Measuring Alignment Bias in Neural Seq2Seq Semantic Parsers

***SEM 2022**

Davide Locatelli and Ariadna Quattoni
Polytechnic University of Catalonia, Barcelona

Motivation

Why alignments?

- Grammar-based parsers modeled NL-MR alignments explicitly
- Neural seq2seq parsers rely on attention to automatically learn them
- **Can attention-based seq2seq parsers handle arbitrary alignments?**
- We introduce GEOALIGNED : GEOQUERY with gold alignment annotations

Alignments

Example 1

INPUT	TARGET
The	Die
cat	Katze
is	ist
on	auf
the	dem
table	Tisch

Alignments

Example 2

INPUT	TARGET
The	Il
cat	gatto
is	è
on	sul
the	ε
table	tavolo

Alignments

Monotonic vs Non-monotonic

- If the words can be aligned linearly the alignment is monotonic
- Non-monotonic alignments allow for reordering

MONOTONIC

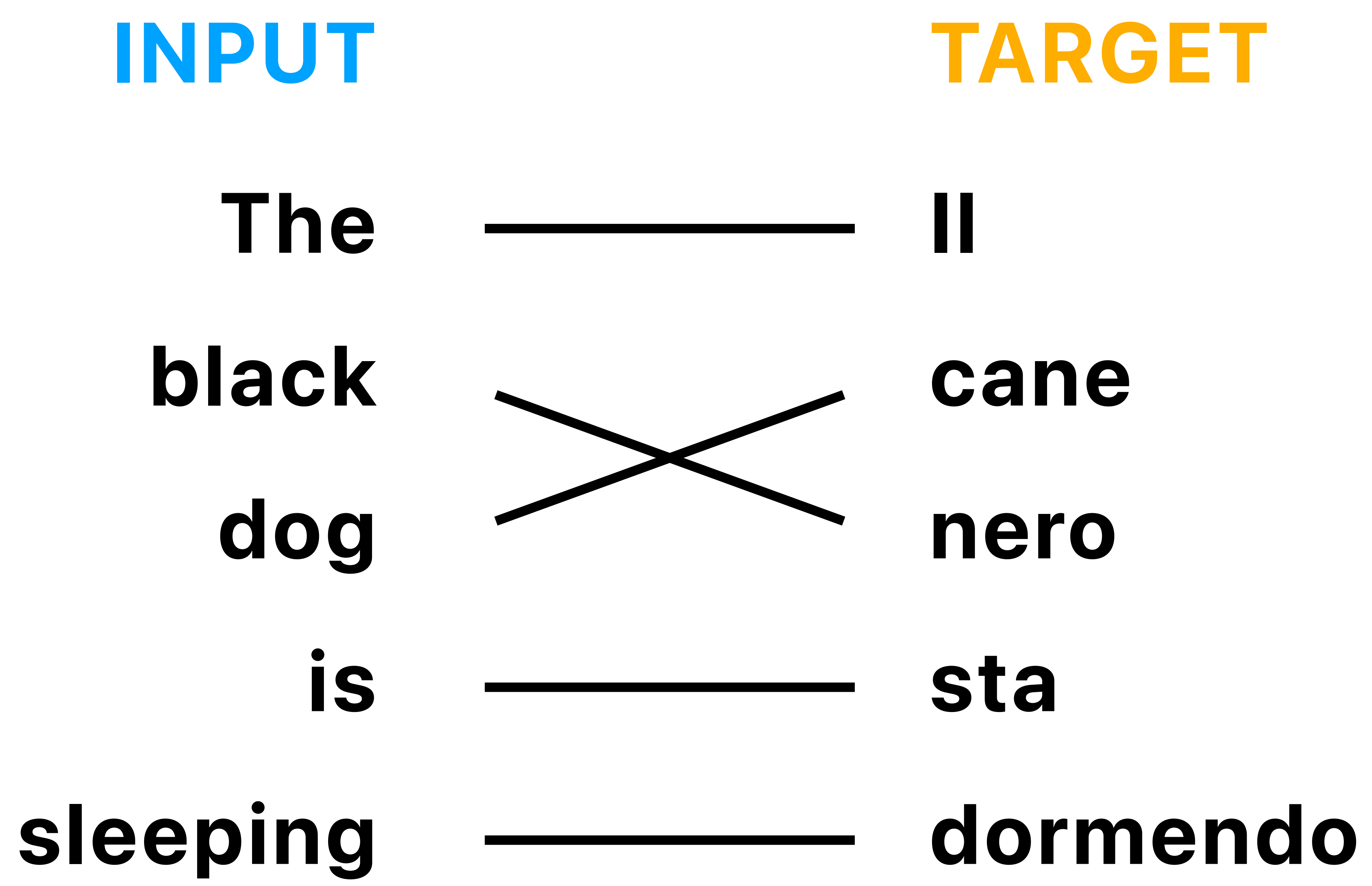
INPUT		TARGET
The	—	Der
black	—	schwarze
dog	—	Hund
is	—	ε
sleeping	—	schläft

