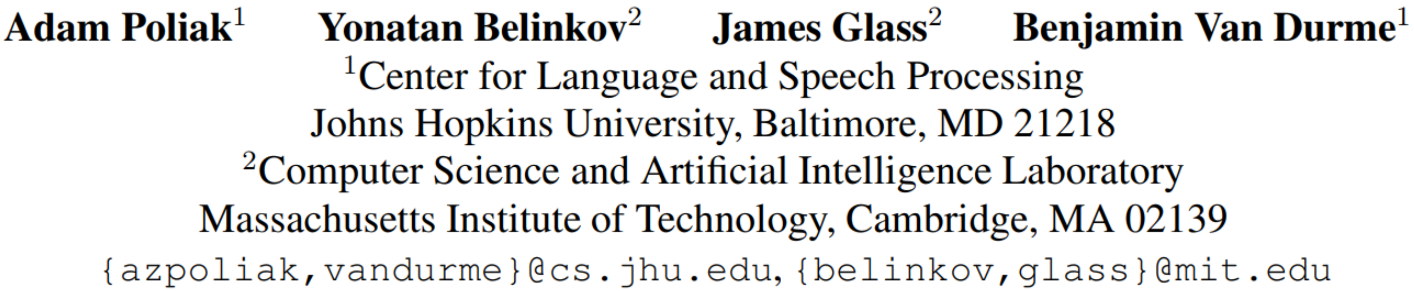
On the Evaluation of Semantic Phenomena in Neural Machine Translation Using Natural Language Inference



Adam Poliak[1] **Yonatan Belinkov**[2] **James Glass**[2] **Benjamin Van Durme**[1]
[1]Center for Language and Speech Processing
Johns Hopkins University, Baltimore, MD 21218
[2]Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology, Cambridge, MA 02139
{azpoliak,vandurme}@cs.jhu.edu, {belinkov,glass}@mit.edu

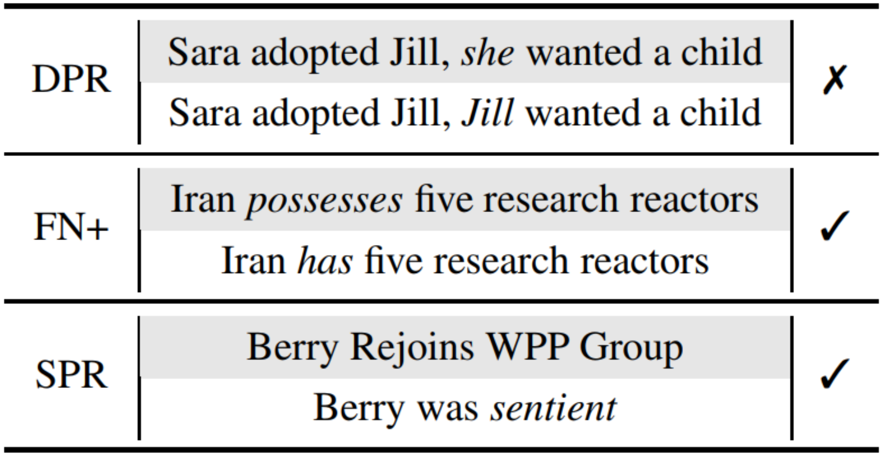
*in NAACL 2018 (short)

Motivation

- NMT’s encoder has learnt sentence representation/meaning.
- Question: how specific semantic phenomena are captured in NMT representations?
- This paper: use recast Natural Language Inference (NLI) to determine whether NMT encoders capture distinct semantic phenomena.

Recast NLI

- NLI: train a classifier to determine if one sentence is supported (entailed) by another.
- White et al. (2017)* propose to recast three semantic phenomena as NLI, and build three NLI datasets.



DPR	Sara adopted Jill, *she* wanted a child Sara adopted Jill, *Jill* wanted a child	✗
FN+	Iran *possesses* five research reactors Iran *has* five research reactors	✓
SPR	Berry Rejoins WPP Group Berry was *sentient*	✓

Figure 1: Example sentence pairs for the different semantic phenomena. DPR deals with complex anaphora resolution, FN+ is concerned with paraphrastic inference, and SPR covers Reisinger et al. (2015)'s semantic proto-roles. ✓ / ✗ indicates that the first sentence entails / does not entail the second.

