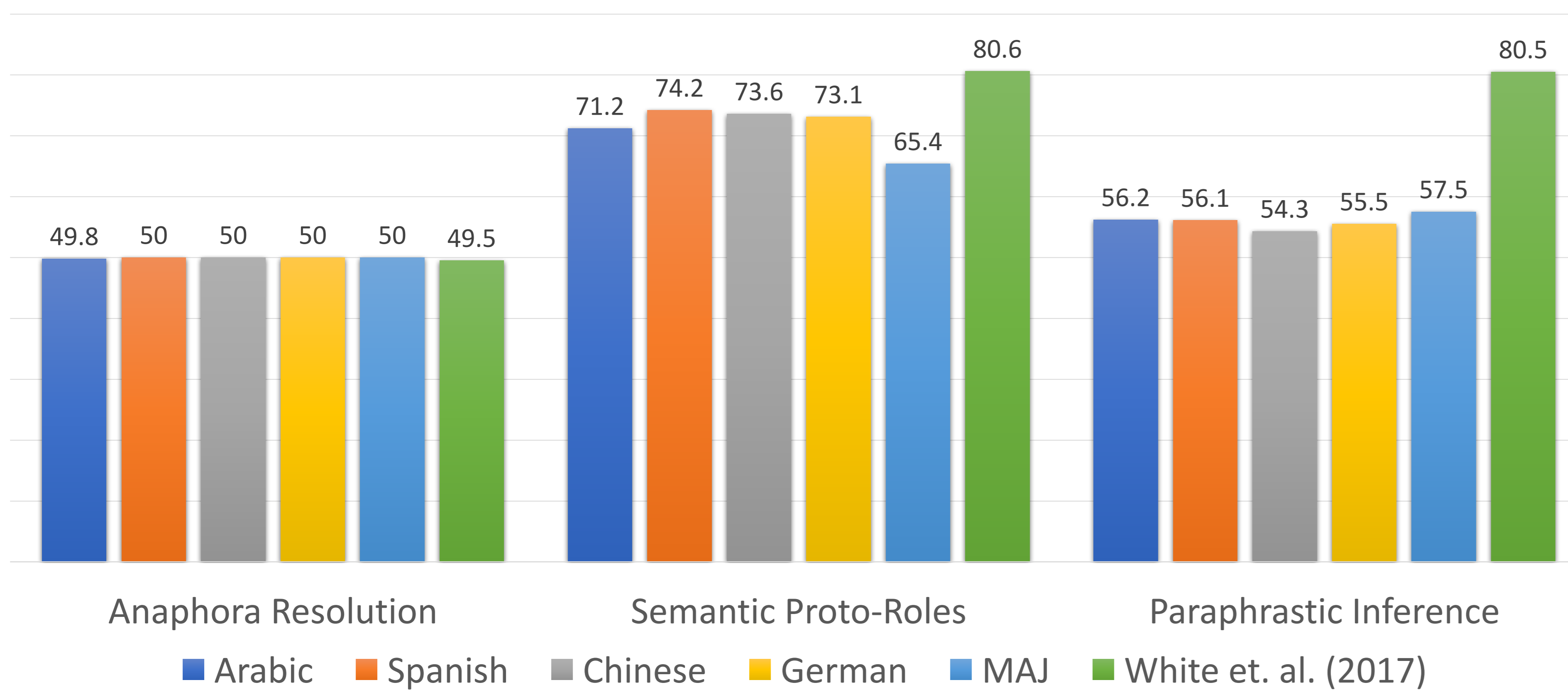
Anaphora Resolution:

Premise: Sara adopted Jill, she wanted a child
 Hypothesis: Sara adopted Jill, Jill wanted a child
not-entailed