Alignments

Monotonic vs Non-monotonic

- If the words can be aligned linearly the alignment is monotonic
- Non-monotonic alignments allow for reordering

NON-MONOTONIC



GeoQuery

Original data

- 880 English questions about US geography paired with MRs
- Several formalisms possible: FOL, SQL, variable-free functions
- Multilingual version includes German, Chinese, Indonesian, Swedish, etc

```
NL: where is Mount Rainier?

MR: answer ( loc ( place ( place_id ( mount_rainier ) ) ) ) )
```

GEOALIGNED
Annotation

- 4 annotators were asked : *for each NL-MR pair in GEOQUERY:*
  1. *Is there a monotonic or non-monotonic alignment?*
  2. *Provide the alignment from the NL to the MR*
- Inter-annotator agreement : 0.83 Cohen's Kappa Statistic
- Disagreement resolution : keep alignment that best matched majority
- 2 native speakers provided new Italian translations of the original GEOQUERY

GEOALIGNED

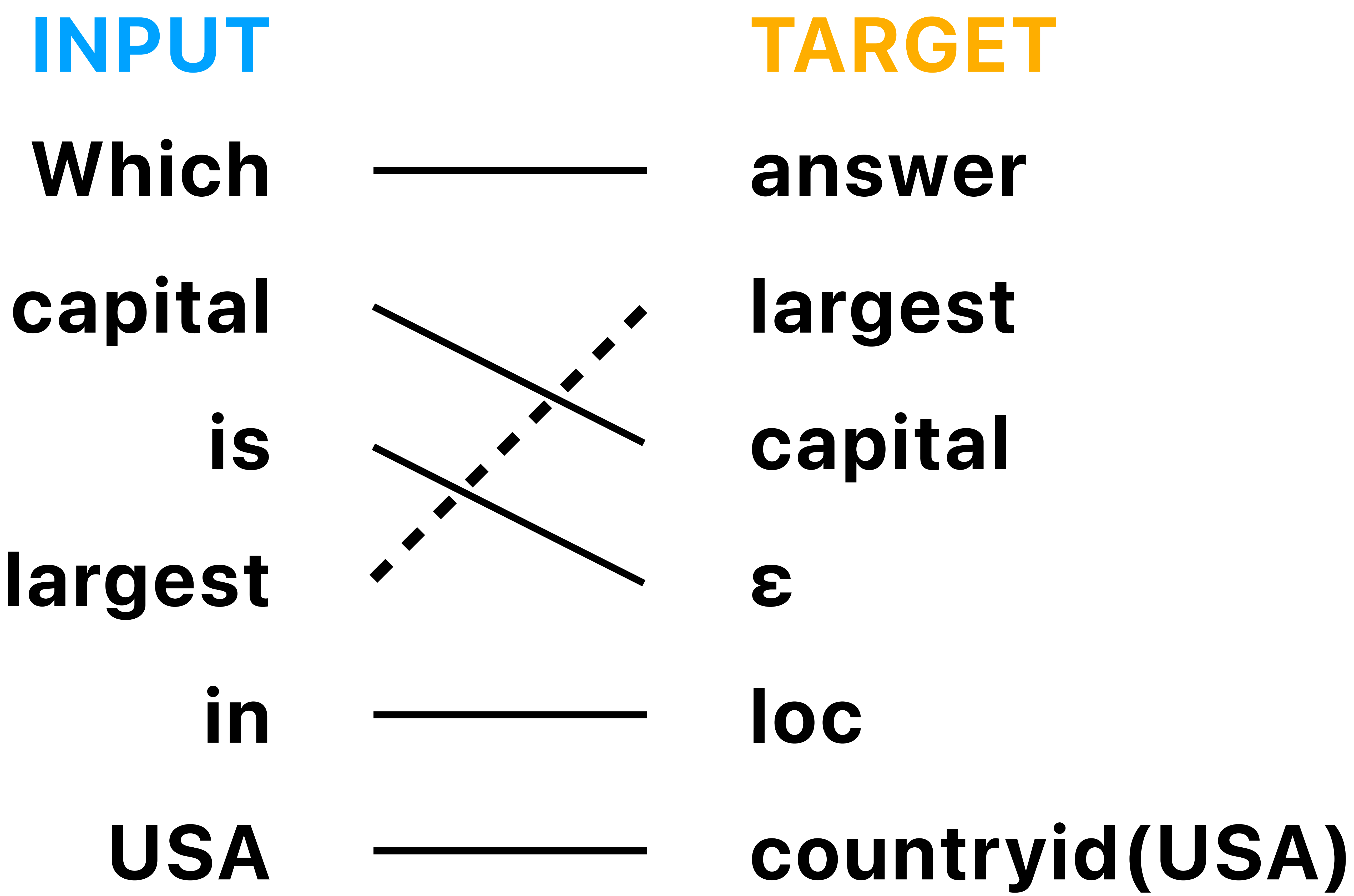
Example 1

MONOTONIC

INPUT		TARGET
Which	—	answer
rivers	—	river
are	—	ε
in	—	loc
Georgia	—	stateid(Georgia)

GEOALIGNED
Example 2

NON-MONOTONIC



GeoAligned
Statistics

Lang	Len	MP	MG	M0	NMR
EN	7.67	0.75	2.52	8.2	2.14
DE	7.72	0.65	2.91	0.55	2.52
IT	7.92	0.52	2.54	1.5	2.23

Table 1: Alignment annotation statistics for different languages. Len is the mean length of input NL sentences, MP is the percentage of monotonic alignments, MG is the average gap in monotonic alignments, M0 is the percentage of monotonic alignments with no gap, and NMR is the average number of words reordered in the non-monotonic alignments.

