Deep learning evaluation using ShapeWorld

Alexander Kuhnle

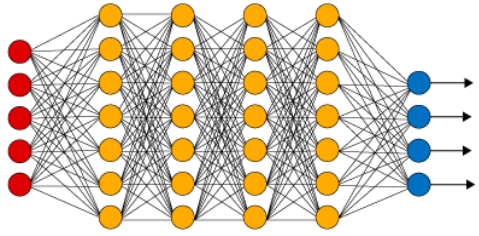
Department of Computer Science and Technology
University of Cambridge

Evaluation methodology

labeled dataset

(>100k data points)

∼90% training split



deep neural network

(recurrent sequence model)

∼10% test split

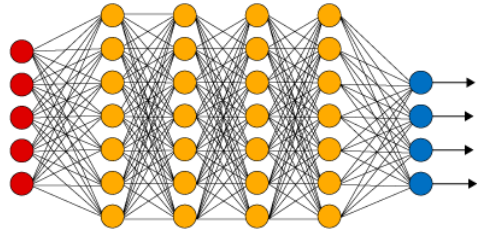
evaluation

Evaluation methodology

data source

(method for obtaining data)

training data



deep neural network

(recurrent sequence model)

feedback & adjustment

series of probing tests

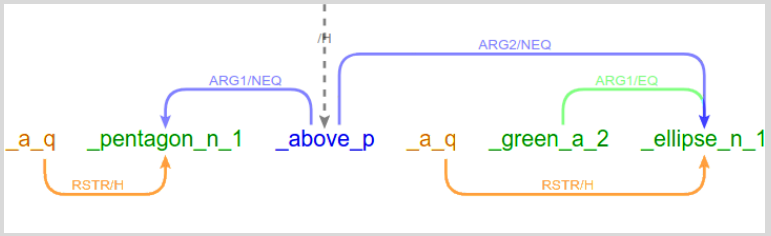
evaluation

ShapeWorld generation framework

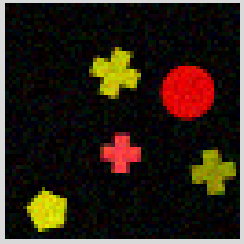
Sampled world model

{ **size**: *64*, **color**: {**name**: *black*, **shade**: *0.0*}, **noise-range**: *0.1*, **entities**:
[{ **shape**: {**name**: *cross*, **extent**: {**x**: *0.10*, **y**: *0.10*}}, **rotation**: *0.06*,
color: {**name**: *yellow*, **shade**: *-0.24*}, **center**: {**x**: *0.47*, **y**: *0.28*} },
{ **shape**: {**name**: *cross*, **extent**: {**x**: *0.08*, **y**: *0.08*}}, **rotation**: *0.76*,
color: {**name**: *red*, **shade**: *0.26*}, **center**: {**x**: *0.49*, **y**: *0.65*} },
{ **shape**: {**name**: *pentagon*, **extent**: {**x**: *0.09*, **y**: *0.08*}}, **rotation**: *0.27*,
color: {**name**: *yellow*, **shade**: *-0.16*}, **center**: {**x**: *0.15*, **y**: *0.91*} },
{ **shape**: {**name**: *circle*, **extent**: {**x**: *0.12*, **y**: *0.12*}}, **rotation**: *0.53*,
color: {**name**: *red*, **shade**: *-0.12*}, **center**: {**x**: *0.80*, **y**: *0.37*} },
{ **shape**: {**name**: *cross*, **extent**: {**x**: *0.09*, **y**: *0.09*}}, **rotation**: *0.73*,
color: {**name**: *yellow*, **shade**: *-0.42*, **center**: {**x**: *0.92*, **y**: *0.73*} }] }

Linguistic representation



Image



Caption

"There is a blue circle."

"Most crosses are yellow."

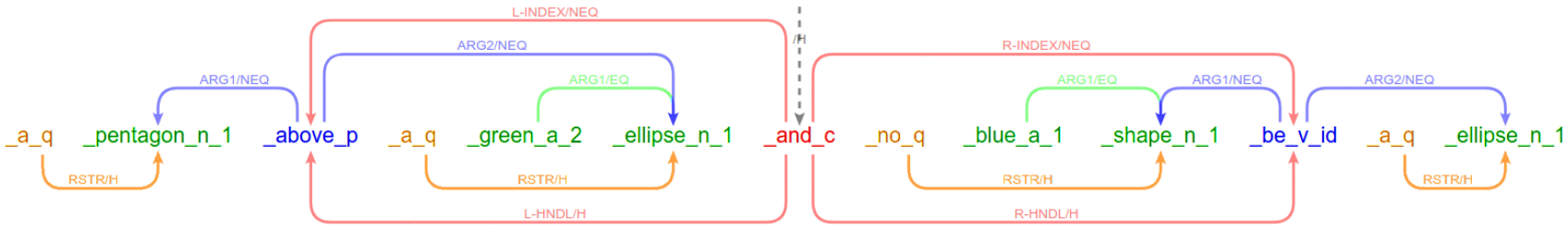
"A pentagon is below a cross."

Agreement?

ShapeWorld: language generation

"A pentagon is above a green ellipse, and no blue shape is an ellipse."

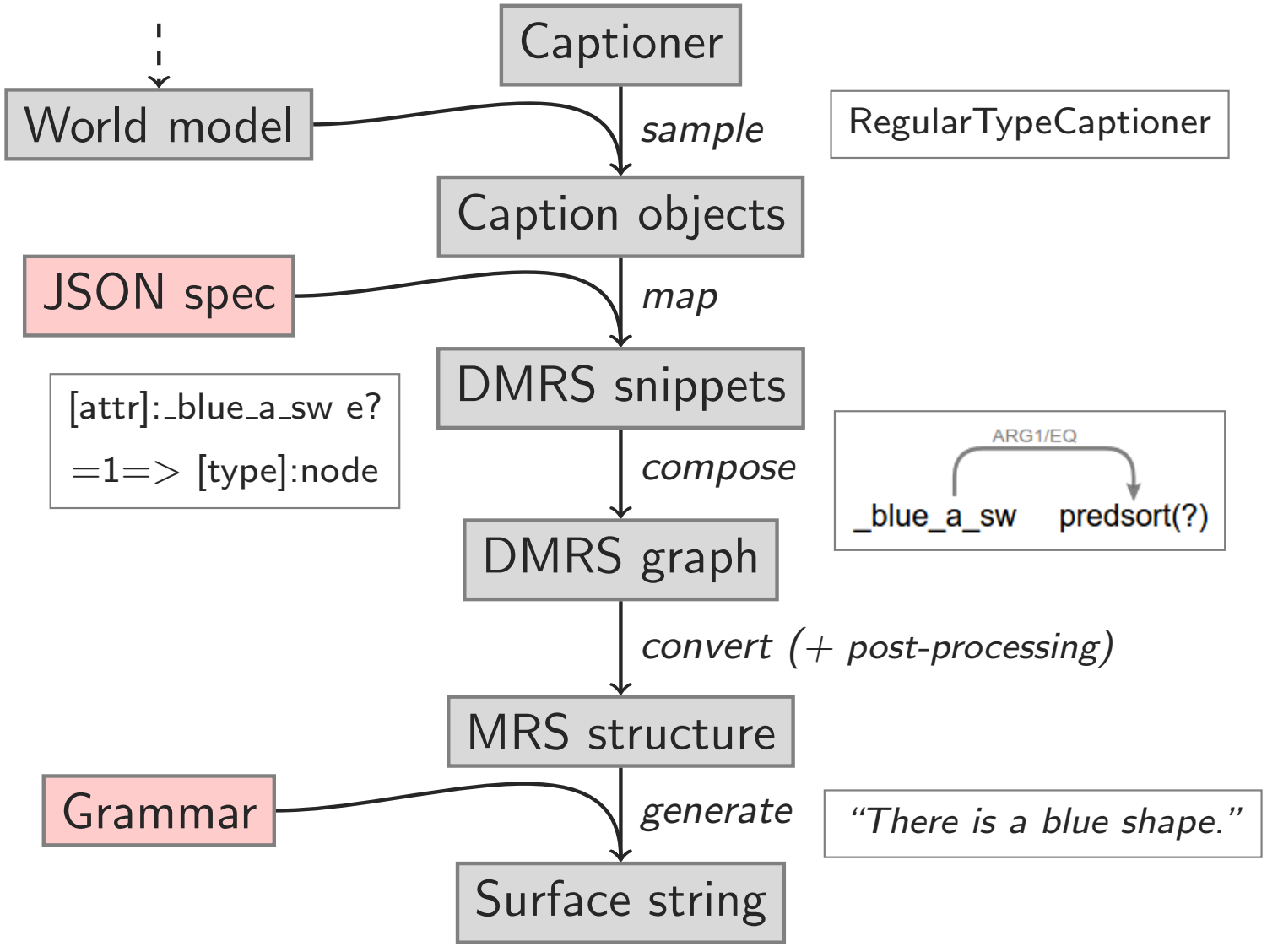
⇑ ERG + ACE realization ⇑



⇑ Internal DMRS mapping ⇑

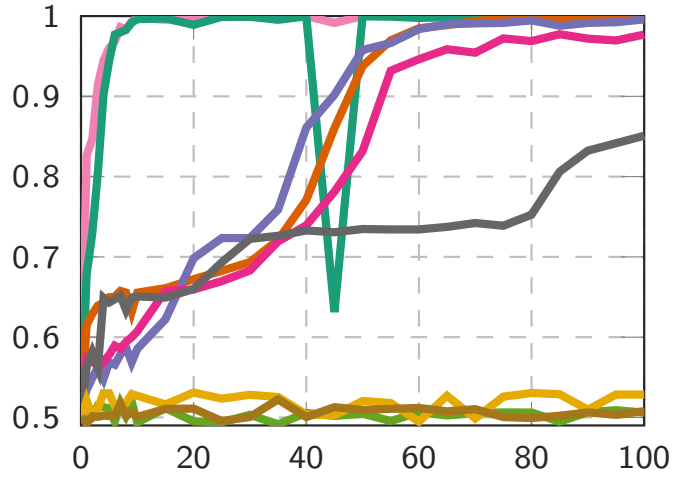
$\exists a$	a.shape=pg	a.y>b.y	$\exists b$	b.color=gr	b.shape=el	$\wedge$	$\neg\exists c$	c.color=bl	true	$c{=}d$	$\exists d$	d.shape=el
$\exists a:$ a.shape=pg		a.y>b.y	$\exists b:$ b.color=gr $\wedge$ b.shape=el			$\wedge$	$\neg\exists c:$ c.color=bl			$c{=}d$	$\exists d:$ d.shape=el	
$\exists a:$ a.shape=pg $\wedge$ [$\exists b:$ b.color=gr $\wedge$ b.shape=el $\wedge$ a.y>b.y]						$\wedge$	$\neg\exists c:$ c.color=bl $\wedge$ [$\exists d:$ d.shape=el $\wedge$ $c{=}d$]					
($\exists a:$ a.shape=pg $\wedge$ [$\exists b:$ b.color=gr $\wedge$ b.shape=el $\wedge$ a.y>b.y]) $\wedge$ ($\neg\exists c:$ c.color=bl $\wedge$ [$\exists d:$ d.shape=el $\wedge$ $c{=}d$])												

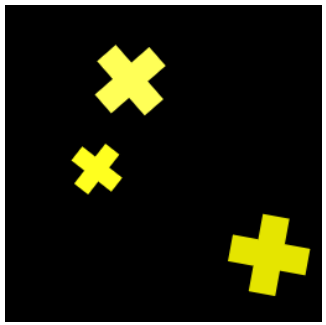
ShapeWorld: language generation



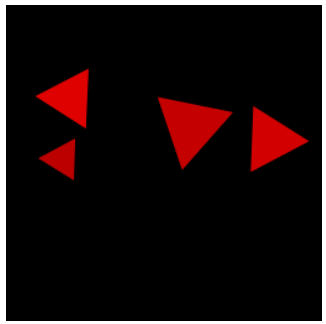
Performance breakdown and generalisation

Dataset	CNN-LSTM		CNN-LSTM-SA		FiLM	
(single-shape)	—		—		100.0	87.2
existential	100.0	81.1	100.0	99.7	100.0	99.9
logical	79.7	62.2	76.5	58.4	99.9	98.9
numbers	75.0	66.4	99.1	98.2	99.6	99.3
quantifiers	72.1	69.1	84.8	80.8	97.7	97.0
(simple-spatial)	81.4	64.8	81.9	57.7	85.1	61.3
relational	—		—		50.6	51.0
implicit-rel	—		—		52.9	53.2
superlatives	—		—		50.8	50.2



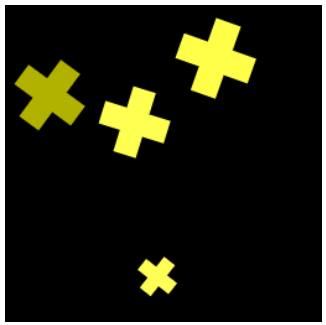


three crosses



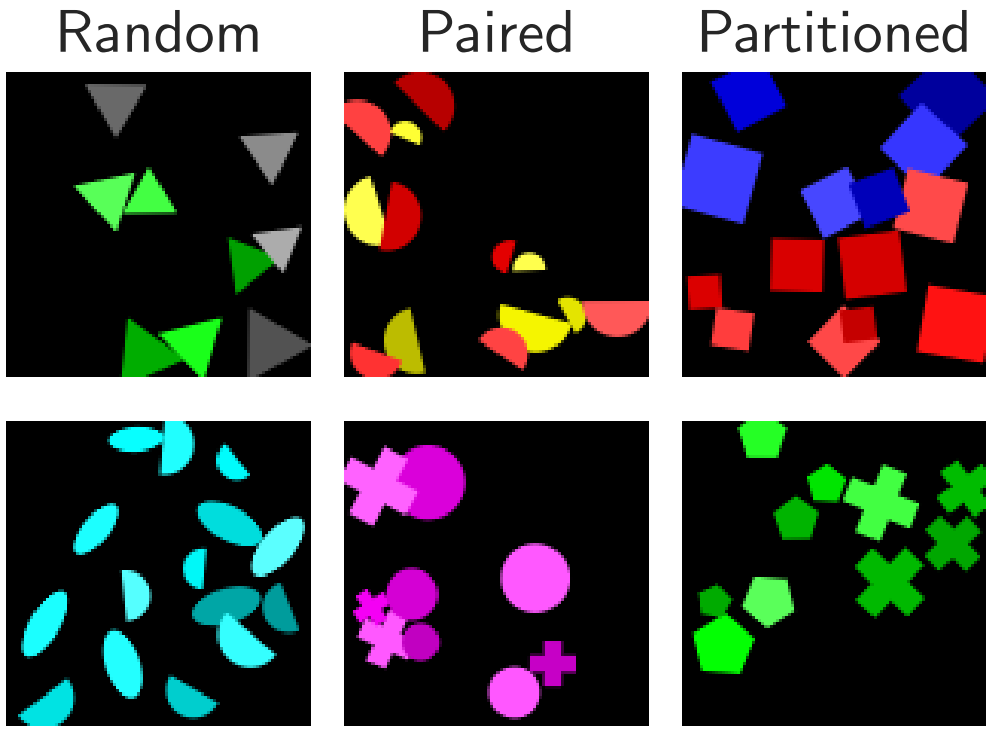
four triangles

$\Longrightarrow$

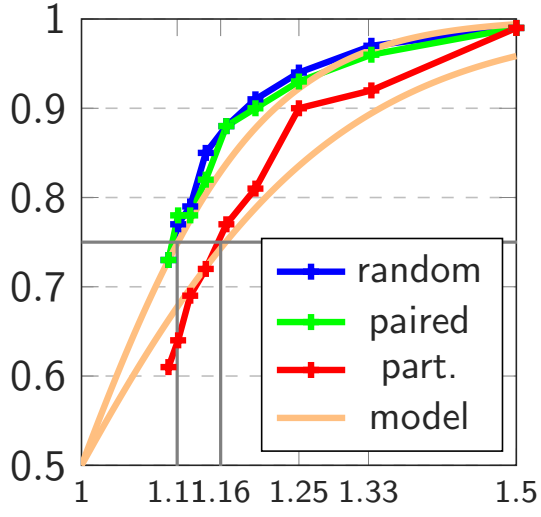


four crosses

Replication of psycholinguistic experiments



+ "More/less than half the shapes are X?"



Intermediate representations and multilingual data

Existential [ObjectType1 Attribute-shape-pentagon]
[Relation-y-rel--1 [ObjectType Attribute-color-green]
Attribute-shape-ellipse]]

"A pentagon is above a green ellipse."

有某一个红色正方形

有一个圆形

有某一个绿色半圆形

有某一个紫色十字形

有某一个红色半圆形

Real-world vs artificial data

real-world data	**vs**	**artificial data**
limited and expensive	$\longleftrightarrow$	unlimited amount
uncontrolled content	$\longleftrightarrow$	configurable content
sparse instance coverage	$\longleftrightarrow$	targeted instance coverage
monolithic benchmark	$\longleftrightarrow$	set of tailored probing tests
test interpolation ability	$\longleftrightarrow$	test extrapolation ability

$\Rightarrow$ **Complementary evaluation paradigms**


Thank you for your attention!
Questions?