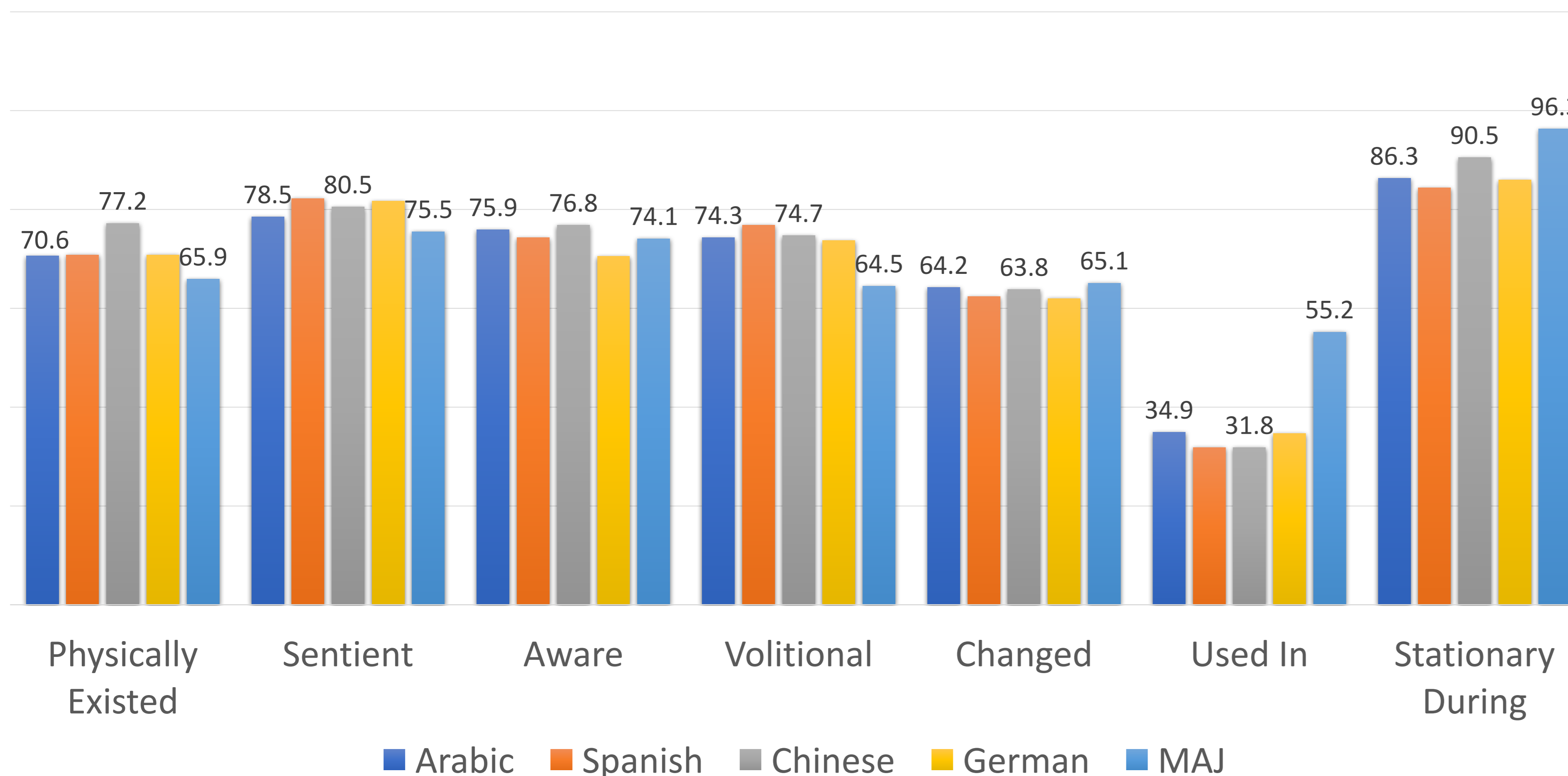
Semantic Proto-Roles:

Premise: Berry rejoins WPP Group
 Hypothesis: Berry was sentient
entailed

NLI Accuracies across Semantic Phenomena



NLI Accuracies across Semantic Proto-Roles



- Captures SPR better than Anaphora Resolution or Paraphrastic Inference
- Performs better on *Proto-Agent* roles compared to *Proto-Patient* roles

Motivating Semantic Phenomena & Translation

Paraphrastic Inference

Map unknown words to paraphrases

Anaphora Resolution

EN: "The parent fed the child since *she* was hungry"

"the child" as *la niña* and not *el niño*

Semantic Proto-Roles

proto-roles may be expressed differently in different languages

Conclusion

- NMT encoders may not capture these semantic phenomena
- NMT train data may not require these types of reasoning
- Test for more semantic phenomena in NMT by recasting more datasets
- Evaluation method may be used to test encoders trained for other NLP tasks

Paper



Code/Data