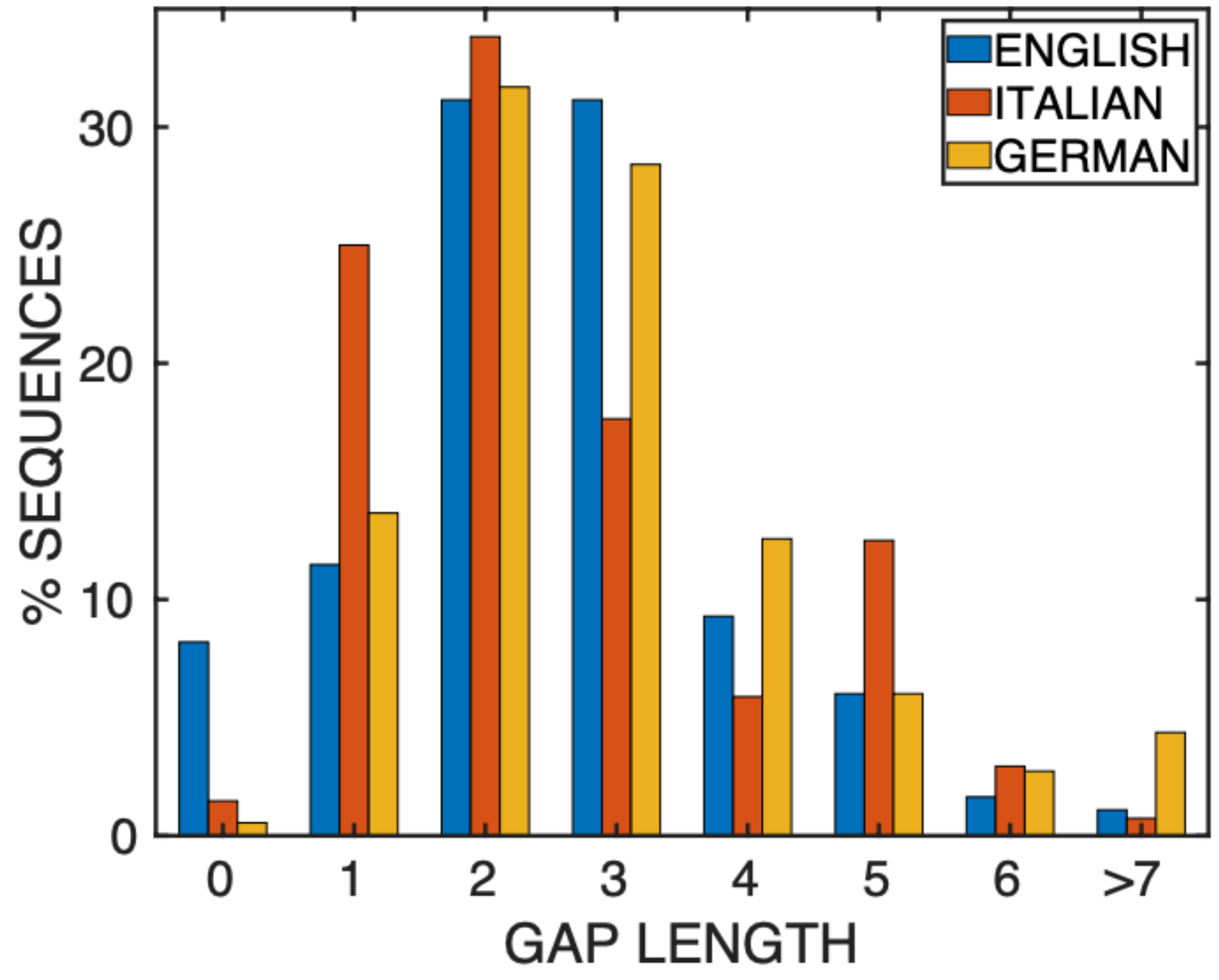


Figure 2: Distribution of gap lengths for the monotonic alignments.

Experiments

Models

- **LSTM Seq2Seq**
  - Bidirectional LSTM encoder
  - Unidirectional LSTM decoder with attention
    - Attention mechanism is then ablated
- **BART Pre-trained Seq2Seq**
  - Bidirectional encoder
  - Left-to-right decoder

Experiments

Results

Lang	Model	Acc	MAcc	NMAcc
En	Lstm	0.83	0.87	0.74
	Lstm-attn	0.75	0.80	0.61
	BART	0.85	0.87	0.80
De	Lstm	0.63	0.73	0.54
	Lstm-attn	0.57	0.69	0.46
It	Lstm	0.77	0.84	0.71
	Lstm-attn	0.71	0.80	0.63

Table 2: Summary of results for the different models and languages: LSTM is the seq2seq model based on a bidirectional LSTM encoder and an LSTM decoder with attention. LSTM-attn ablates the attention layer in the decoder. Acc reports the overall accuracy for each model, MAcc and NMAcc are the accuracy over sequences with monotonic and non-monotonic alignments respectively.

