UNIVERSITY OF HULL

Using linguistic prior knowledge for deep learning

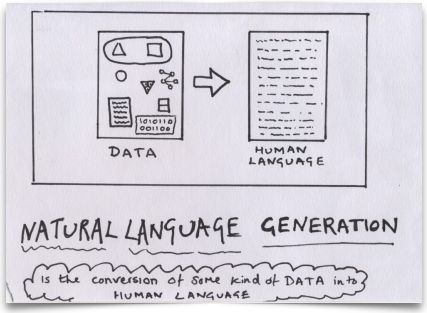
Dr Nina Dethlefs
n.dethlefs@hull.ac.uk

14 Feb 2018

Seminar: Bangor University

UNIVERSITY OF HULL

Natural Language Generation

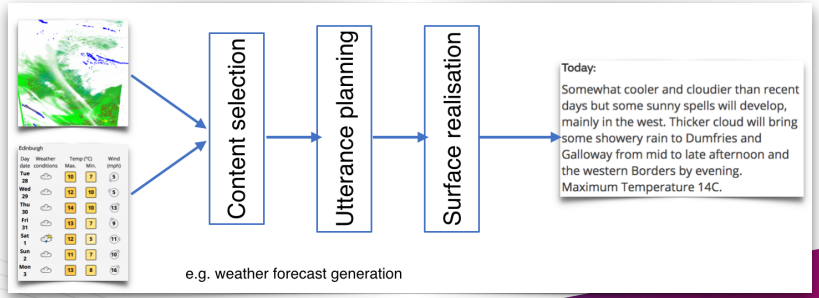


https://www.flickr.com/photos/triptech/7690700352

UNIVERSITY OF HULL

Natural Language Generation (NLG)

Generating natural language from non-linguistic representations.



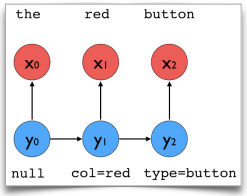
UNIVERSITY OF HULL

Building NLG Systems

3 Approaches.

Mapping rules: human experts help system developers design sophisticated rules to generate text from data in a specific domain.

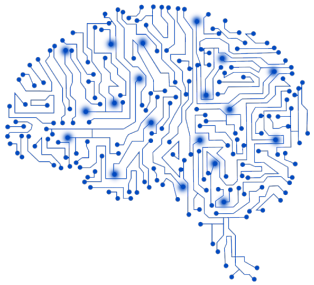
Statistical approaches: learn a mapping from aligned data or (more recently) from unaligned data.



Deep learning: learn a direct mapping from inputs to outputs.

UNIVERSITY OF HULL

Deep Learning



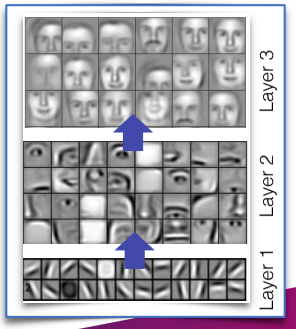
UNIVERSITY OF HULL

Deep Learning - Basic ideas

Models learn increasingly abstract representations of inputs - at different layers of abstraction - and recognise patterns in unstructured data.

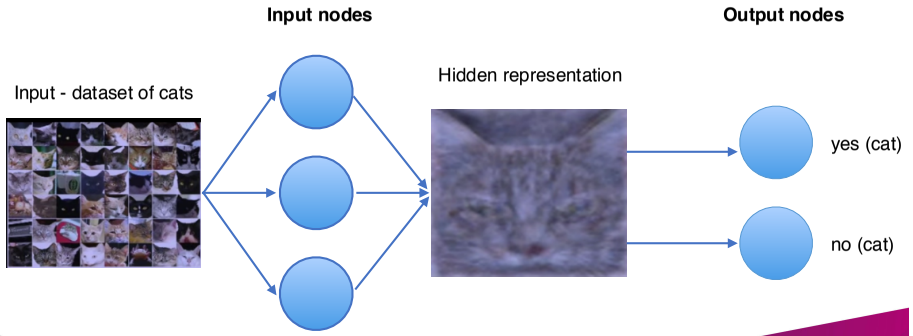
Applications: self-driving cars, natural language processing, image/video recognition, medical imaging, stock market prediction, …

Computationally intensive.



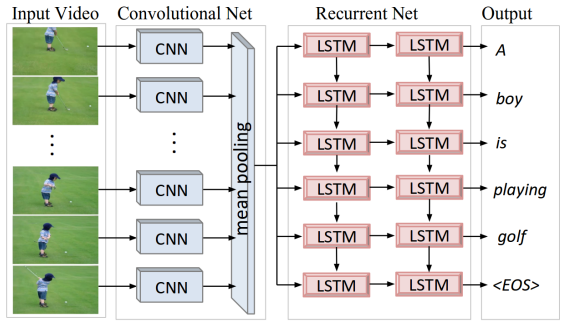
UNIVERSITY OF HULL

Imagine processing - abstract features



UNIVERSITY OF HULL

Video caption generation



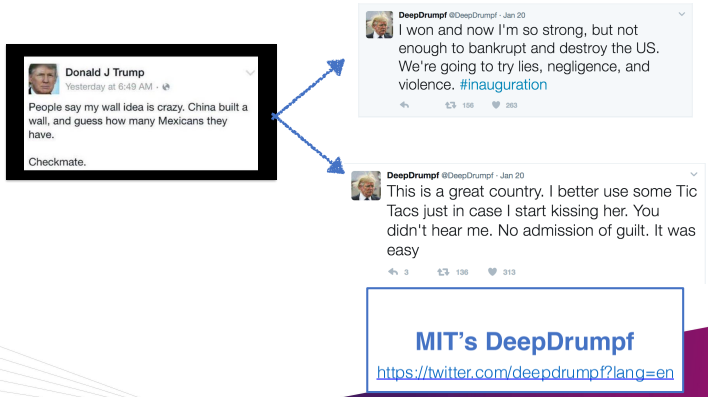
Venugopalan, et al., (2015), "Translating Videos to Natural Language Using Deep Recurrent Neural Networks", CVPR.

UNIVERSITY OF HULL

Sequence generation

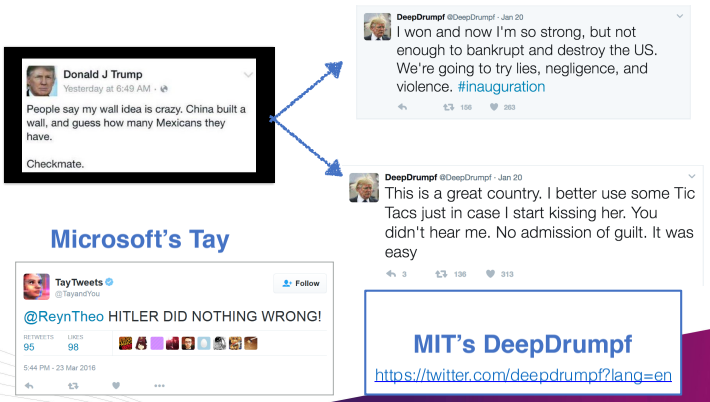
UNIVERSITY OF HULL

Sequence generation



UNIVERSITY OF HULL

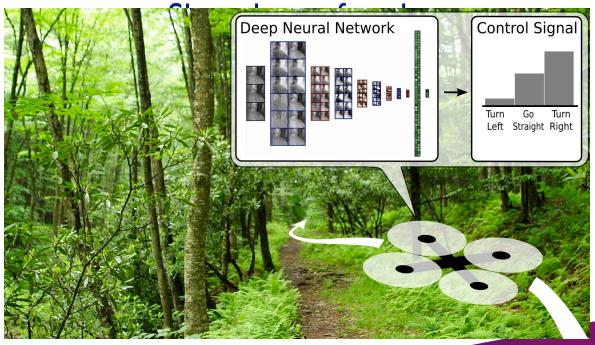
Sequence generation



UNIVERSITY OF HULL

Flying drones

Drones learn to fly unsupervised in new environments.

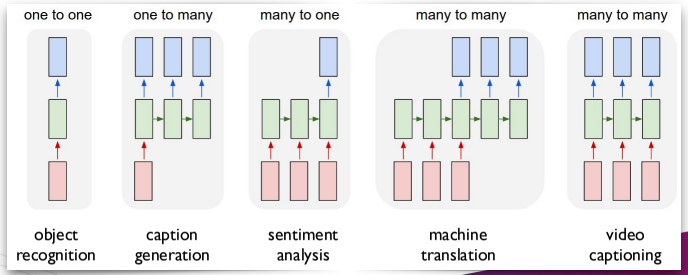


Guisti, et al., (2016), "A Machine Learning Approach to Visual Perception of Forest Trails for Mobile Robots", IEEE RA-L

UNIVERSITY OF HULL

The “best” neural network

Different types of neural nets for different tasks…



UNIVERSITY OF HULL

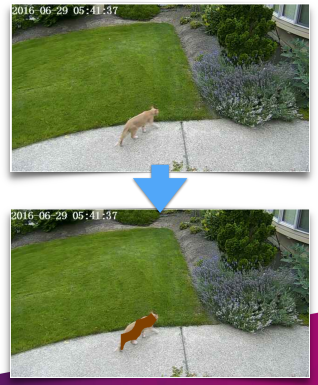
“Folk wisdoms” of deep learning

- To train good models, we need a lot of data
- Data need to be “balanced”
- Very general rules of thumb for parameter tuning
- Generally run on a GPU and / or in a distributed way
- (Don’t expect to understand what the model learnt and why)

UNIVERSITY OF HULL

Deep learning - summary

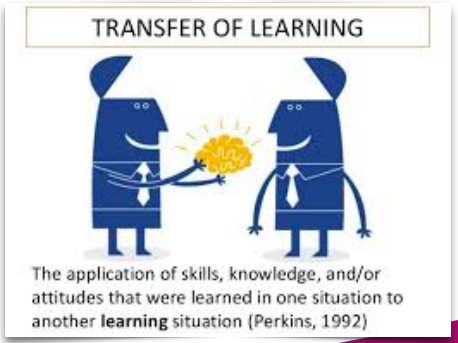
- Deep learning works - given the right setup and conditions - and can be applied to many useful tasks
- Different neural networks are suitable for different tasks, including language tasks
- Lots of parameters and space for improvement…



https://blogs.nvidia.com/blog/2016/07/07/deep-learning-cats-lawn/

UNIVERSITY OF HULL

Domain Transfer for Natural Language Generation



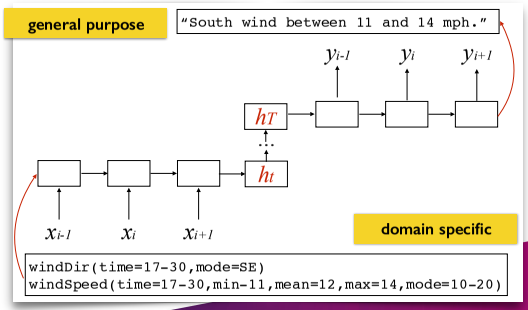
Nina Dethlefs (2017) Domain Transfer for Deep Natural Language Generation from Abstract Meaning Representations. IEEE Computational Intelligence Magazine.

UNIVERSITY OF HULL

Deep learning for NLG

Learn a conditional mapping from inputs to outputs.

The Encoder-Decoder LSTM: encodes and input sequence into an abstract hidden representation and learns to decode it to a (different) output sequence.

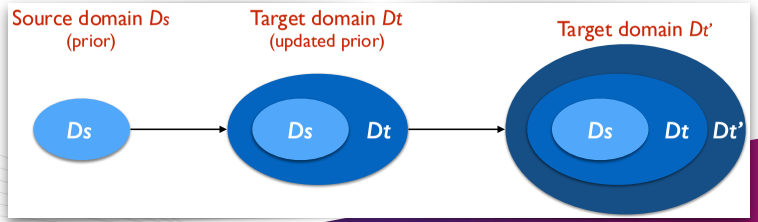


UNIVERSITY OF HULL

Motivation

Develop **NLG systems** from datasets relatively quickly (also) for new / flexible / dynamic domains.

Idea of this work: instead of collecting ever more data to make deep learning work, can we structure datasets differently to make the most of them?



UNIVERSITY OF HULL

Domain Adaptation in NLP

Not a new idea in other areas of NLP: machine translation, NER, capitalisation, shallow parsing.

Approaches often reuse prior knowledge.

Transfer probs from a source to a target domain, and then **adapt** them (Chelba and Acero, 2006; Daume 2007, many more since)

Normally rely on a common feature representation.

Recent work on NLG across domains using synthetic data and shared input representations (Wen et al., 2016).

UNIVERSITY OF HULL

(Part of) The Problem with NLG

1a. Beijing restaurant is serving Chinese food.
1b. Beijing restaurant is a nice place. It serves Chinese food.
1c. Beijing restaurant is a Chinese restaurant.

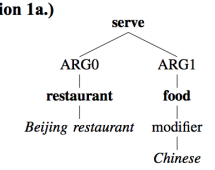
different ways of saying the same thing?

Flat semantic representation (for realizations 1a.-1c.)

```
inform(name='beijing_restaurant', food=chinese)
```

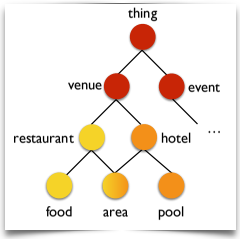
AMR and corresponding tree representation (for realization 1a.)

```
(s / serve
     :arg0 (n / restaurant
              :name 'beijing_restaurant')
     :arg1 (f / food :mod chinese)
)
```



UNIVERSITY OF HULL

(Part of) The Problem with NLG



Underspecified semantic forms may not just pretend that these are the same:

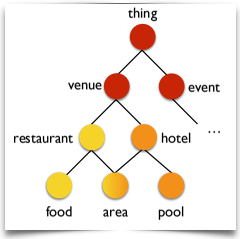
1. *Beijing restaurant serves Chinese food.*
2. *Beijing restaurant is a nice place. It serves Chinese food.*
3. *Beijing restaurant is a Chinese restaurant.*

but also that they're not nothing to with:

4. *The blue cube touches the red sphere.*

UNIVERSITY OF HULL

(Part of) The Problem with NLG



Underspecified semantic forms may not just pretend that these are the same:

1. Beijing restaurant serves Chinese food.
2. Beijing restaurant is a nice place. It serves Chinese food.
3. Beijing restaurant is a Chinese restaurant.

but also that they're not nothing to with:

The blue cube touches the red sphere

UNIVERSITY OF HULL

Learning Constructions

Constructions in cognitive grammar are conventionalised form-meaning pairs at different levels of abstraction: (idioms), partially lexically filled, fully general phrases.

1. Elmo the car gopping. (SOV)
2. Dacking Elmo the car. (VSO)
3. Elmo blicking the car. (SVO)

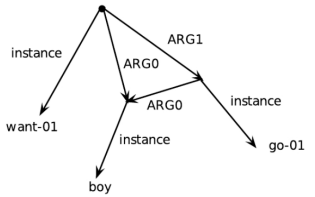
Goldberg, 2005

Can we use constructions in our **x** to **y** mappings?

UNIVERSITY OF HULL

Abstract Meaning Representations

```
(w / want-01
:ARG0 (b / boy)
:ARG1 (g / go-01
      :ARG0 b))
```



Semantic representations as directed graphs — representing a shared set of semantic roles, named entities, modality, negation, questions, quantities, etc.

UNIVERSITY OF HULL

Dealing with words

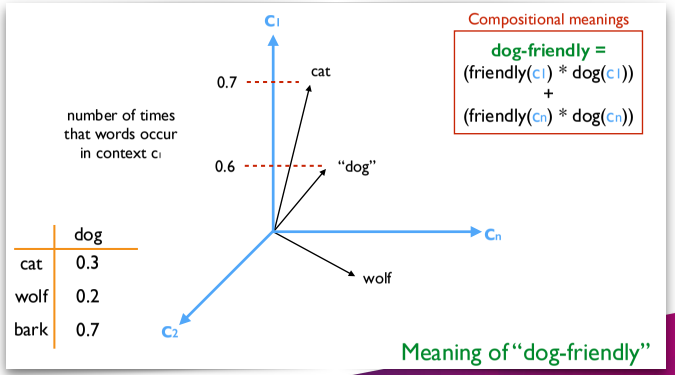
AMRs help us find structural similarities but don't solve the "unseen word problem" — for example, to a neural net, these are complete different:

The Kirin	is	a child-friendly restaurant
The Kirin	is	a dog-friendly restaurant
The Kirin	is	an exquisite restaurant

UNIVERSITY OF HULL

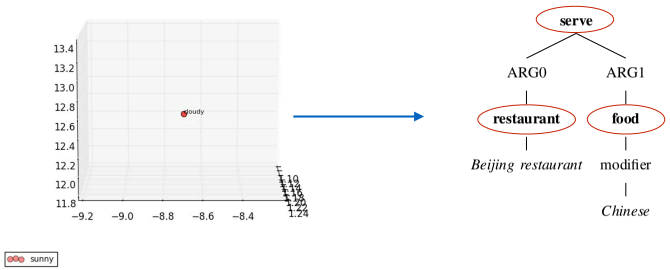
Words - Distributional Semantics

The meaning of a word is defined by "the company it keeps" (aka its co-occurrence statistics).



UNIVERSITY OF HULL

Making AMRs with Word Embeddings



Semantic classes have the same identifier if they occur in the same context.

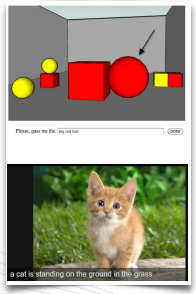
UNIVERSITY OF HULL

Domain Adaptation revisited

Hypothesis: using AMRs + deep learning, we can learn enough lexical-syntactic patterns to generate language in unseen target domains.

Source and target domains

Image recognition
(caption generation)



Navigation
(planning)

Tourist info
(dialogue)

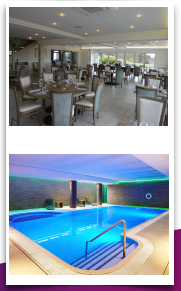
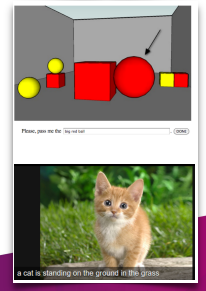


Image recognition

Source domain: GRE3D7 objects in a spatial arrangement. Referring expressions and spatial relations “The blue cube next to the yellow sphere”.

Target domain: REFCOCO image descriptions “a cute little cat in the grass”. Similar but more verbs.

Image recognition
(caption generation)



Navigation (planning)

Source domain: GIVE2 navigation in a spatial virtual environment “go left through the door and press the red button”. Referring expressions, spatial language, imperatives.

Target domain: SAIL navigation in real environments. Similar but more varied.

https://www.youtube.com/watch?v=5-vsSnFX6kI

Navigation
(planning)

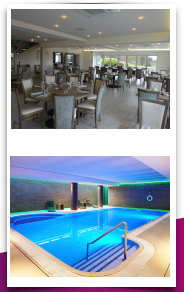
Tourist information (dialogue)

Source domain: SFXRestaurants spoken dialogue system utterances enquiring about user preferences and making recommendations.

Target domain: SFXHotels spoken dialogue system in a similar domain but using different slots and slightly different distribution and sample size.

https://www.youtube.com/watch?v=EZwO8xv2REI&t=23s

Tourist info
(dialogue)



UNIVERSITY OF HULL

Datasets - statistics

Dataset	Examples	Vocab. size	Ave. len.	NPs	SRs	Trans. Cl.	Intr. Cl.	Rel. Cl.	Imperatives	Conj.
GRE	4,480	195	3.50	4,969	1,084	45	0	15	0	3
REFCOCO	142,051	10,046	3.50	302,703	54,484	384	6,634	794	229	2,329
GIVE	1,756	467	3.30	706	1,716	294	17	4	1,043	100
SAIL	816	295	7.95	598	895	218	64	5	656	135
SFXR	6,198	2,058	12.91	6,094	956	2,597	872	598	3	905
SFXH	6,384	1,061	12.13	6,403	1,192	2,247	938	761	1	850

Similar linguistic patterns in datasets but with different distributions.

UNIVERSITY OF HULL

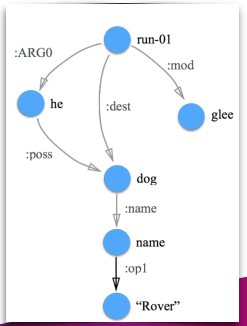
Experimental setup

Three experiments:

In-domain training: train and test in the same domain.

Out-of-domain training: train in one domain, test in another.

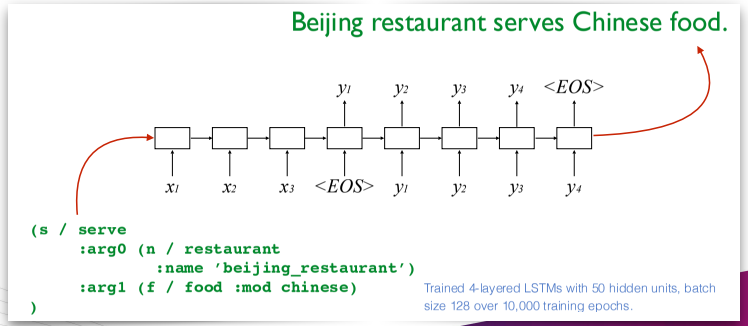
Cross-domain training: train in target domain, but use prior knowledge from a different domain.



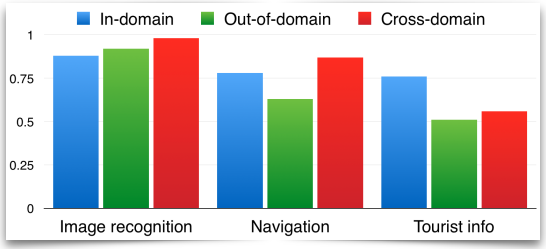
UNIVERSITY OF HULL

(Deep) Learning from AMRs

Improve performance of NLG by leveraging general-purpose linguistic knowledge.



Case study: Results



- Out-of-domain: 75%-100% of original performance
- Cross-domain: up to 110% of original performance.

UNIVERSITY OF HULL

Results: Linguistic Patterns

Complex NP with spatial relation and temporal adverb

```
(e1 / obj :time (n / now)
          :domain (b1 / obj :mod property
        :location (l / obj
          :op1 ( n / on [near, by])))  )
```

"Now, the blue circle on the green sphere."

"Now, the green button by the window."

"Now, an Indian restaurant near Pacific Heights."

UNIVERSITY OF HULL

Results: Linguistic Patterns

Imperative clause construction (with relative clause)

```
(e1 / event :arg0 (y / you)
        :arg1 (b1 / obj :mod property
        :location (w /obj :op1 (o1 / on [in]) ))
 :mode imperative )
```

"Click the red button (that is) on the wall."

"Try Chinese restaurant Kirin in the Pacific Heights area."

UNIVERSITY OF HULL

Results: Linguistic Patterns

Transitive clause construction

```
(e1 / event :arg0 (b1 / obj :mod property)
:arg1 (b2 / obj :mod property))
```

"The yellow sphere (that is) touching the blue box."

"Source restaurant serves Italian food."

UNIVERSITY OF HULL

Results: similarity and error

Objective evaluation using BLEU-3 and -4 score (n-gram similarity with human examples) and semantic error.

No significant differences.

in-domain	GRE-HUMAN	0.90	0.88	0.016
	REFCOCO	0.88	0.82	0.02
	GIVE	0.79	0.78	0.09
	SAIL	0.82	0.77	0.04
	SFXR	0.81	0.76	0.10
	SFXH	0.75	0.69	0.08
out-of-domain	GRE trained from GIVE	0.94	0.92	0.04
	REFCOCO trained from GRE	0.91	0.84	0.02
	GIVE trained from GRE	0.23	0.16	0.29
	SAIL trained from GIVE	0.68	0.63	0.10
	SFXH trained from SFXR	0.61	0.51	0.12
prior	GRE with GIVE prior	0.99	0.98	0.0
	REFCOCO with GRE prior	0.96	0.77	0.01
	GIVE with GRE prior	0.89	0.87	0.06
	SAIL with GIVE prior	0.80	0.72	0.05
	SFXH with SFXR prior	0.74	0.56	0.08

UNIVERSITY OF HULL

Results: naturalness

Subjective evaluation with 204 human judges from AMT, rating 3425 utterances.

"The utterance is natural i.e. could have been produced by a human"

Prior knowledge is better than pure domain transfer. No other significant differences.

	SYSTEM	NATURALNESS
Human	GRE-HUMAN	2.97 (4)
	REFCOCO	3.43 (4)
	GIVE	3.77 (4)
	SAIL	4.36 (5)
	SFXR	4.17 (4)
	SFXH	3.93 (4)
in-domain	GRE-HUMAN	3.12 (3)
	REFCOCO	3.45 (4)
	GIVE	2.76 (2)
	SAIL	4.02 (4)
	SFXR	3.67 (4)
	SFXH	3.95 (4)
out-of-domain	GRE trained from GIVE	3.11 (3)
	REFCOCO trained from GRE	3.20 (3)
	GIVE trained from GRE	2.03 (2)
	SAIL trained from GIVE	2.95 (3)
	SFXH trained from SFXR	3.45 (3)
prior	GRE with GIVE prior	3.29 (3)
	REFCOCO with GRE prior	3.41 (4)
	GIVE with GRE prior	3.09 (3) *
	SAIL with GIVE prior	3.44 (4) *
	SFXH with SFXR prior	3.58 (4)

UNIVERSITY OF HULL

Did it work?

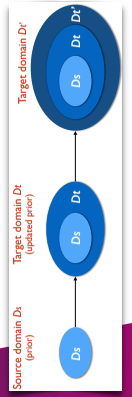
Yes:

Model learns lexical-syntactic patterns from AMRs that are not domain-specific

Out-of-domain results are acceptable in some domains, similar to in-domain training in others

Learning from source domain priors gives us up to 10% extra performance

Limitations:
AMR annotation is an overhead.



UNIVERSITY OF HULL

Conclusions

Deep learning works well, but only with enough data of the right kind.

More "intelligence" needed in algorithms — transfer of prior knowledge across tasks.

More control needed to constrain decisions — linguistic knowledge is one way to achieve this for NLP tasks.

(Synthetic data seems a bad idea)

NVIDIA

diffbot