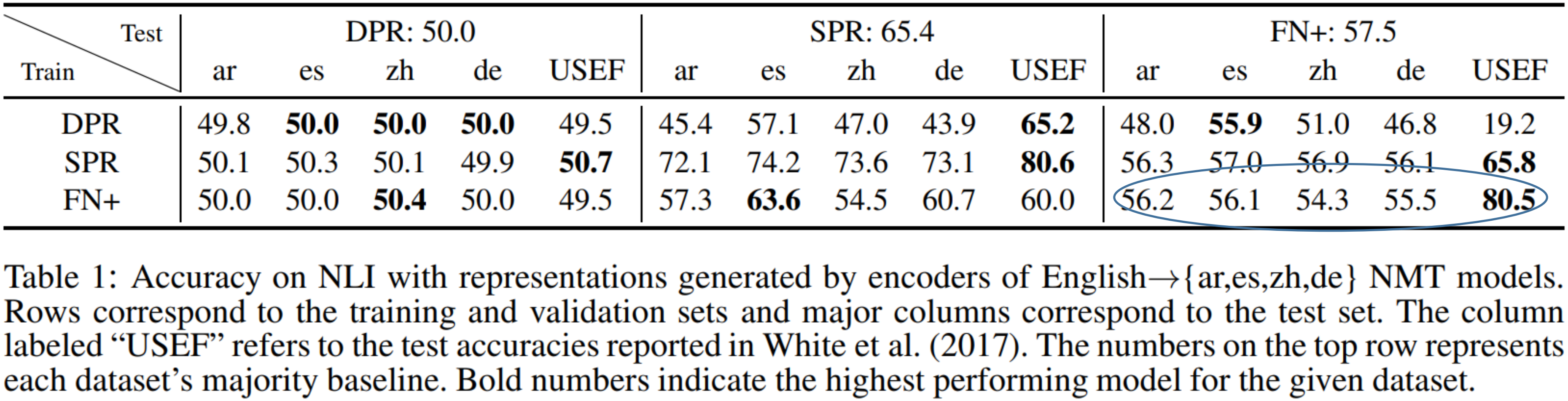
*White et. al, "Inference is Everything: Recasting Semantic Resources into a Unified Evaluation Framework", in IJCNLP 2017

Method

- Train Bi-LSTM with attention on several parallel corpus.
  - English → {Arabic (ar), Spanish (es), Chinese (zh), and German (de)}
- Pass each sentence through the trained NMT encoder to extract respective vector representation.
- Feed the sentence pair representation into a classifier with a softmax layer (2/3 labels).

Results

- Main results:



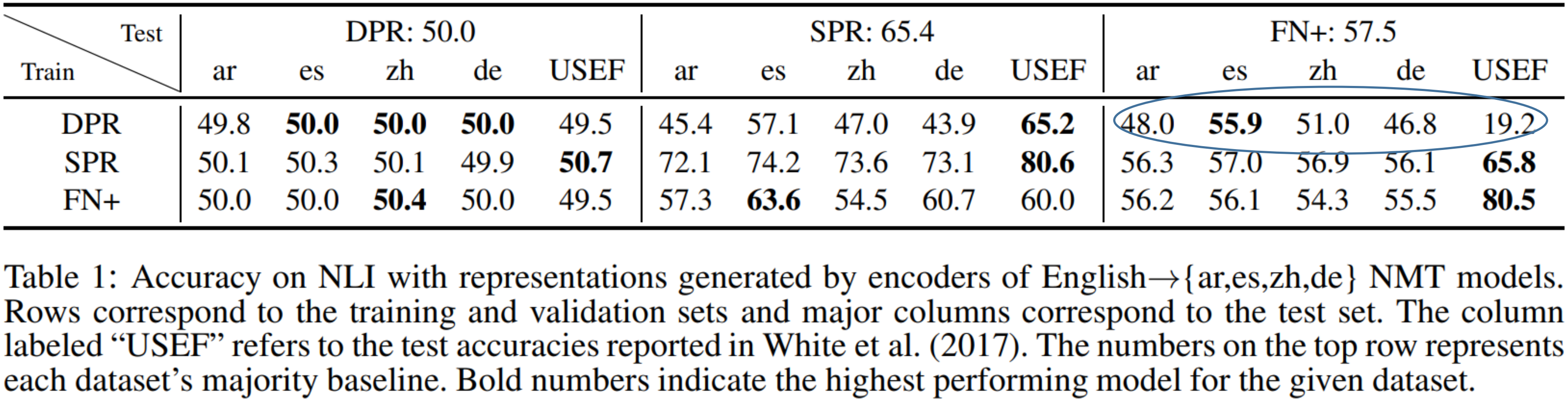
Test / Train	DPR: 50.0 ar	es	zh	de	USEF	SPR: 65.4 ar	es	zh	de	USEF	FN+: 57.5 ar	es	zh	de	USEF
DPR	49.8	**50.0**	**50.0**	**50.0**	49.5	45.4	57.1	47.0	43.9	**65.2**	48.0	**55.9**	51.0	46.8	19.2
SPR	50.1	50.3	50.1	49.9	**50.7**	72.1	74.2	73.6	73.1	**80.6**	56.3	57.0	56.9	56.1	**65.8**
FN+	50.0	50.0	**50.4**	50.0	49.5	57.3	**63.6**	54.5	60.7	60.0	56.2	56.1	54.3	55.5	**80.5**

Table 1: Accuracy on NLI with representations generated by encoders of English→{ar,es,zh,de} NMT models. Rows correspond to the training and validation sets and major columns correspond to the test set. The column labeled "USEF" refers to the test accuracies reported in White et al. (2017). The numbers on the top row represents each dataset's majority baseline. Bold numbers indicate the highest performing model for the given dataset.

- Paraphrastic entailment (FN+)
  Poor performance
  FN+ might not be an ideal dataset.

Results

- Main results:



Test / Train	DPR: 50.0					SPR: 65.4					FN+: 57.5				
	ar	es	zh	de	USEF	ar	es	zh	de	USEF	ar	es	zh	de	USEF
DPR	49.8	**50.0**	**50.0**	**50.0**	49.5	45.4	57.1	47.0	43.9	**65.2**	48.0	**55.9**	51.0	46.8	19.2
SPR	50.1	50.3	50.1	49.9	**50.7**	72.1	74.2	73.6	73.1	**80.6**	56.3	57.0	56.9	56.1	**65.8**
FN+	50.0	50.0	**50.4**	50.0	49.5	57.3	**63.6**	54.5	60.7	60.0	56.2	56.1	54.3	55.5	**80.5**

Table 1: Accuracy on NLI with representations generated by encoders of English→{ar,es,zh,de} NMT models. Rows correspond to the training and validation sets and major columns correspond to the test set. The column labeled "USEF" refers to the test accuracies reported in White et al. (2017). The numbers on the top row represents each dataset's majority baseline. Bold numbers indicate the highest performing model for the given dataset.

- Anaphora entailment (DPR)

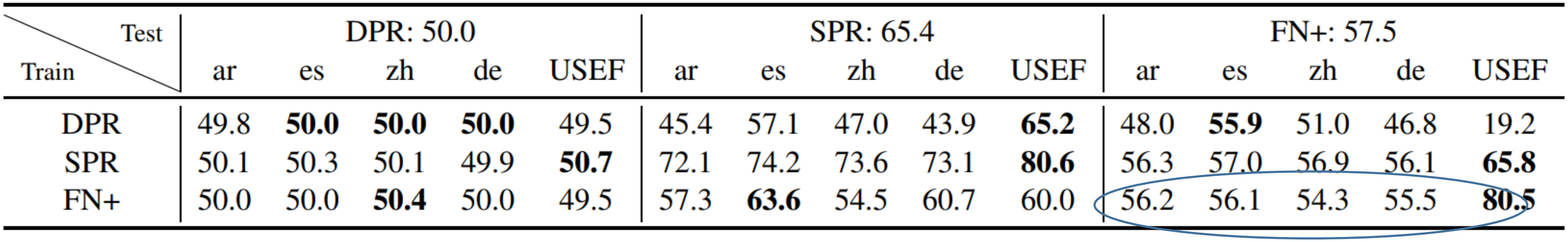
  Poor performance

  DPR tests whether a model contains common sense knowledge.

  Good performance when trained on DPR for paraphrastic entailment.

Results

- Main results:



Test Train	DPR: 50.0					SPR: 65.4					FN+: 57.5				
	ar	es	zh	de	USEF	ar	es	zh	de	USEF	ar	es	zh	de	USEF
DPR	49.8	**50.0**	**50.0**	**50.0**	49.5	45.4	57.1	47.0	43.9	**65.2**	48.0	**55.9**	51.0	46.8	19.2
SPR	50.1	50.3	50.1	49.9	**50.7**	72.1	74.2	73.6	73.1	**80.6**	56.3	57.0	56.9	56.1	**65.8**
FN+	50.0	50.0	**50.4**	50.0	49.5	57.3	**63.6**	54.5	60.7	60.0	56.2	56.1	54.3	55.5	**80.5**

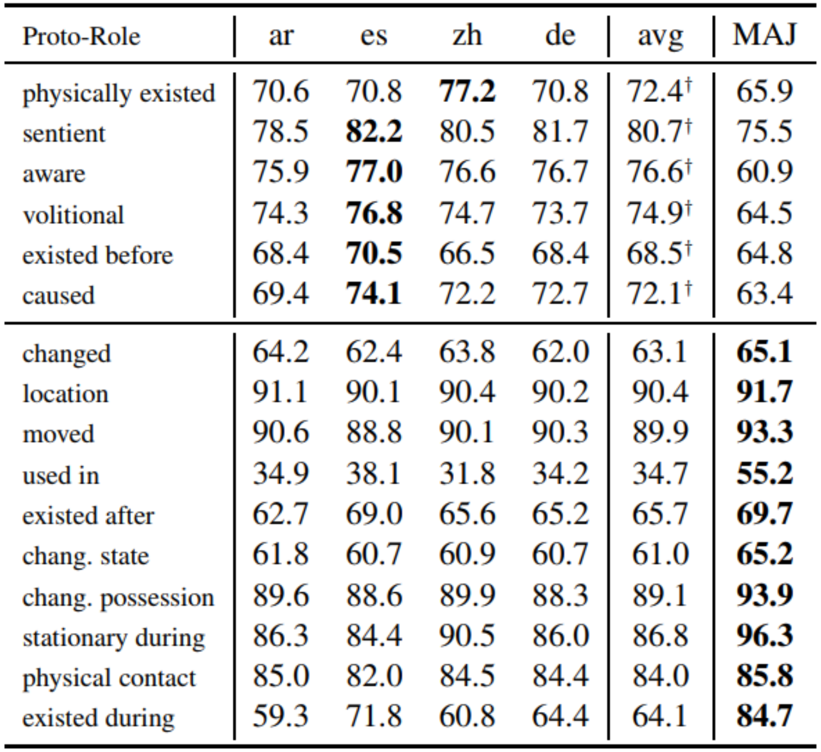
- Proto-role entailment (SPR)

  Outperform the majority baseline but are below USEF.

- Accuracies for each proto-role:

  The 6 proto-roles are more associated with proto-agents than proto-patients.

- Target Language: en-es performs better.



Proto-Role	ar	es	zh	de	avg	MAJ
physically existed	70.6	70.8	**77.2**	70.8	72.4†	65.9
sentient	78.5	**82.2**	80.5	81.7	80.7†	75.5
aware	75.9	**77.0**	76.6	76.7	76.6†	60.9
volitional	74.3	**76.8**	74.7	73.7	74.9†	64.5
existed before	68.4	**70.5**	66.5	68.4	68.5†	64.8
caused	69.4	**74.1**	72.2	72.7	72.1†	63.4
changed	64.2	62.4	63.8	62.0	63.1	**65.1**
location	91.1	90.1	90.4	90.2	90.4	**91.7**
moved	90.6	88.8	90.1	90.3	89.9	**93.3**
used in	34.9	38.1	31.8	34.2	34.7	**55.2**
existed after	62.7	69.0	65.6	65.2	65.7	**69.7**
chang. state	61.8	60.7	60.9	60.7	61.0	**65.2**
chang. possession	89.6	88.6	89.9	88.3	89.1	**93.9**
stationary during	86.3	84.4	90.5	86.0	86.8	**96.3**
physical contact	85.0	82.0	84.5	84.4	84.0	**85.8**
existed during	59.3	71.8	60.8	64.4	64.1	**84.7**

Summarization

- Inspected whether distinct types of semantics are captured by NMT encoders.
- NMT encoders might learn the most about semantic proto-roles, do not focus on anaphora resolution, and may poorly capture paraphrastic inference.
- The target-side language affects how well an NMT encoder captures these semantic phenomena.