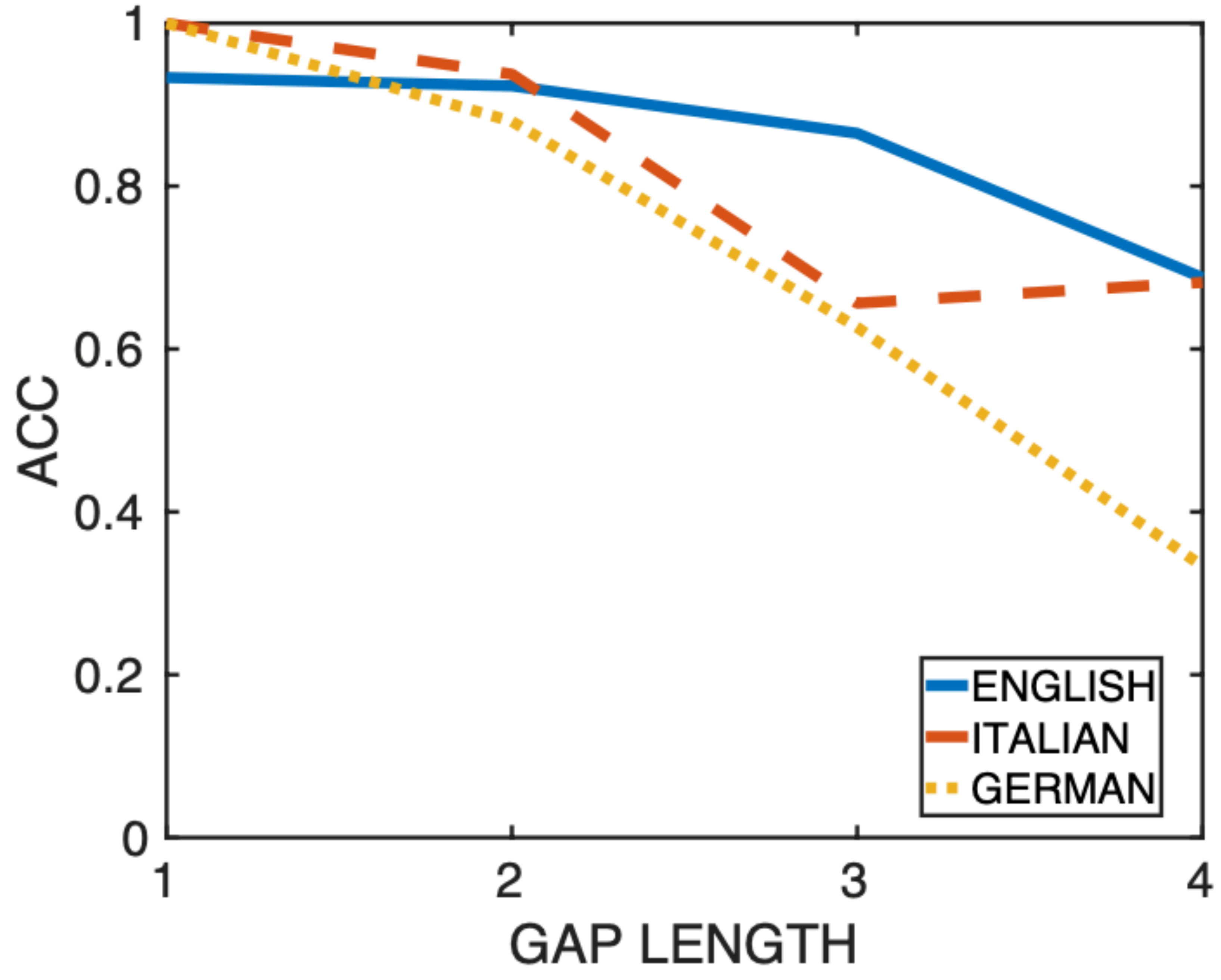


Figure 3: Accuracy for monotonic examples as a function of gap length.

Experiments

Analysis

Lang	Align	Model	1T	2T	Other
En	M	Lstm	0.46	0.19	0.32
	NM	Lstm	0.24	0.15	0.61
	M	BART	0.67	0.25	0.08
	NM	BART	0.29	0.17	0.54
De	M	Lstm	0.72	0.08	0.20
	NM	Lstm	0.32	0.27	0.41
It	M	Lstm	0.72	0.05	0.23
	NM	Lstm	0.43	0.18	0.39

Table 3: Statistics of qualitative analysis on prediction errors. Align indicates the type of alignment: M stands for monotonic, NM for non-monotonic. 1T is the proportion of examples requiring a one-token correction without reordering. Similarly, 2T is for two-token corrections without reordering. Other is the proportion of examples requiring more complex corrections of three or more tokens, occasionally with reordering.

Summary

- A significant portion of GEOQUERY examples are monotonically aligned
- Models have a harder time with non-monotonic alignments
- Attention improves performance especially over non-monotonic sequences
- Pre-training helps the model correct hard mistakes

- GEOALIGNED can be used to :
  1. analyze performance of semantic parsers based on alignment complexity
  2. train semantic parsers that model alignments explicitly

Future work

- More alignment classifications
- Different MR formalisms
- Explicit use of alignment annotations in training

- So many of the alignments are monotonic. Does this mean that the dataset is simple, or that there are simple phenomena in semantic parsing that we should be able to capture?
- Not all MR formalisms are equal. When you design a semantic parsing dataset, should you pick an MR that is easier to align?

Questions?

Thanks!

Measuring Alignment Bias in Neural Seq2Seq Semantic Parsers

Davide Locatelli
Universitat Politècnica de Catalunya
Campus Nord, Barcelona
davide.locatelli@upc.edu

Ariadna Quattoni
Universitat Politècnica de Catalunya
Campus Nord, Barcelona
aquattoni@cs.upc.edu

Data and paper: