Imitation learning, zero-shot learning and automated fact checking

—

Andreas Vlachos
http://andreasvlachos.github.io/

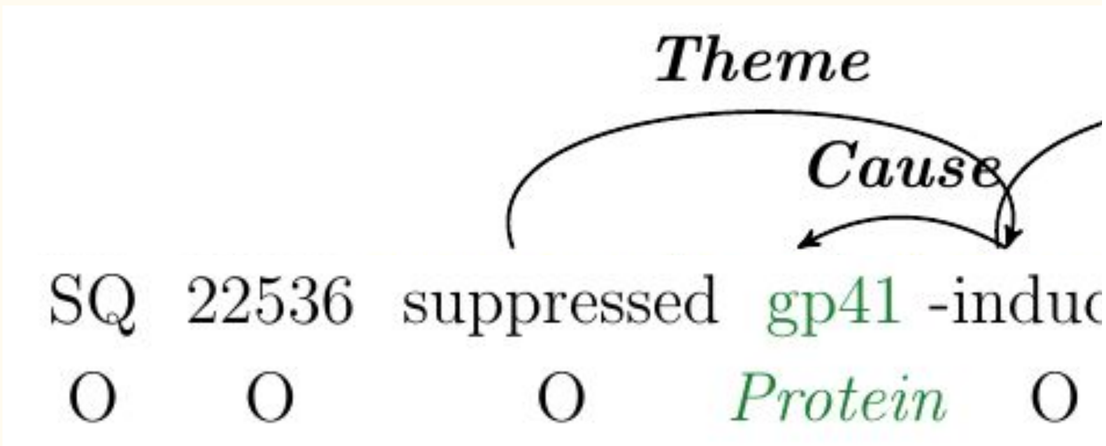
Introduction

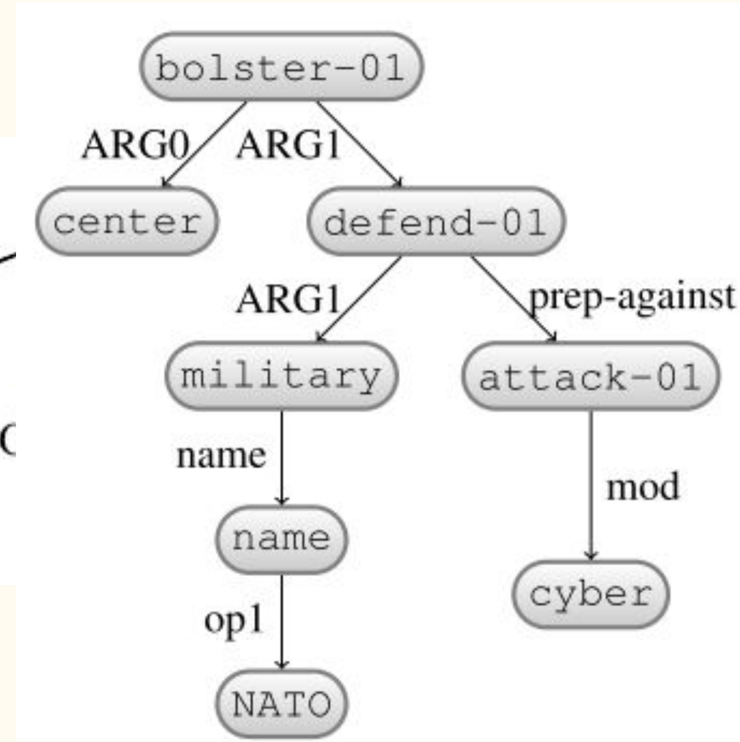
- Lecturer at the Department of Computer Science at the University of Sheffield
  - Member of the natural language processing (NLP) and machine learning (ML) research groups

- Research in developing ML methods for:
  - natural language **understanding**: convert text into (machine-readable) meaning representations
  - natural language **generation**: convert meaning representations into (human-readable) text
  - applications encompassing both directions

Research Context

Natural Language Understanding (NLU)





- Named entity recognition (Vlachos et al., PSB 2006)
- Relation extraction (Vlachos and Craven, CoNLL 2011)
- Semantic parsing (Goodman et al., ACL 2016)

Natural Language Generation (NLG)

INPUT:

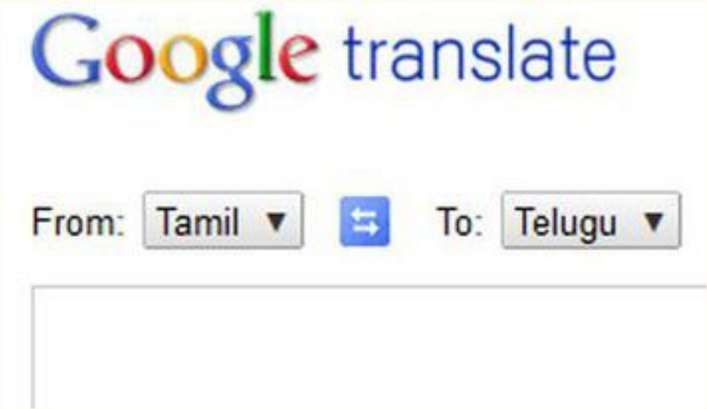
```
predicate= INFORM
name = "The Saffron Brasserie"
type = placetoeat
eattype = restaurant
area = riverside, "addenbrookes"
near = "The Cambridge Squash", "The Mill"
```

OUTPUT:

The Saffron Brasserie is a restaurant at the side of the river near the Cambridge Squash and the Mill in the area of Addenbrookes

- SOTA on 3 datasets (Lampouras and Vlachos, Coling 2016)
- NN-based system most fluent among 20 systems in End2End NLG (Chen et al., 2018)

Applications encompassing both directions



- Translation Quality Estimation (Beck et al., WMT 2016)
- Digital Personal Assistants (Vlachos and Clark, TACL2014)
- Automated Fact Checking (Vlachos and Riedel, Computational Social Science and NLP 2014)

Machine Learning for Natural Language

Learning from data allows us to adapt rapidly to:

- language evolution
- different applications

Compared to rule-based approaches:

- wider coverage
- weighted feature combinations
- feature learning with neural networks/deep learning
  - reuse models across tasks (trade-off between feature engineering vs architecture engineering)
  - facilitate focus on novel tasks

This talk

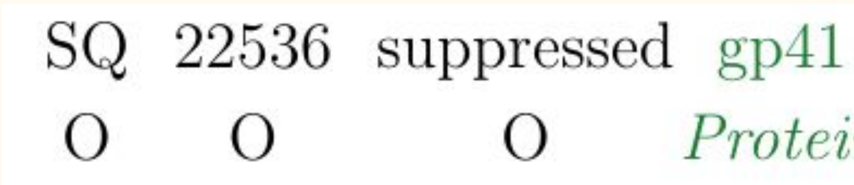
- Improved structure prediction with **imitation learning**

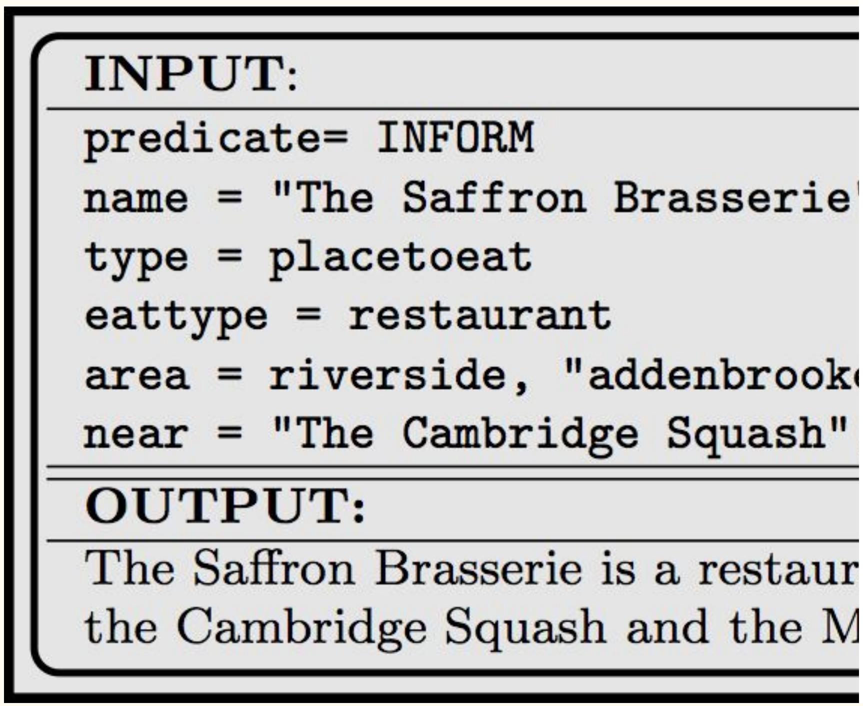
- Ability to predict labels unseen during training using **zero-shot learning with neural networks**

- A challenge to advance ML, NLP and artificial intelligence: **automated fact checking**

Imitation learning for structured prediction

Structured prediction in NLP is everywhere

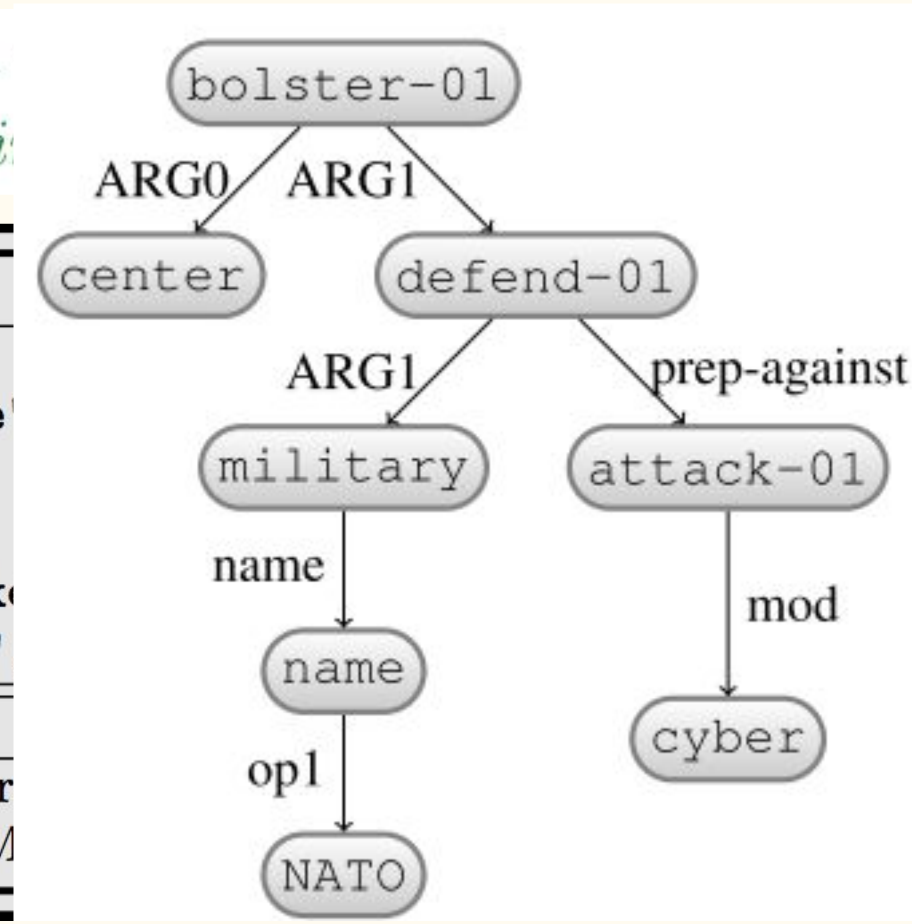



INPUT:

```
predicate= INFORM
name = "The Saffron Brasserie
type = placetoeat
eattype = restaurant
area = riverside, "addenbrook
near = "The Cambridge Squash"
```

OUTPUT:

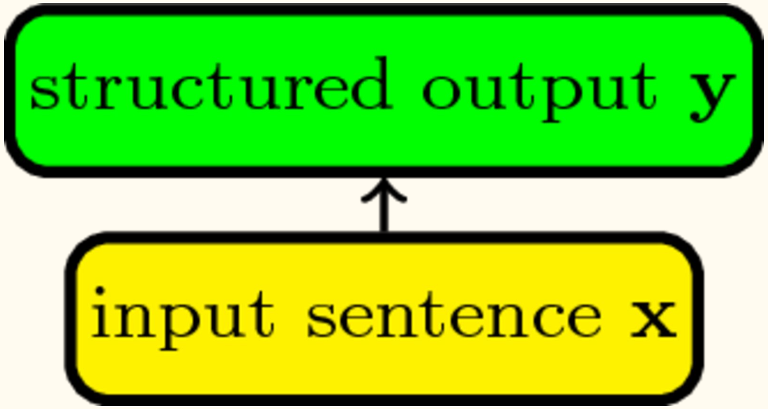
The Saffron Brasserie is a restaur
the Cambridge Squash and the M



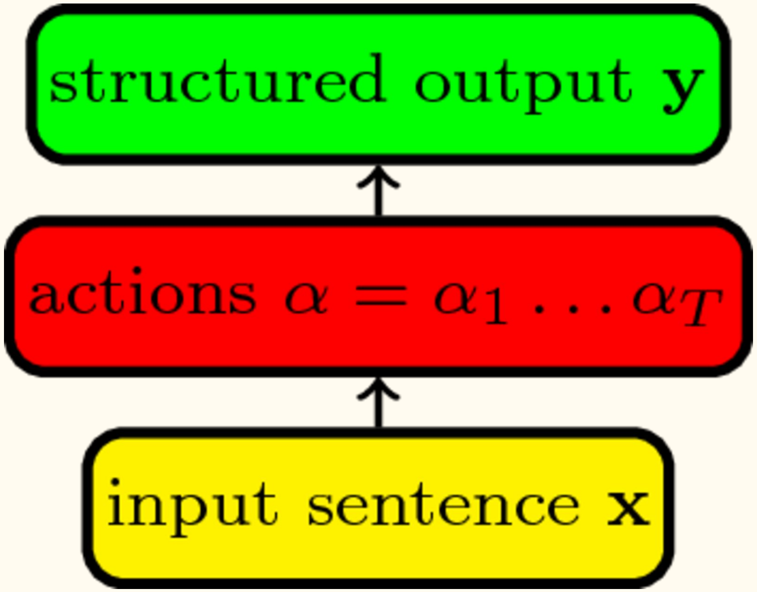
Sequences of labels, words and graphs combining them

Imitation learning for structured prediction

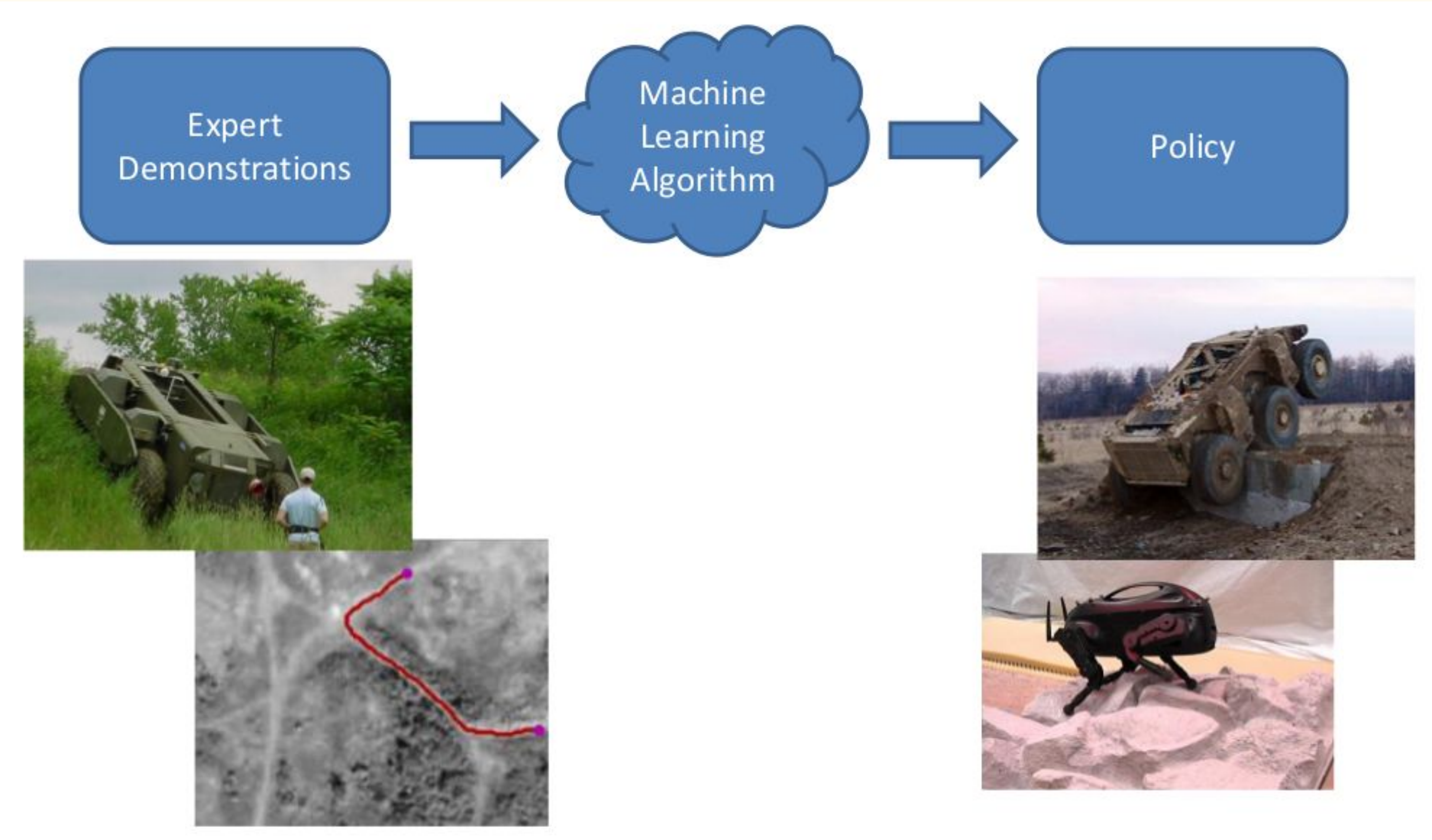
- Assume human-annotated input-output (**x**,**y**) for **supervised** training



- Train a classifier to predict the **actions** ($\boldsymbol{\alpha}$) constructing the output **y**
- Actions not annotated; imitation learning is **semi-supervised**



Imitation learning in robotics

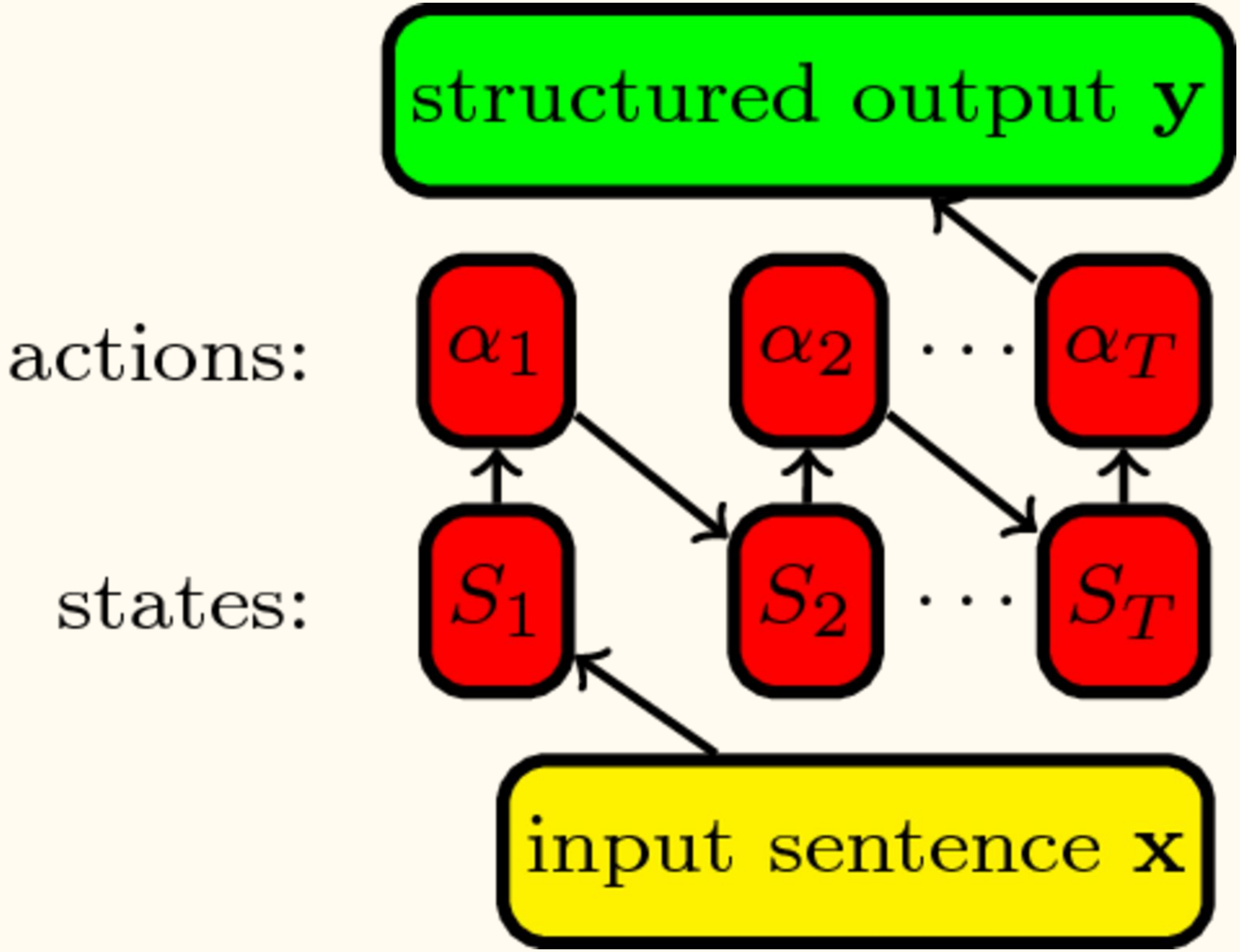


Meta-learning: better model ($\approx$policy) by generating better training data from expert demonstrations

Relation to reinforcement learning

- Both reinforcement and imitation learning learn a classifier/policy to maximize reward
- Learning in imitation learning is facilitated by an **expert**

Breaking output into actions constructing it



Incremental structured prediction

A classifier **f** predicting actions to construct the output:

$$\hat{\mathbf{y}} = output\left(\begin{aligned} \hat{\alpha}_1 &= \arg\max_{\alpha \in \mathcal{A}} f(\alpha, \mathbf{x}), \\ \hat{\alpha}_2 &= \arg\max_{\alpha \in \mathcal{A}} f(\alpha, \mathbf{x}, \hat{\alpha}_1), \cdots \\ \hat{\alpha}_N &= \arg\max_{\alpha \in \mathcal{A}} f(\alpha, \mathbf{x}, \hat{\alpha}_1 \ldots \hat{\alpha}_{N-1}) \end{aligned} \right)$$

✓ Use our favourite classifier
✓ No need to enumerate all possible outputs
✓ No modelling restrictions on features

x Prone to error propagation
x Classifier not trained w.r.t. task-level loss

Imitation learning

Improve incremental structured prediction by:

- addressing error-propagation
- training wrt the task-level loss function

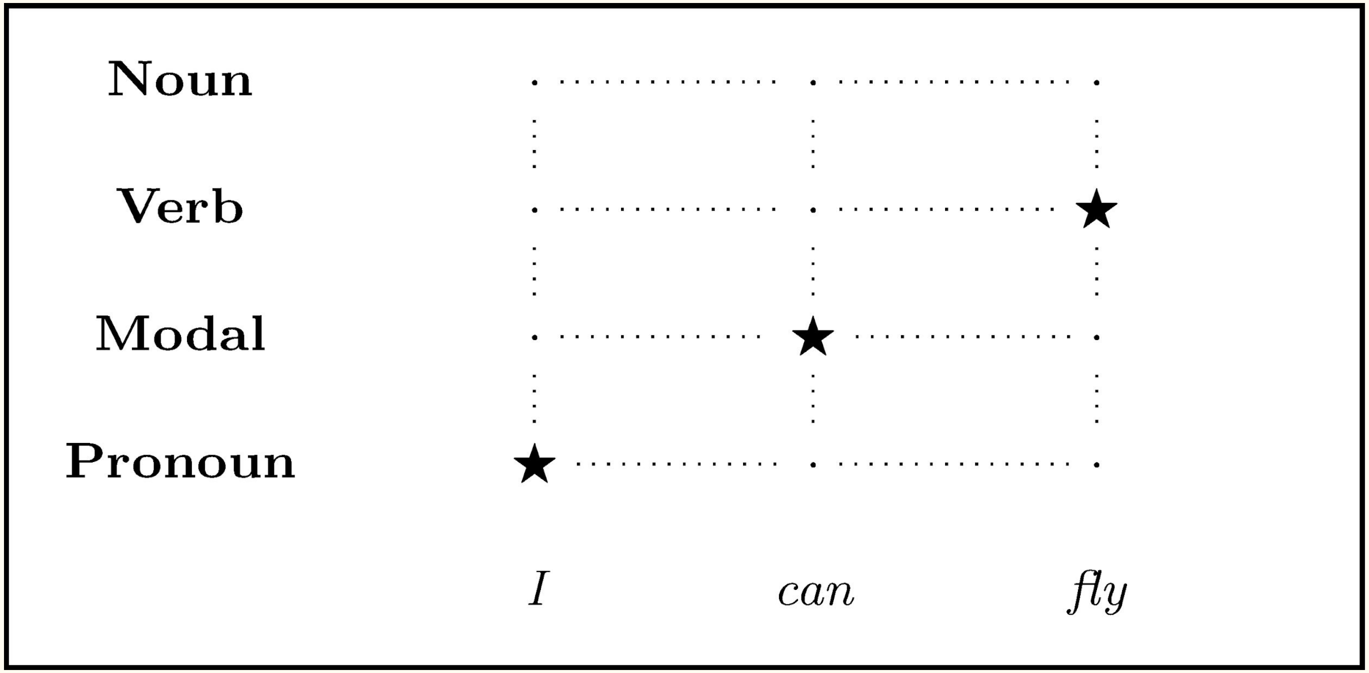
Meta-learning: use our favourite classifier and features, but generate better training data

Can handle more complex problems than joint inference approaches:

- no output enumeration ⇛ no need for dynamic programming
- no dynamic programming ⇛ no modelling restrictions such as Markov assumptions used in conditional random fields, etc.

Imitation learning for part of speech tagging

Human annotated tags:

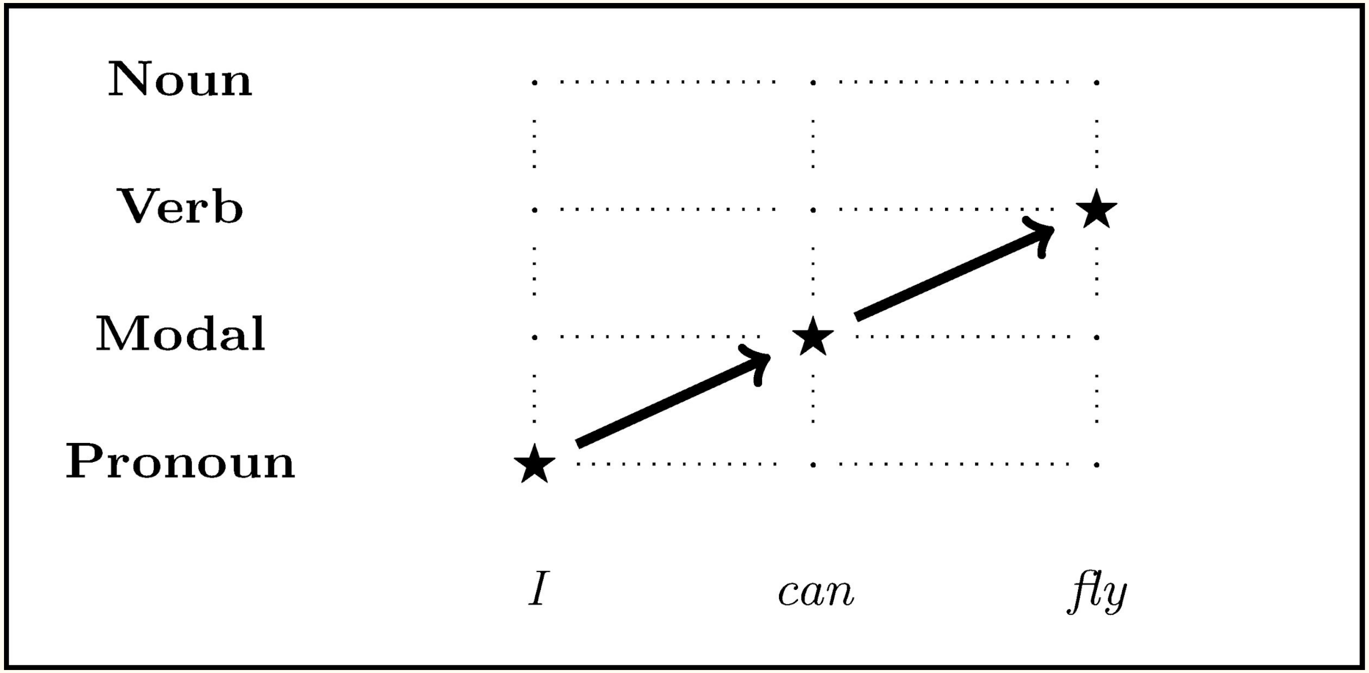


expert policy: at each word return the correct tag

loss: number of incorrect tags

Imitation learning for part of speech tagging

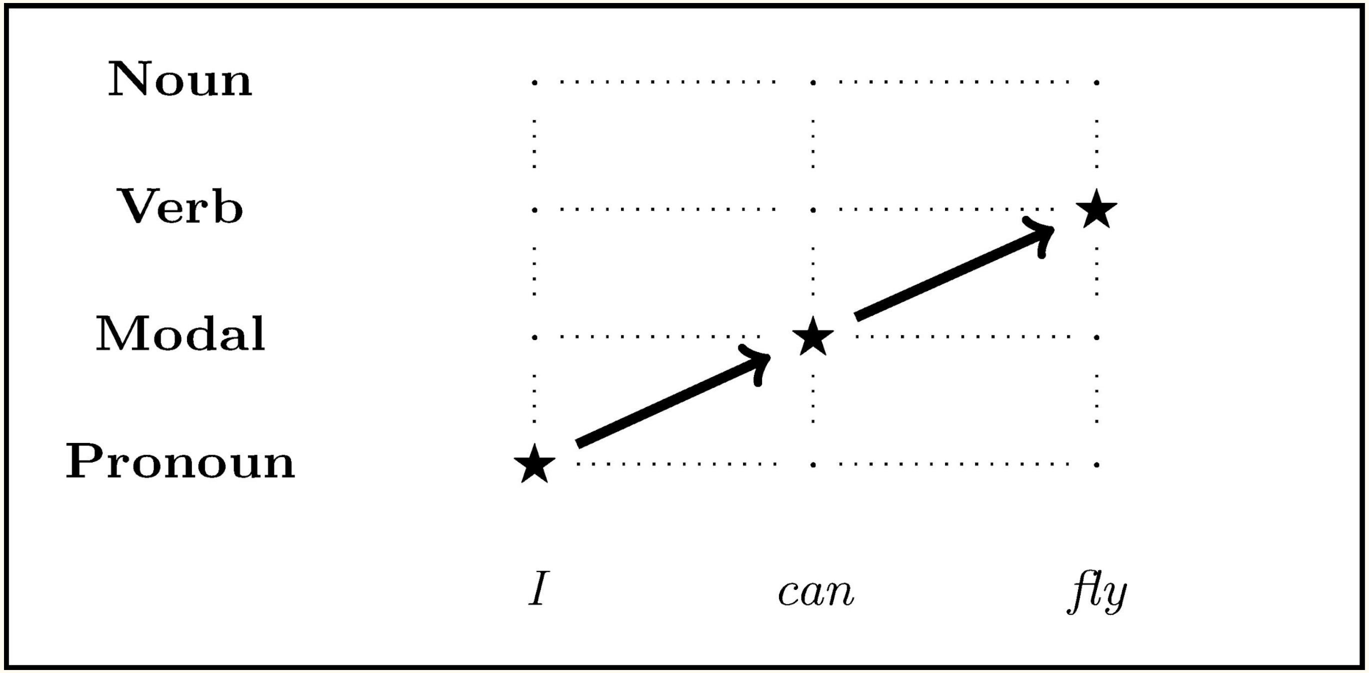
Standard incremental structured prediction:



word	**label**	**features**
I	Pronoun	token=I, prev=NULL ...
can	Modal	token=can, prev=Pronoun ...
fly	Verb	token=fly, prev=Modal ...

Imitation learning for part of speech tagging

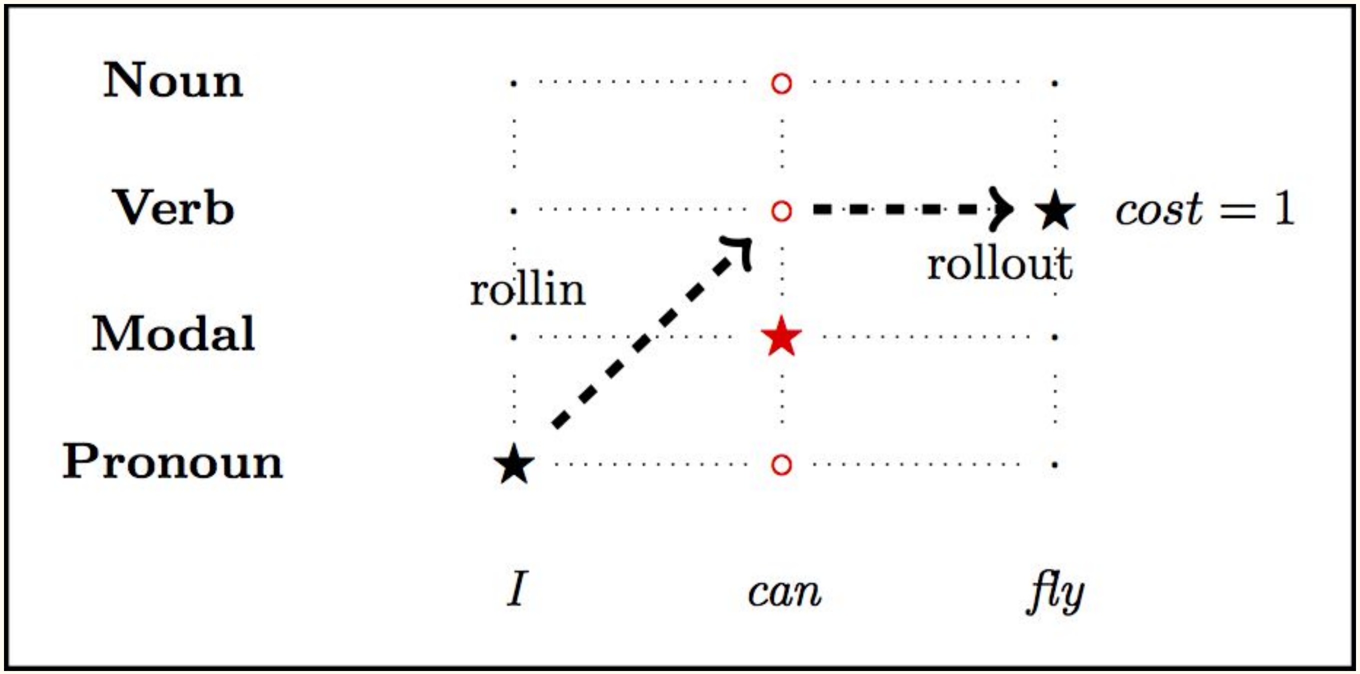
Labels as costs:



word	Pronoun	Modal	Verb	Noun	features
I	0	1	1	1	token=I, prev=NULL...
can	1	0	1	1	token=can, prev=Pronoun...
fly	1	1	0	1	token=fly, prev=Modal...

Imitation learning for part of speech tagging

Breaking down action costing:



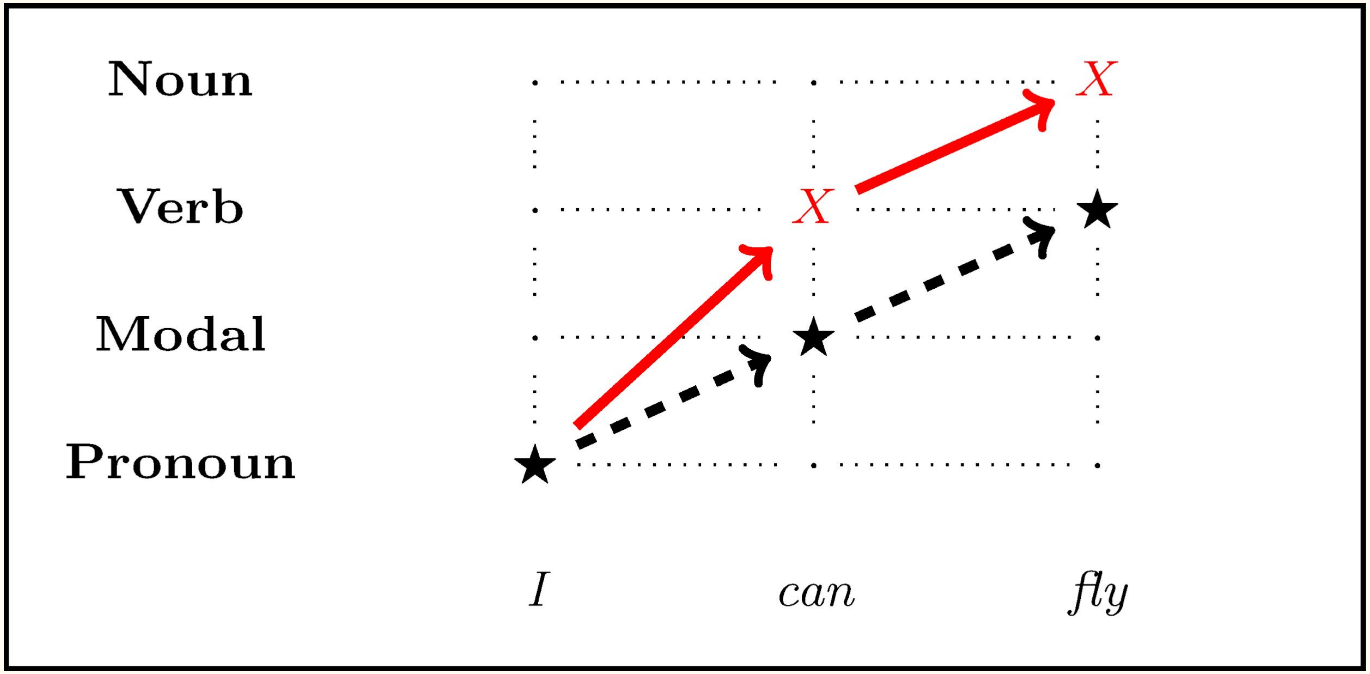
- **rollin** to obtain a trajectory through the sentence
- for each label:
  - **rollout** to complete the output prediction
  - **cost** the complete output with the task loss

If rollin and rollout with the expert policy then correct labels have 0 cost, incorrect have 1

word	Pronoun	Modal	Verb	Noun	features
can	1	0	1	1	token=can, prev=Pronoun...

Imitation learning for part of speech tagging

Mixed rollins/rollouts with the expert policy and the classifier



word	Pronoun	Modal	Verb	Noun	features
I	0	1	1	1	token=I, prev=NULL...
can	1	0	**2**	1	token=can, prev=Pronoun...
fly	1	1	0	1	token=fly, **prev=Verb**...

Back to learning how to drive

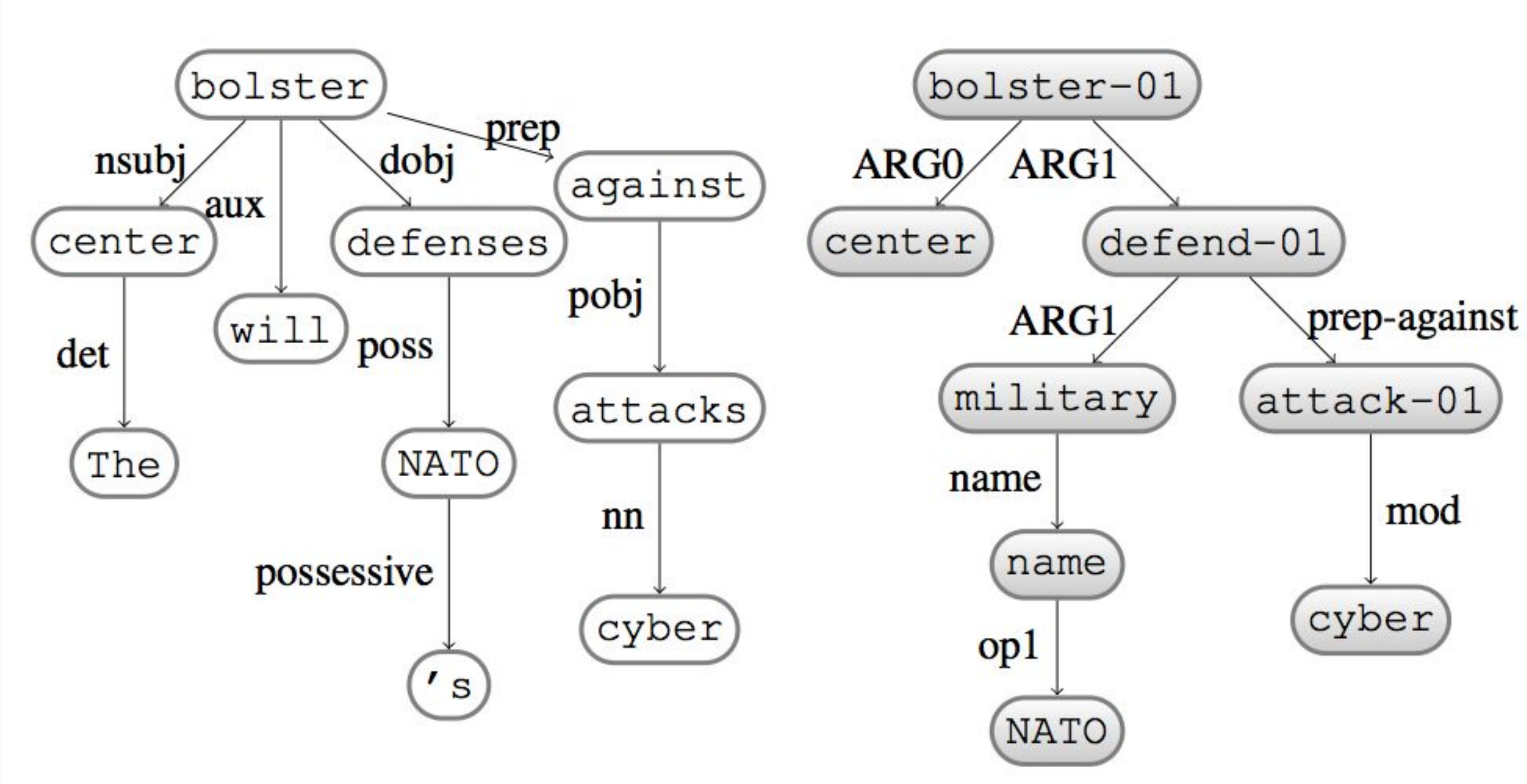
- Instead of observing the expert drive, let the classifier drive
- The expert gives the correct actions given the classifier's ones
- The classifier is allowed to explore the effect of its own actions

Imitation learning for NLP

- Explores only the parts of the search space likely to be encountered ⇛ **applicable to complex outputs**
- Training data generation mixing expert and classifier ⇛ **addresses error propagation**
- Task loss only used on complete outputs ⇛ can train against **non-decomposable loss functions** such as BLEU, ROUGE, etc.
- Addresses a fundamental limitation of incremental predictors, including **recurrent neural networks**

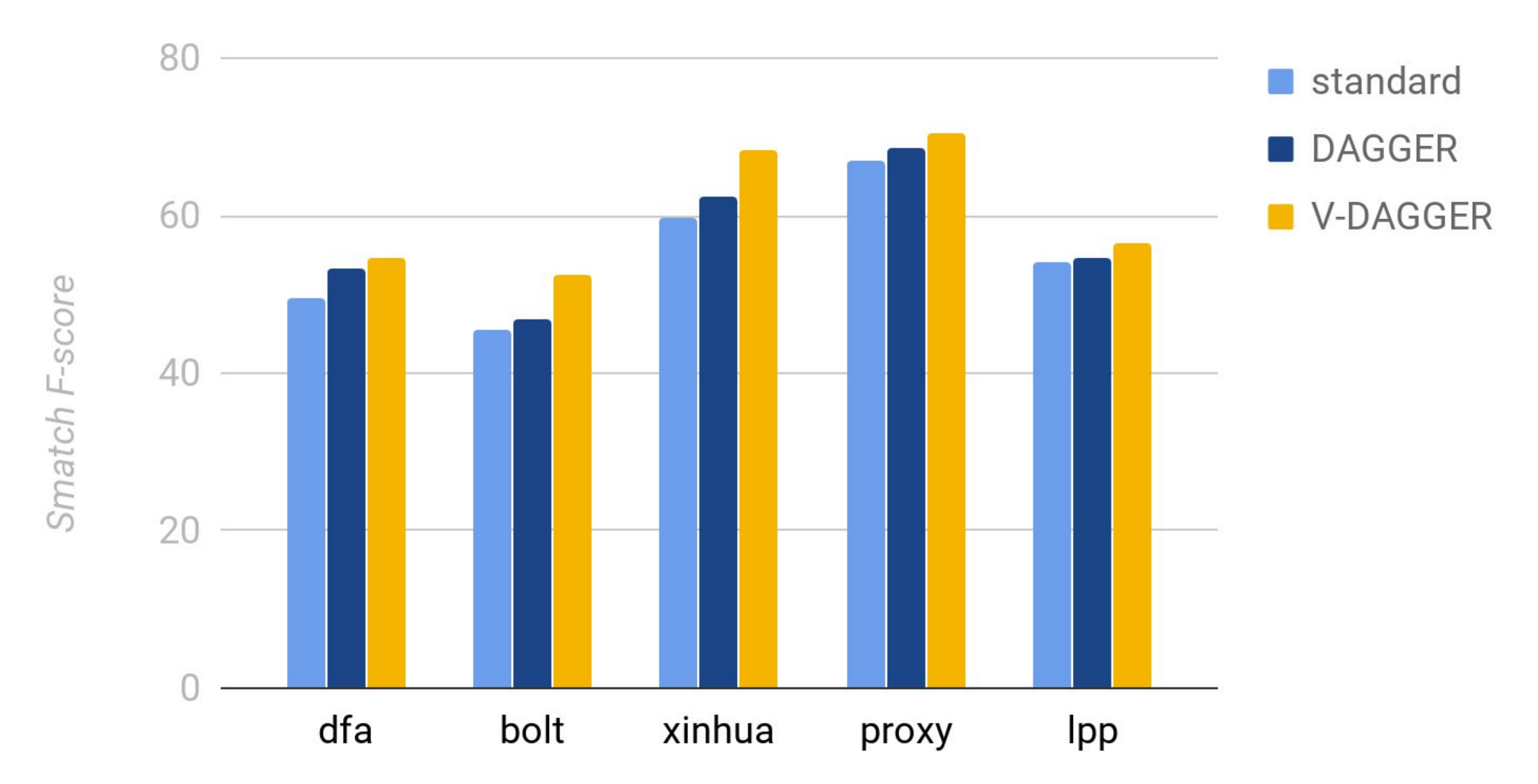
More in our EACL 2017 tutorial, but now some real applications

Imitation learning for semantic parsing



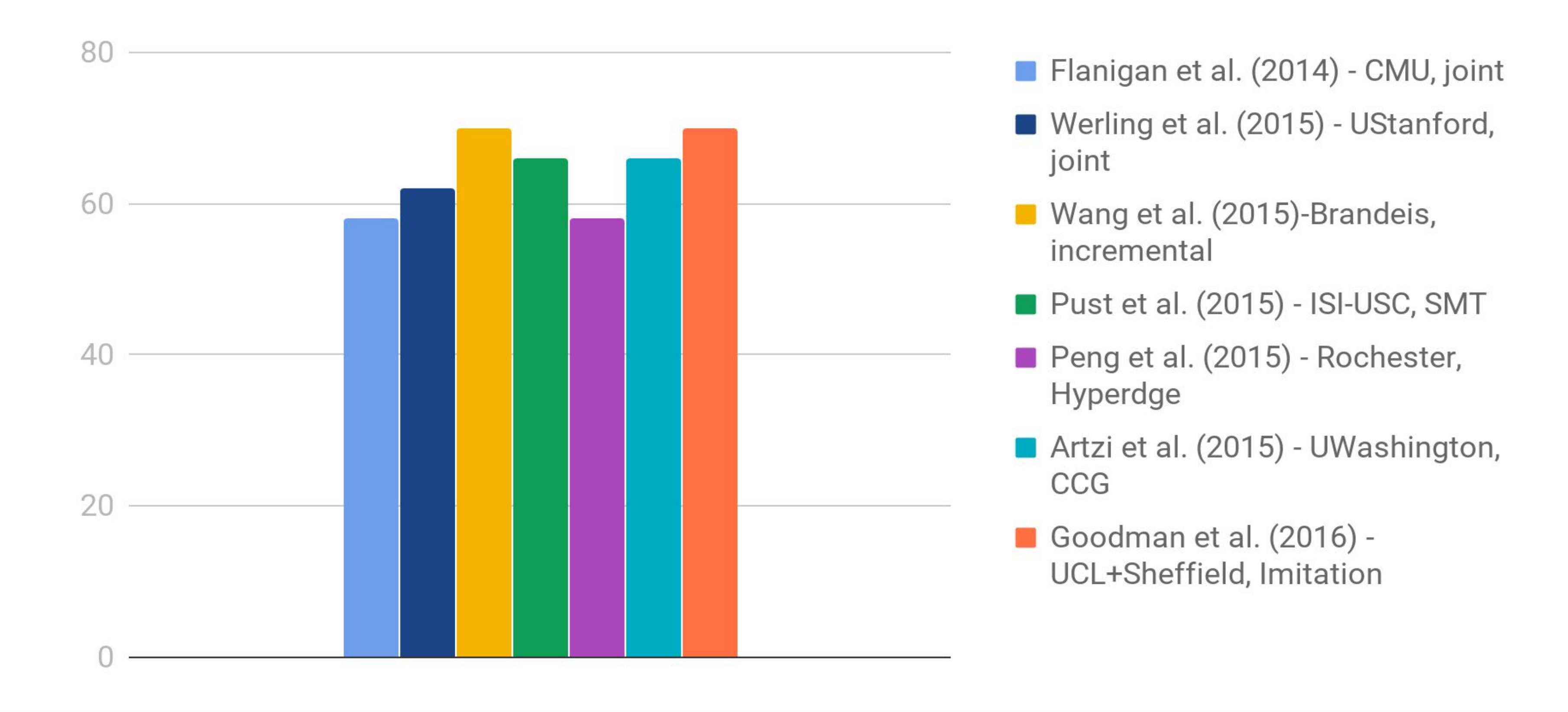
- Convert a syntax tree to a meaning graph
- Long complex action sequences (>100 actions, 10K labels)
- Used in many applications: summarization, generation, etc.

Imitation learning benefits



- DAGGER uses rollins (Ross et al., AISTATS 2011)
- V-DAGGER uses roll-in/-outs (Vlachos and Clark, TACL 2014)

Semantic Parsing Evaluation



- Best reported results (Goodman et al., ACL 2016)
- No external resources used, just the training data
- Docker image of parser downloaded >100 times

Imitation learning for Language Generation

INPUT:

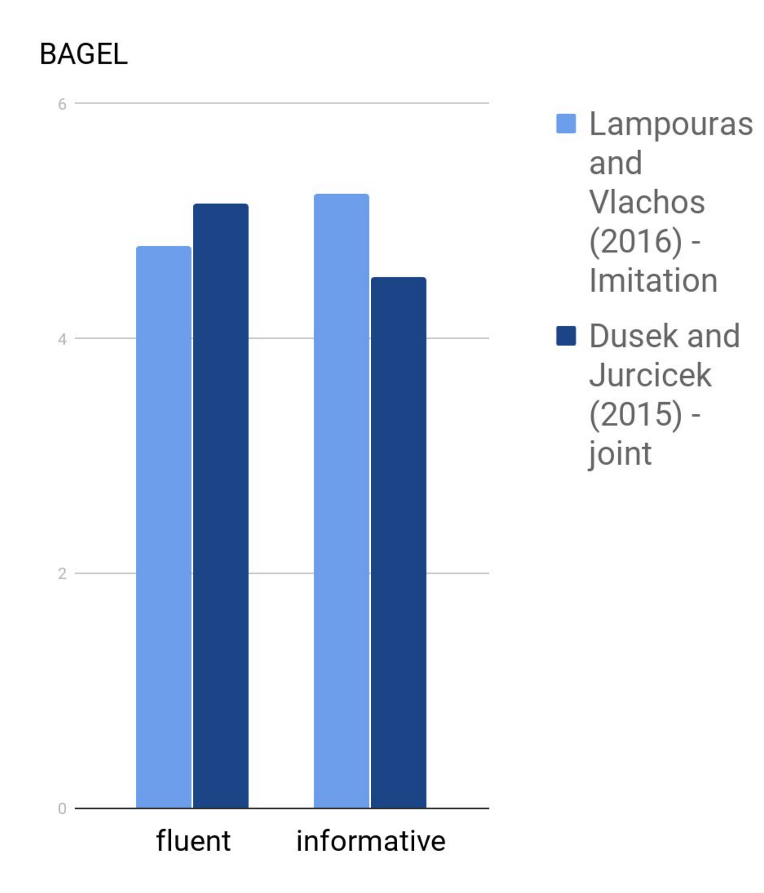
```
predicate= INFORM
name = "The Saffron Brasserie"
type = placetoeat
eattype = restaurant
area = riverside, "addenbrookes"
near = "The Cambridge Squash", "The Mill"
```

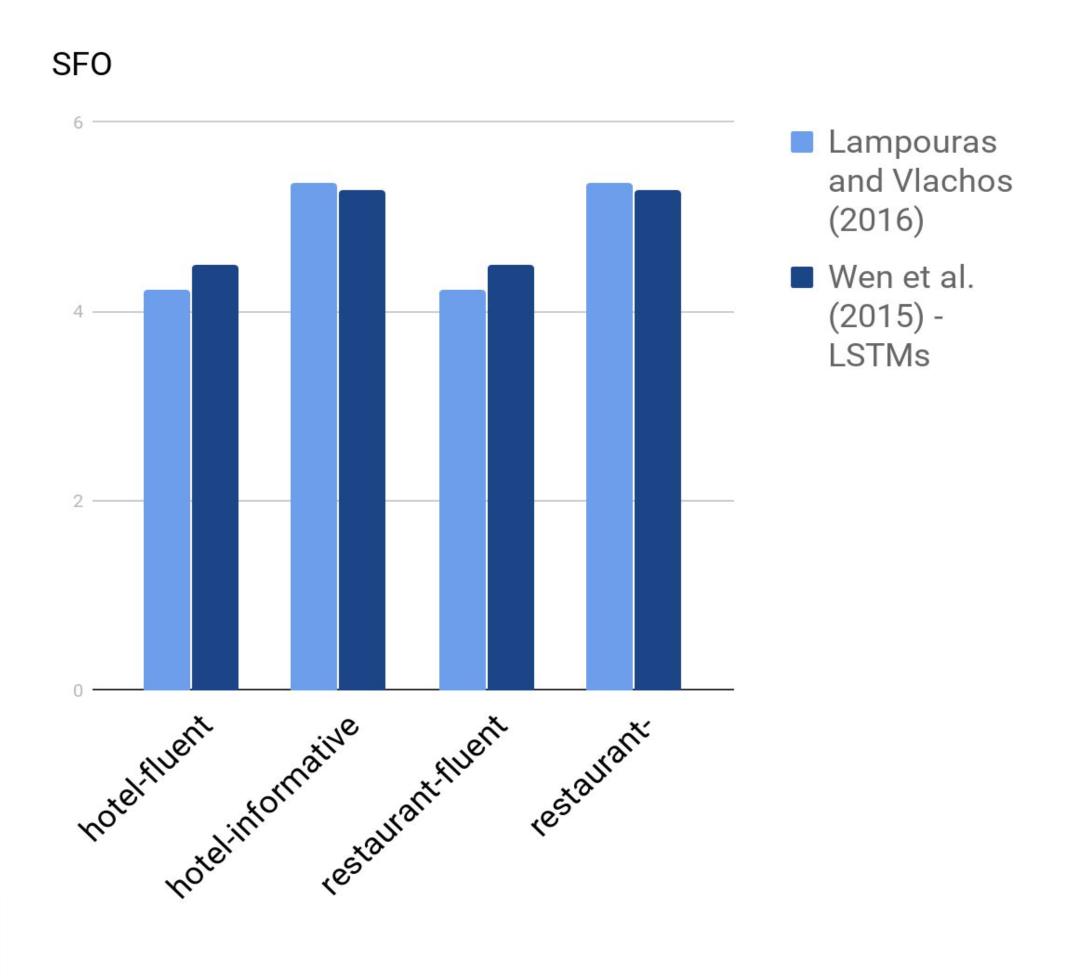
OUTPUT:

The Saffron Brasserie is a restaurant at the side of the river near the Cambridge Squash and the Mill in the area of Addenbrookes

- Reversed semantic parsing, similar to machine translation (MT)
- Unlike MT, labeled data is rather limited

Language Generation - Human Evaluation





- SOTA on three datasets (Lampouras and Vlachos, 2016)
- No rules, re-ranking or templates, just two classifiers

More imitation learning applications

Own work:

- Biomedical Event Extraction (Vlachos and Craven, CoNLL2011)
- Language Understanding for Digital Personal Assistants (Vlachos and Clark, TACL 2014)
- Knowledge Base Population (Augenstein et al., EMNLP 2015)
- Machine Translation Quality Estimation (Beck et al., WMT 2016)

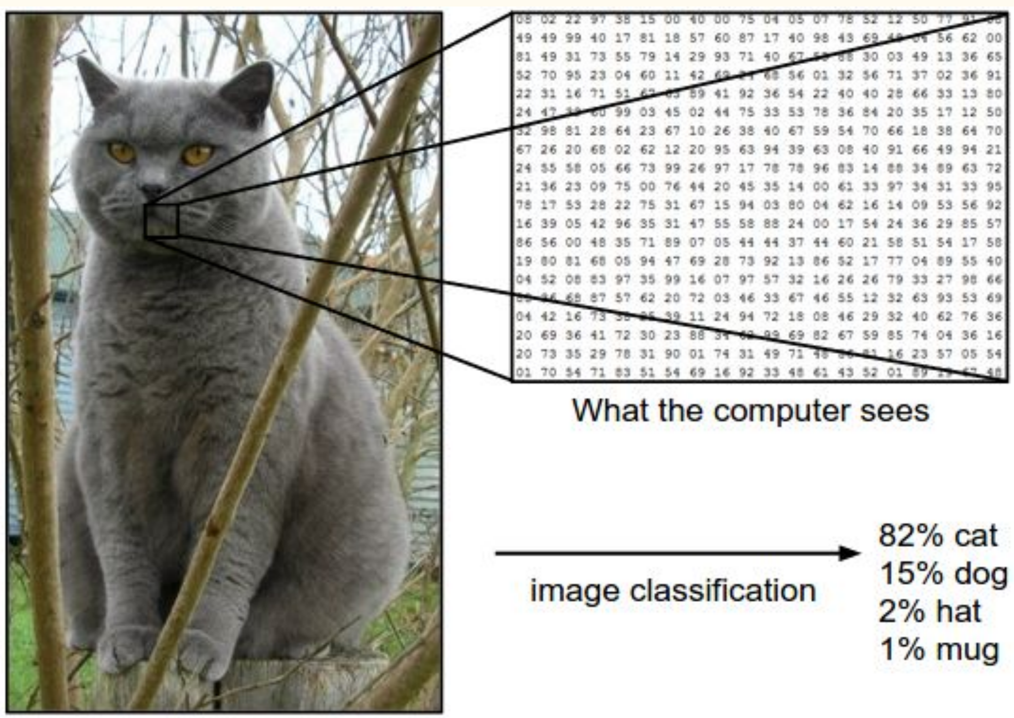
Others:

- Syntactic dependency parsing
  - Dynamic oracles (Goldberg and Nivre, Coling 2012)
  - LSTM-based (Ballesteros et al., EMNLP 2016)
  - Popular spacy.io NLP toolkit
- Coreference resolution (Clark and Manning, ACL 2015)

Zero-shot learning with neural networks

Zero-shot learning

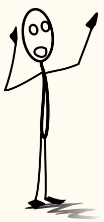
ML models typically can predict only labels they saw in the training data, e.g. a model trained on cats and dogs can’t recognize birds



Zero shot learning explores how to predict labels unseen in training

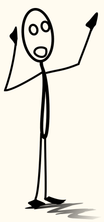
Stance classification

Given a target concept, e.g. **abortion** or **Hillary Clinton**, decide whether a text is **positive/negative/neutral** towards the target:



No more Hillary!

Can we learn a model for targets unseen in training?



No more Hillary!

Zero-Shot Stance Classification

Standard supervised learning:

$$\hat{y} = \underset{y \in \mathcal{Y}}{\arg\max}(\mathbf{w}_t^y \cdot \phi(x))$$

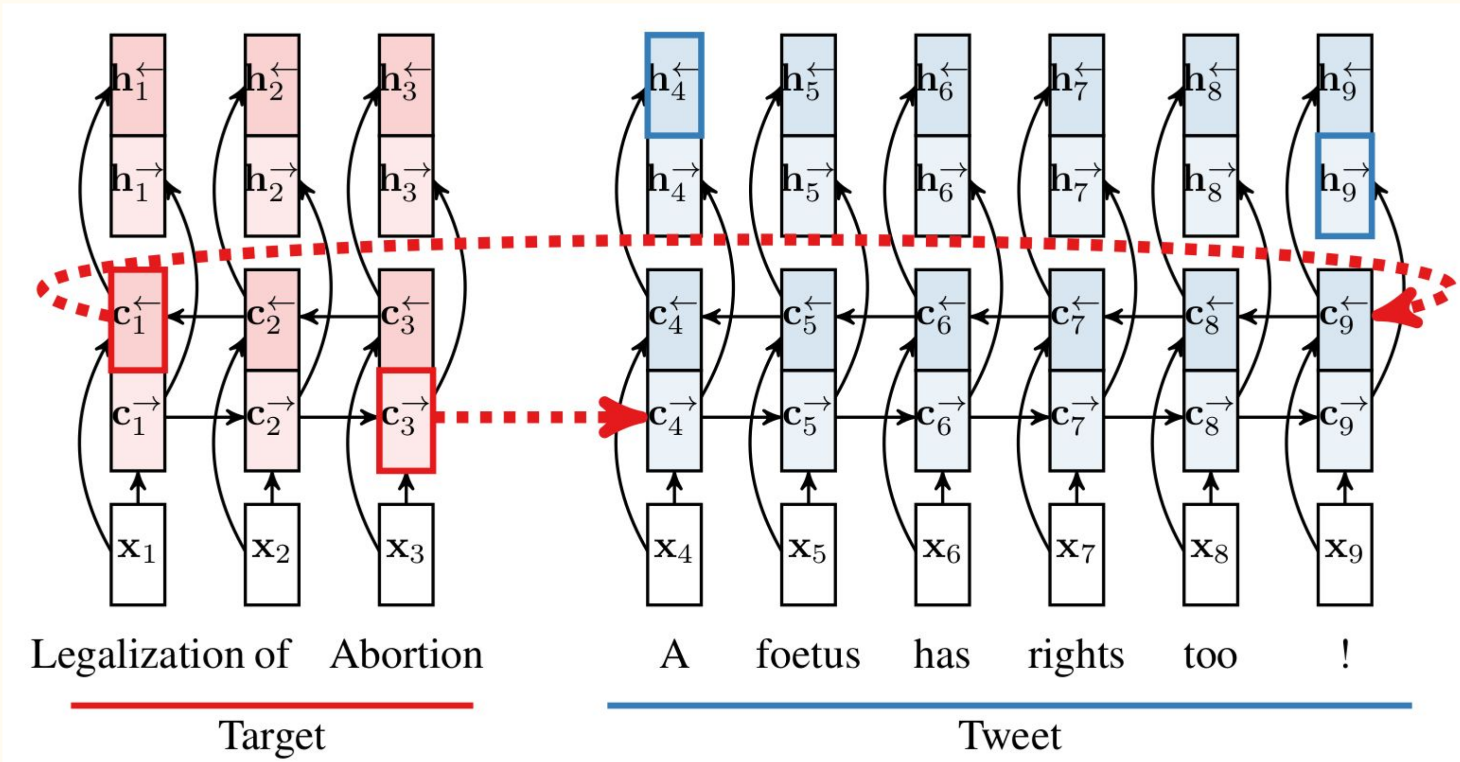
- learn weights $\mathbf{w}$ for each label $\boldsymbol{y}$ and target $\boldsymbol{t}$ assuming a feature construction ϕ for tweet x (e.g. bag-of-words)
- fails for new targets (Trump vs Hillary)

Idea: use the target $\boldsymbol{t}$ in feature construction ϕ

$$\hat{y} = \underset{y \in \mathcal{Y}}{\arg\max}(\mathbf{w}^y \cdot \phi(x, t; \theta))$$

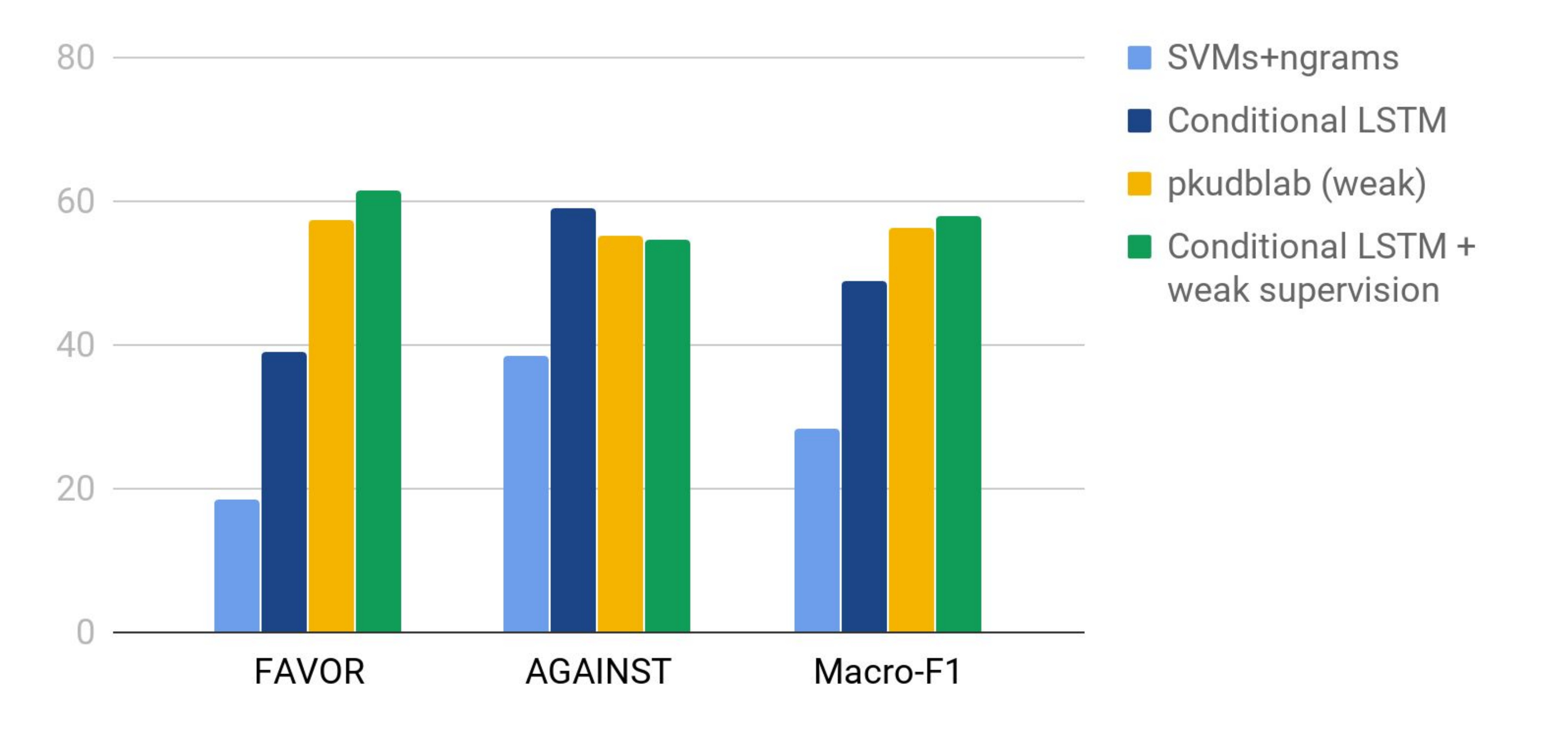
Learn the parameters $\boldsymbol{\theta}$ constructing the feature representation jointly with $\mathbf{w}$ using Long Short Term Memory Networks (LSTMs)

Stance Classification with Conditional LSTMs



- One LSTM encodes the target, another LSTM the tweet
- The representation of the tweet is **conditioned** on the target
- Same tweet-different target ⇒ **different stance**

Results



- Train on stance-annotated tweets for 5 targets, test on Trump
- State-of-the-art results without training data for target and with weak supervision (Augenstein et al., EMNLP 2016)

Zero-shot Relation Classification

Relation	Subject (**X**)	Object (**Y**)	Text (Premise)	Description (Hypothesis)
religious_order	Lorenzo Ricci	Society of Jesus	**X** (August 1, 1703 – November 24, 1775) was an Italian Jesuit, elected the 18th Superior General of the **Y**.	***X*** *was a member of the group* ***Y***
director	Kispus	Erik Balling	**X** is a 1956 Danish romantic comedy written and directed by **Y**.	*The director of* ***X*** *is* ***Y***
designer	Red Baron II	Dynamix	**X** is a computer game for the PC, developed by **Y** and published by Sierra Entertainment.	***Y*** *is the designer of* ***X***

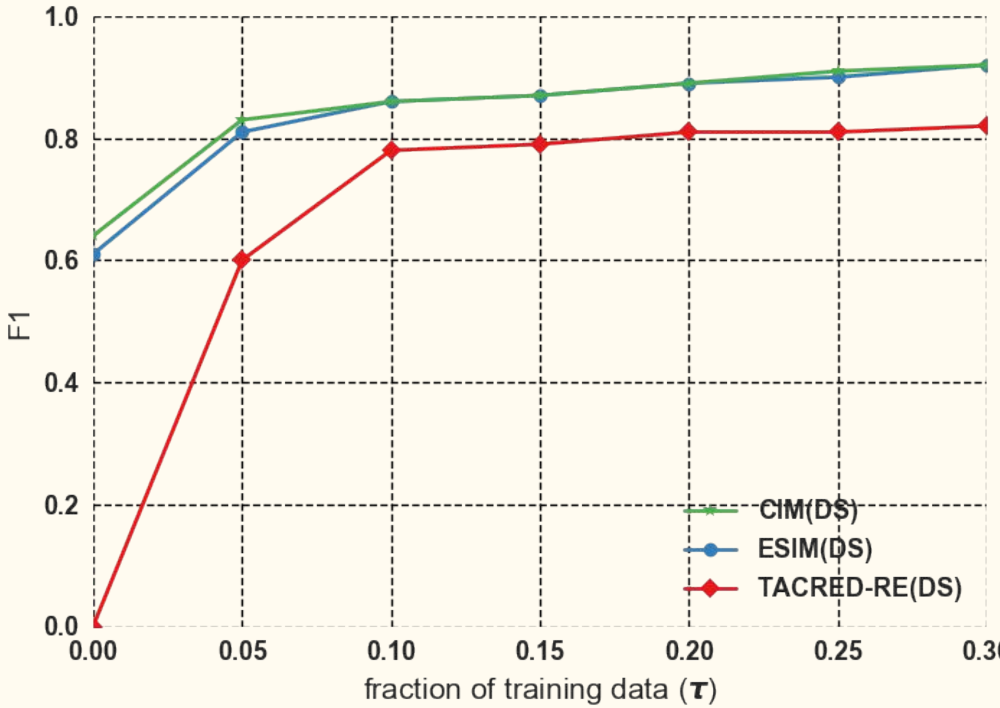
Extended relation classification using descriptions instead of labeled data (Obamuyide and Vlachos, under review):

- Given training for `director` relation, we can predict `designer`
- Formulated the task textual entailment (sentence-pair classification)

Results

Dataset	Model	F1 (%)
LMU-RC	ESIM	20.16
	CIM	**22.20**
UW-RE	ESIM	61.32
	CIM	**63.58**

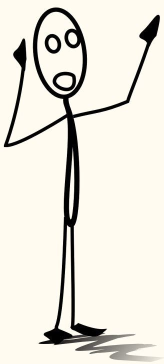
- Good results on two datasets, improved using conditional encoding
- Can use labeled training data if available



Automated fact checking

A new challenge for AI: Automated fact-checking

The United Kingdom has ten times Italy's number of immigrants.



Country/ Immigration	Italy	UK
2014	4.92M	5.05M
2015	5.01M	5.42M
2016	5.03M	5.64M

FALSE: We find no data to support this claim. The UK does not have "ten times Italy's number of immigrants".

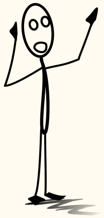
(Vlachos and Riedel, 2014)

What do we want from automated fact-checking?

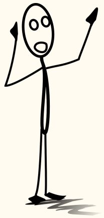
- Verdict justification, a.k.a. algorithmic transparency
  - Can’t convince otherwise
  - Need to check their correctness
- Generalization to different domains (economy, health, etc.)
- Learn with (relatively) little data

(Vlachos and Riedel, 2014)

What claims should we fact-check?



Syrian refugees are not properly vetted or tracked by the FBI once in the US



Leaving the EU would put 3M jobs at risk

- Does the source of the claim matter?
- Does the linguistic style matter?

Evidence for or against a claim

False

Claim: Doctors confirmed the first case of death by genetically modified food

Tagged: Fake News Hoaxes World News Daily Report

Resolved Added Mar 9

It originated on a fake news website and is therefore false. Emergent is as of now the only site to offer a full debunking.

Sources

Sources Tracked: **3** Total Shares: **62,188**

For	Against
2	1
Shares	Shares
60,596	**1,592**

Results

- 300 claims from debunking website www.emergent.info
- Automated stance classification with 73% accuracy *(Ferreira and Vlachos, 2016)*
- Advisor to the Fake News Challenge with 50 participants

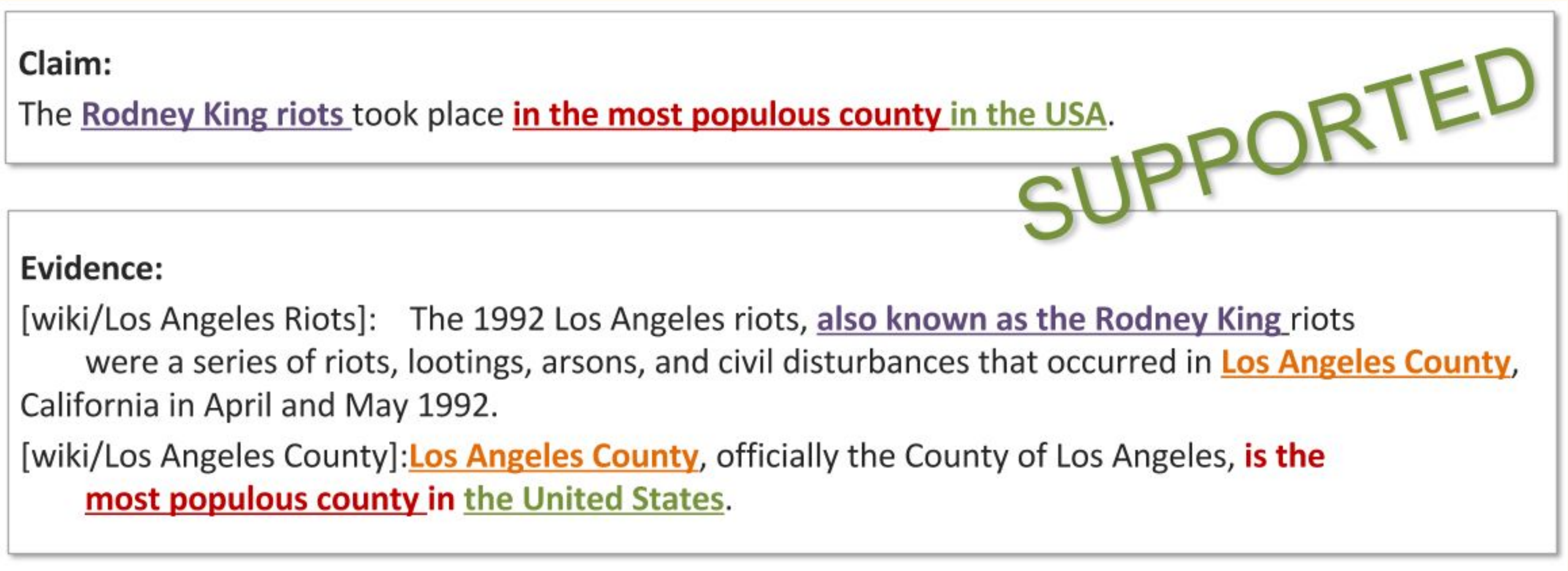
New datasets needed

AI successes follow dataset availability *(Wissner-Gross, 2016)*

Year	Breakthroughs in AI	Datasets (First Available)	Algorithms (First Proposed)
1994	Human-level spontaneous speech recognition	Spoken Wall Street Journal articles and other texts (1991)	Hidden Markov Model (1984)
1997	IBM Deep Blue defeated Garry Kasparov	700,000 Grandmaster chess games, aka "The Extended Book" (1991)	Negascout planning algorithm (1983)
2005	Google's Arabic- and Chinese-to-English translation	1.8 trillion tokens from Google Web and News pages (collected in 2005)	Statistical machine translation algorithm (1988)
2011	IBM Watson became the world Jeopardy! champion	8.6 million documents from Wikipedia, Wiktionary, Wikiquote, and Project Gutenberg (updated in 2010)	Mixture-of-Experts algorithm (1991)
2014	Google's GoogLeNet object classification at near-human performance	ImageNet corpus of 1.5 million labeled images and 1,000 object categories (2010)	Convolution neural network algorithm (1989)
2015	Google's Deepmind achieved human parity in playing 29 Atari games by learning general control from video	Arcade Learning Environment dataset of over 50 Atari games (2013)	Q-learning algorithm (1992)
Average No. of Years to Breakthrough:		**3 years**	**18 years**

300 claims are not enough to learn fact checking

Fact Extraction and VERification (FEVER)



- 200K claims verified on Wikipedia (Thorne et al., NAACL 2018)
- 3-way classification:
    - The claim is **SUPPORTED** by the evidence
    - The claim is **REFUTED** by the evidence
    - **NOT ENOUGH INFORMATION** in Wikipedia to verify it

Annotation process details

- 50 annotators, all native speakers, trained by the authors or more experienced annotators
- Fixed Wikipedia dump to avoid changes in labels
- One annotator constructs the claim, different annotator verifies it
- Dedicated user interfaces were developed for the task
- Guidelines were refined through pilot studies
- Advised to spend 2-3 minutes per claim
- Instructed to avoid using their own world knowledge: “Shakira is Canadian” is NOT ENOUGH INFORMATION

Annotation findings

- 0.68 in Fleiss Kappa inter-annotator agreement on 3.4K claims
- 96.12% precision and 74.84% recall in evidence retrieval: measured against annotators who were not time-constrained
- Claims were 7.9 tokens long
- Multi-sentence evidence was chosen for 28.04% of the claims
- Evidence from different pages was chosen for 11.47%
- 7.6% of the mutated claims were excluded due to being too vague/ambiguous
- Final verification by the authors: 91.2% correct on 227 claims.

Results

Unlike previous tasks and datasets, **evidence** matters:

- a correct label with incorrect supporting evidence is wrong
- a simple approach using TF-IDF-based similarity for evidence selection and LSTMs for labeling the claim given the evidence achieved 31.87% acc. (50.91% ignoring evidence)

Room for improvement:

Fact Extraction and Verification (FEVER) shared task

- EMNLP 2018 workshop with Amazon Research Cambridge and Imperial College
- Interest from academics, industry and journalists and **you**?

Research summary

- Imitation learning for structured prediction in NLP
- Zero-shot learning with neural networks
- Automated fact-checking (see our Coling 2018 survey)

Other work:

- active learning (CSL 2008)
- Bayesian non-parametric approaches for NLP (PhD)
- syntax-based neural language models (ACL 2015, with Piotr Mirowski from Google DeepMind)
- authorship attribution with neural networks (EACL 2017, Coling 2018)

Thanks to my collaborators and sponsors

Looking forward to Cambridge from October!

Questions?