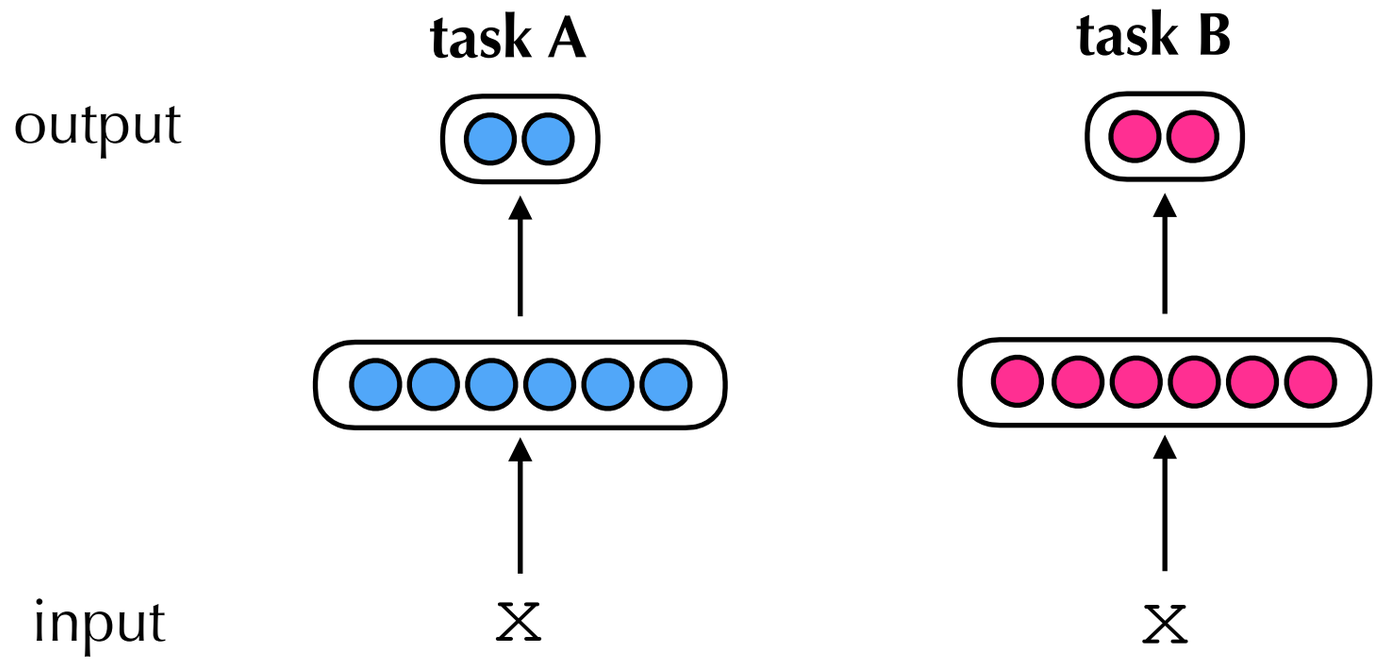
When is multi-task learning effective?

Semantic sequence prediction under varying data conditions

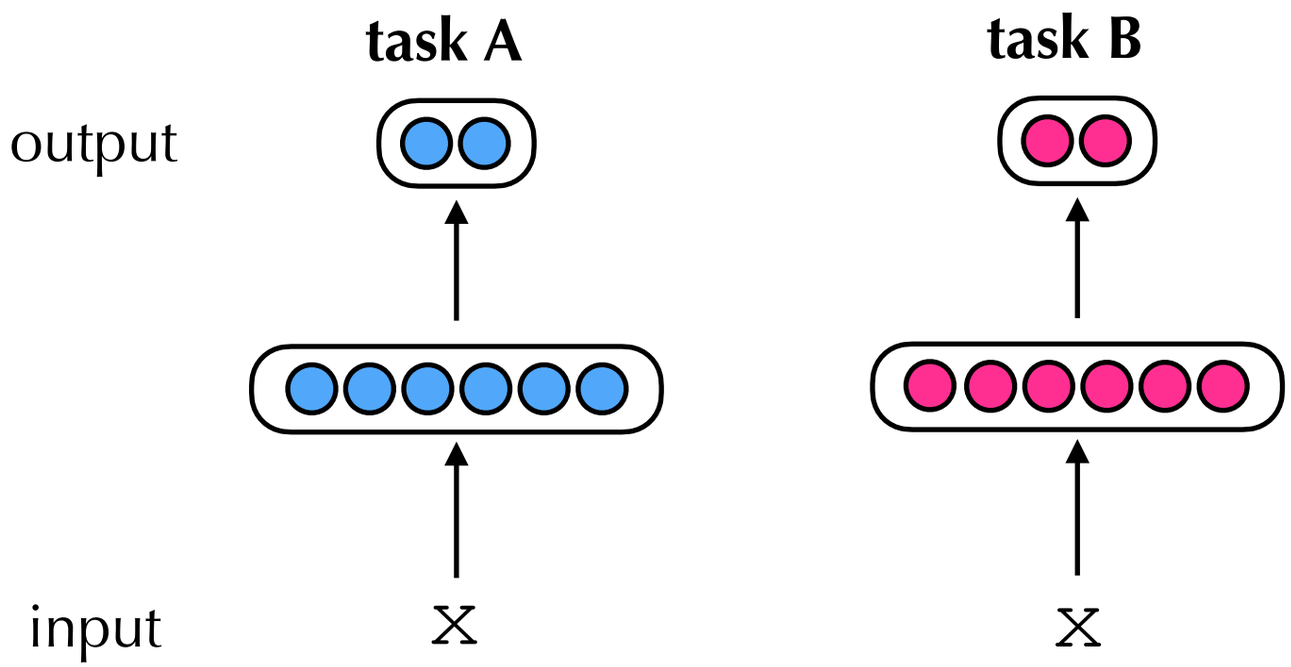
Héctor Martínez Alonso and Barbara Plank
INRIA (France) and Univ. of Groningen (Netherlands)

Typical single-task learning



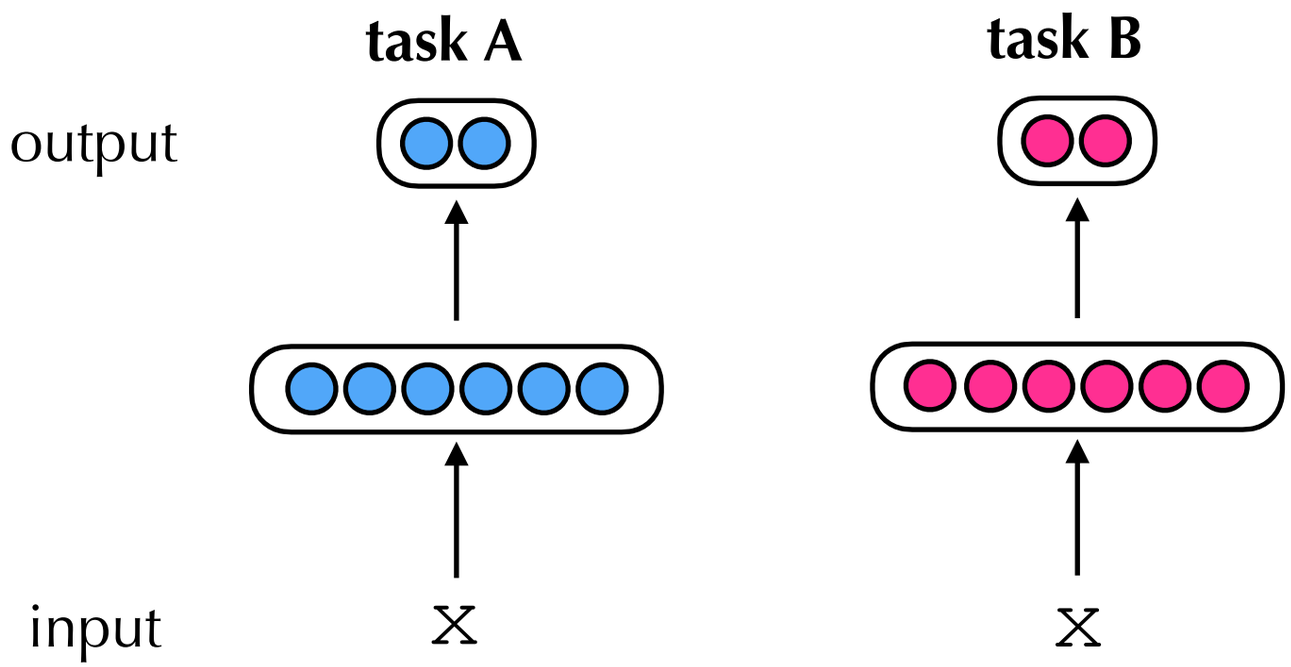
Can we do better?

Multi-task Learning (MTL): Key Idea



single-task learning (STL)

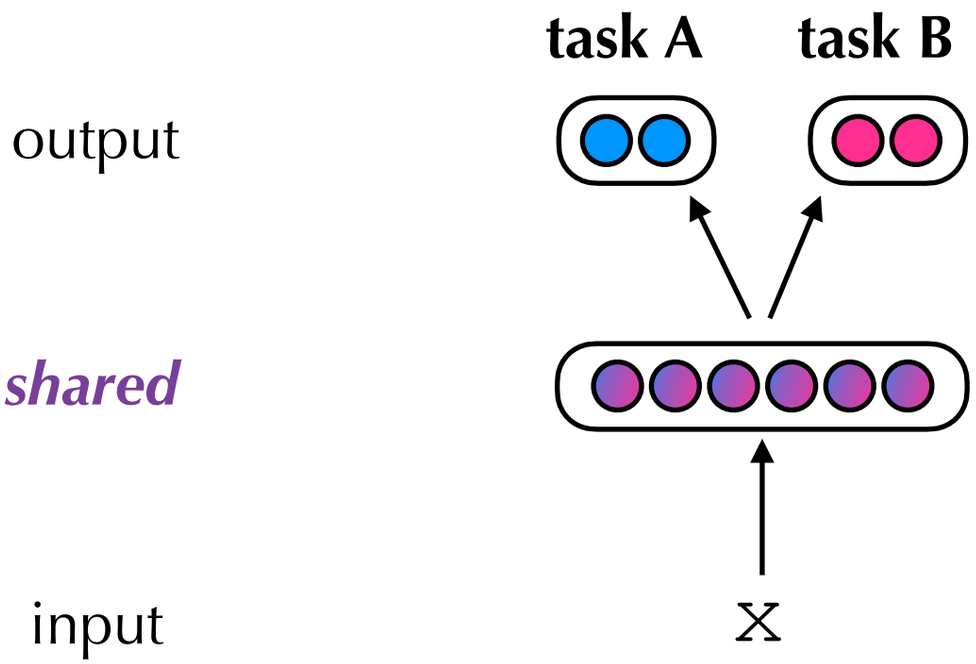
Multi-task Learning (MTL): Key Idea



single-task learning (STL)

"**learning tasks in parallel** while using a **shared representation**; what is learned for each task **can help other tasks be learned better**" (Caruana, 1997)

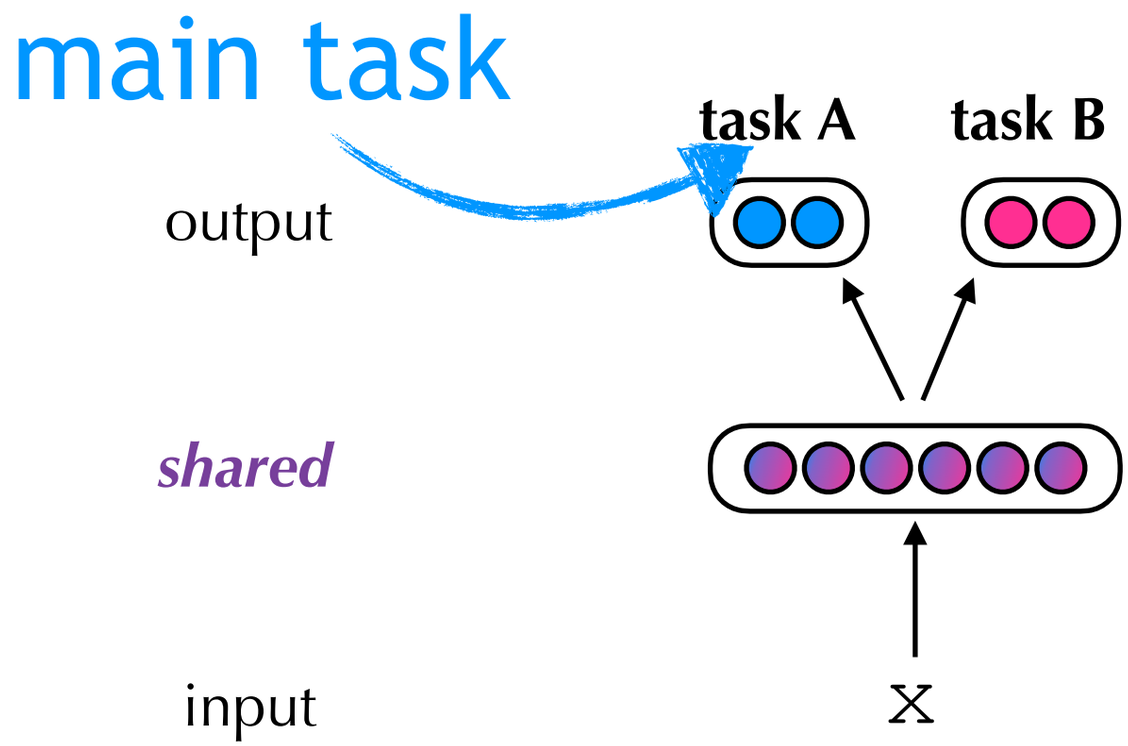
Multi-task Learning (MTL): Key Idea



multi-task learning (MTL)

"**learning tasks in parallel** while using a **shared representation**; what is learned for each task **can help other tasks be learned better**" (Caruana, 1997)

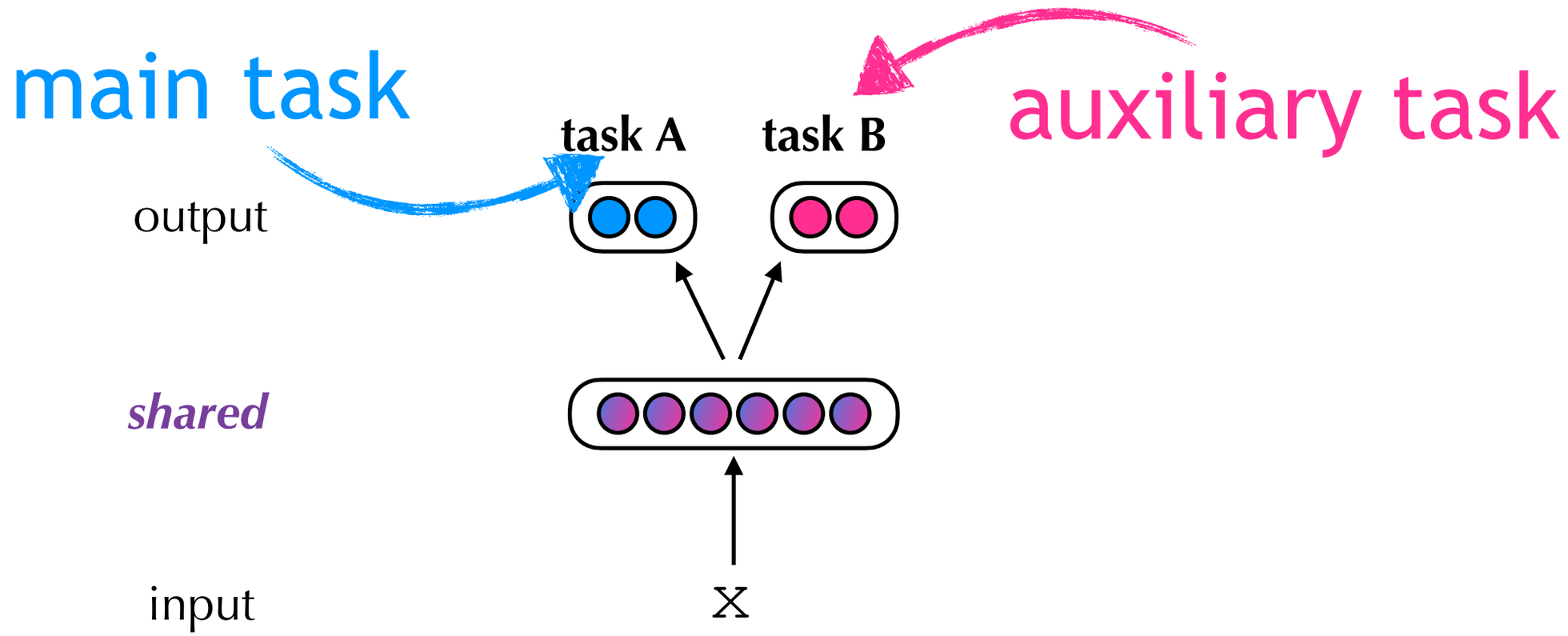
Multi-task Learning (MTL): Key Idea



multi-task learning (MTL)

"**learning tasks in parallel** while using a **shared representation**; what is learned for each task **can help other tasks be learned better**" (Caruana, 1997)

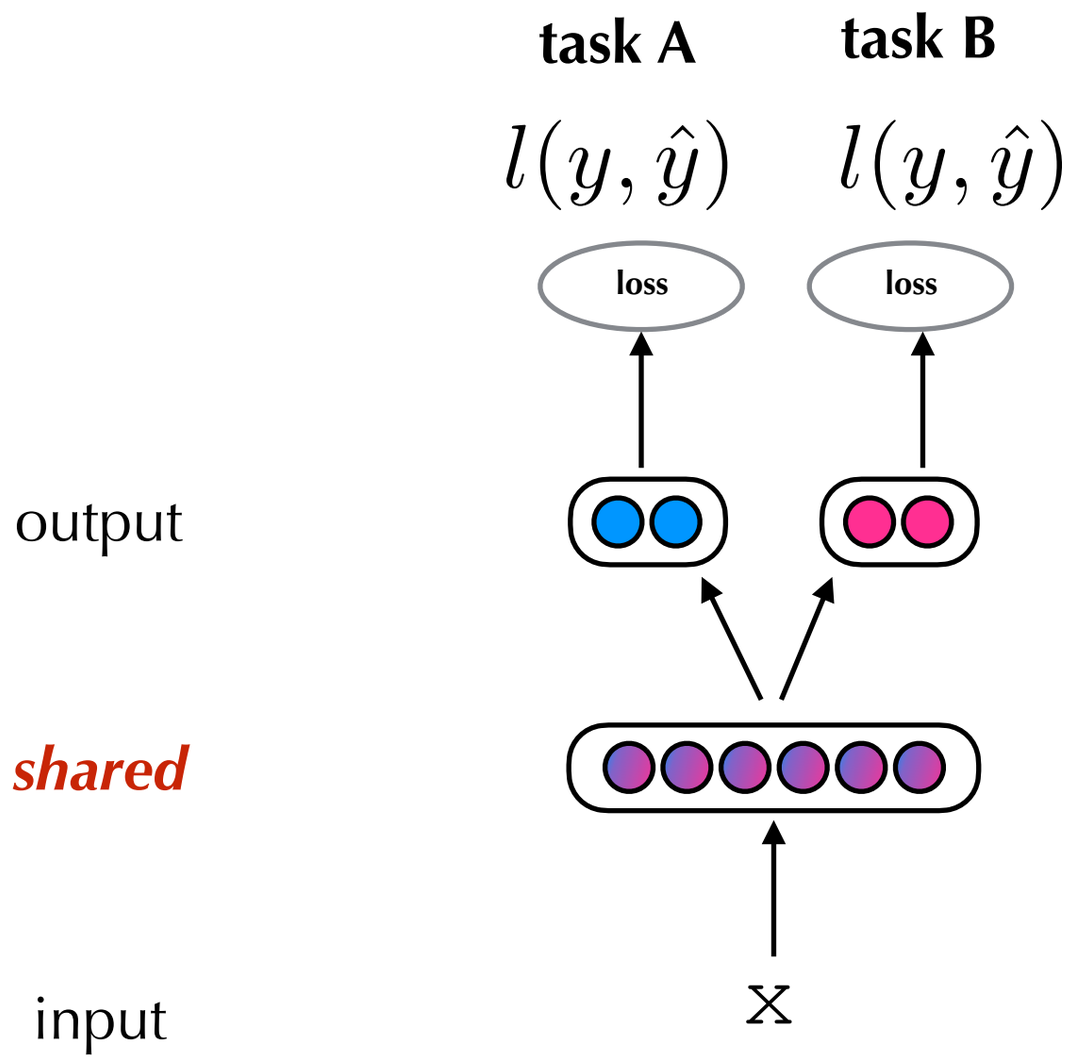
Multi-task Learning (MTL): Key Idea



multi-task learning (MTL)

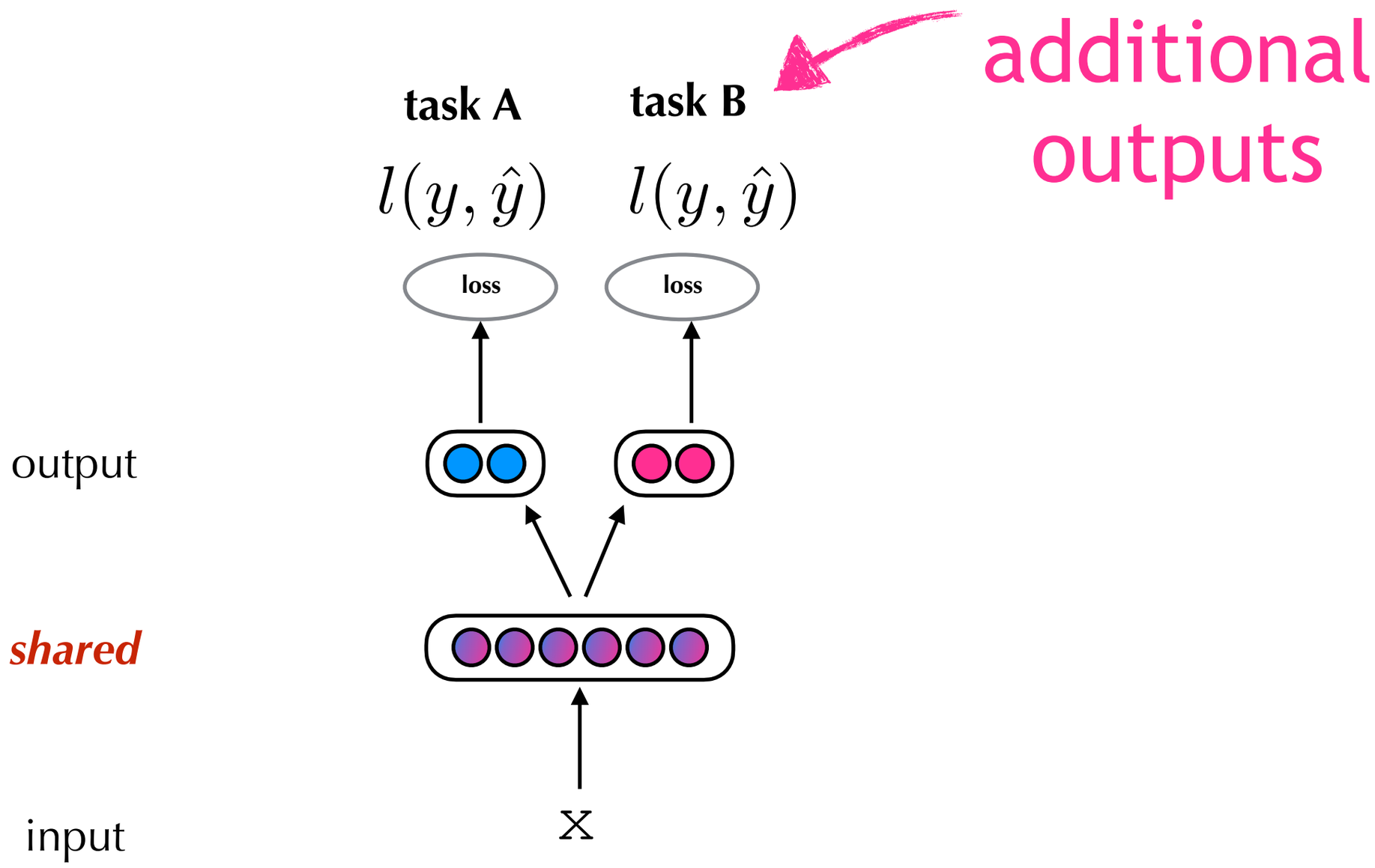
"**learning tasks in parallel** while using a **shared representation**; what is learned for each task **can help other tasks be learned better**" (Caruana, 1997)

MTL in Neural Networks (NNs)



NNs make MTL particularly attractive/easy

MTL in Neural Networks (NNs)



NNs make MTL particularly attractive/easy

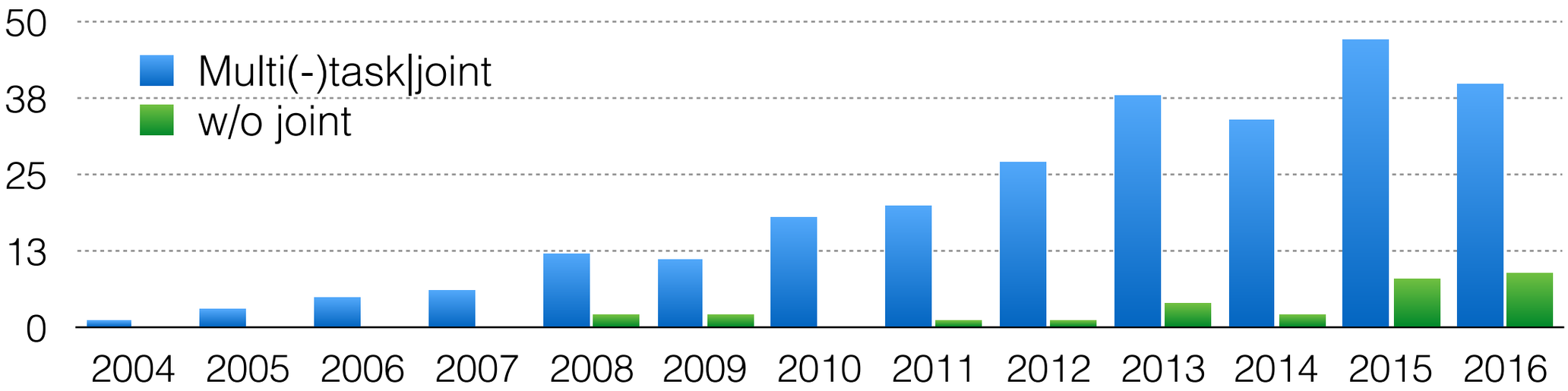
Multi-task learning in NLP

Titles in ACL anthology (from 2004)

*(incl. variants of RNN/CNNs and excl. deep parsing)

Multi-task learning in NLP

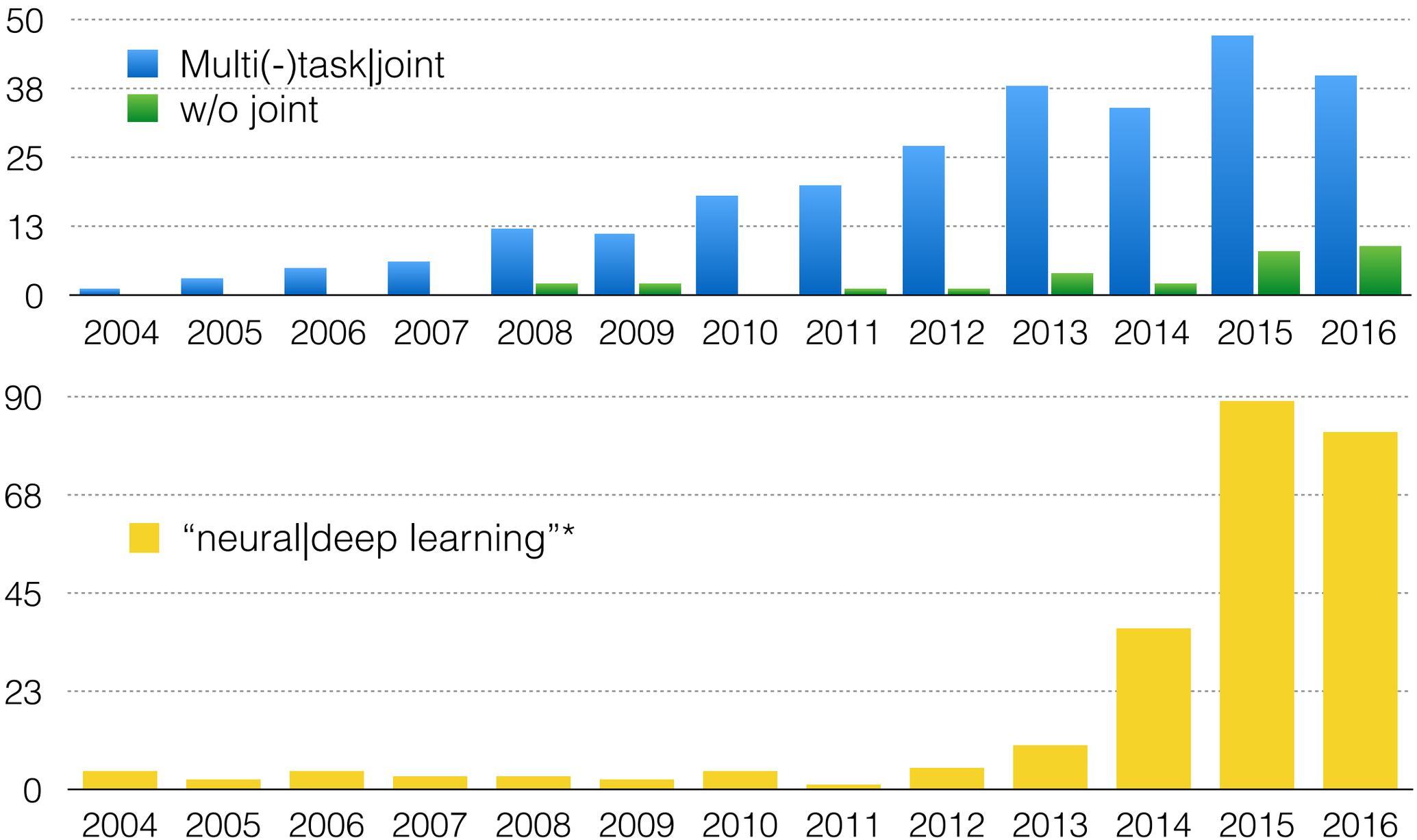
Titles in ACL anthology (from 2004)



*(incl. variants of RNN/CNNs and excl. deep parsing)

Multi-task learning in NLP

Titles in ACL anthology (from 2004)

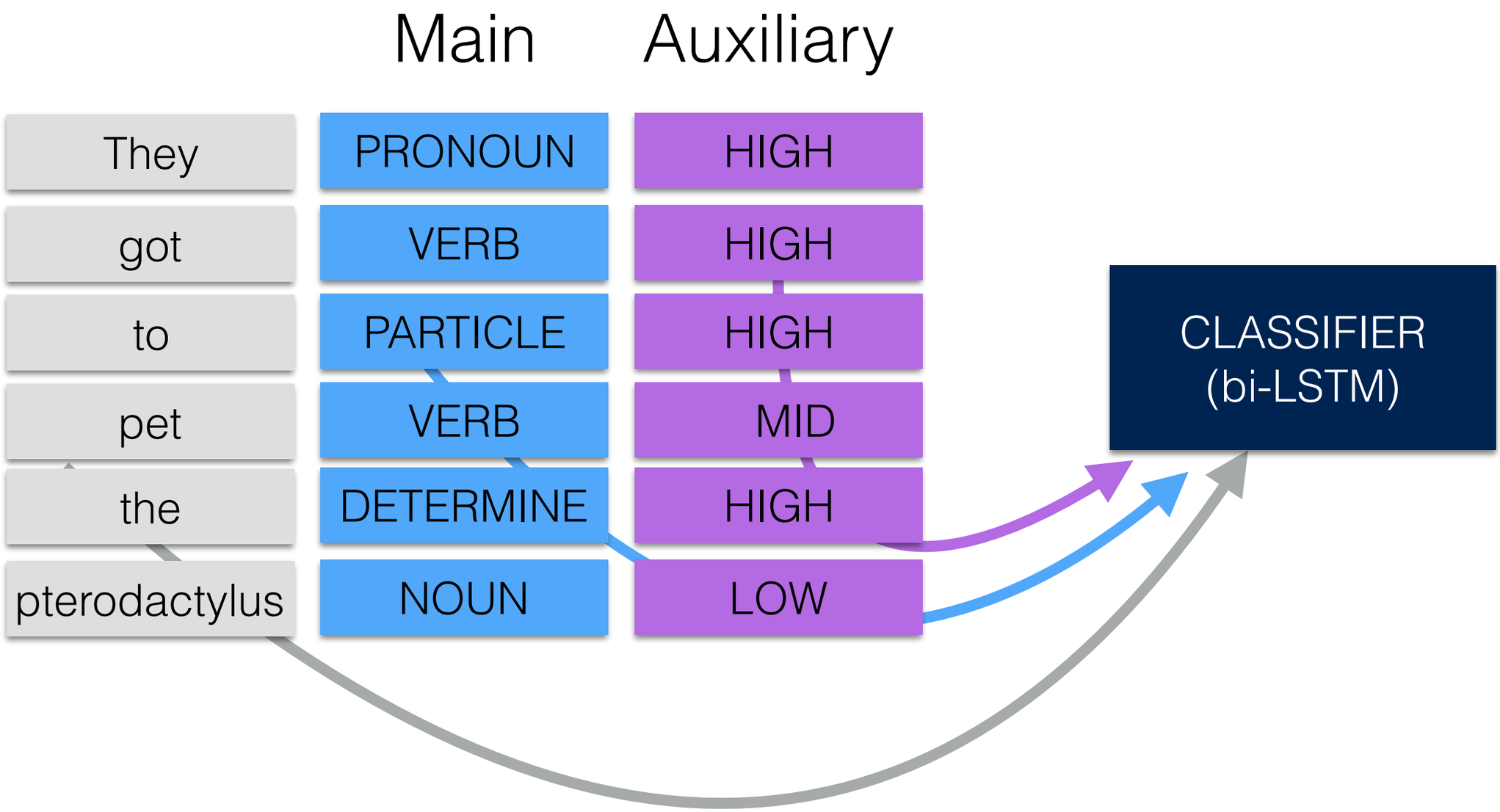


*(incl. variants of RNN/CNNs and excl. deep parsing)

Perspectives on MTL

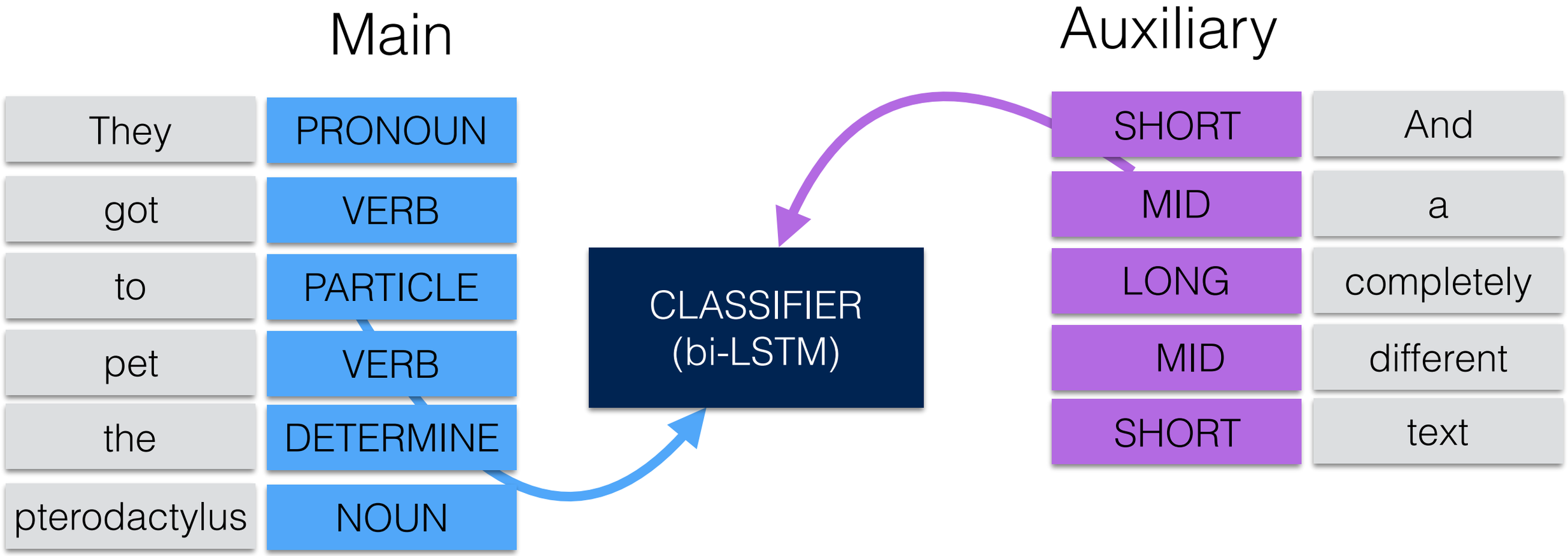
MTL: learning from distinct views

e.g., **FreqBin** of Plank et al. (2016),
Braud et al. (2016)



MTL: learning from distinct sources

e.g., from cognitive behavioral data
Klerke et al. (2016), Plank (2016)



However, *when* does it work?

However, *when* does it work?

E.g. POS+Word Freq in Plank et al. (2016)

However, *when* does it work?

Semantic sequence prediction under varying data conditions

Information-Theoretic Measures

Information-Theoretic Measures

Label distribution Y:

Information-Theoretic Measures

Label distribution Y:

- **Entropy** H(Y)

Information-Theoretic Measures

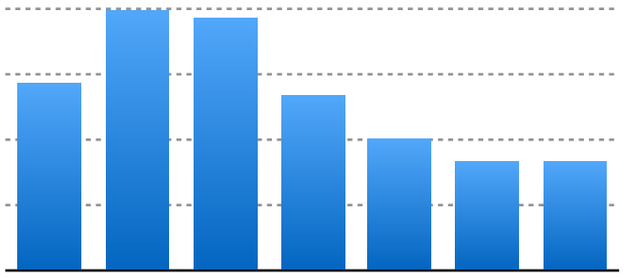
Label distribution Y:

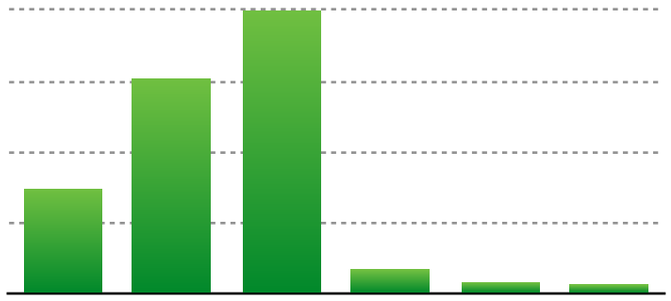
- **Entropy** H(Y)
- **Kurtosis** k(Y) (tailedness)

Information-Theoretic Measures

Label distribution Y:

- **Entropy** H(Y)
- **Kurtosis** k(Y) (tailedness)

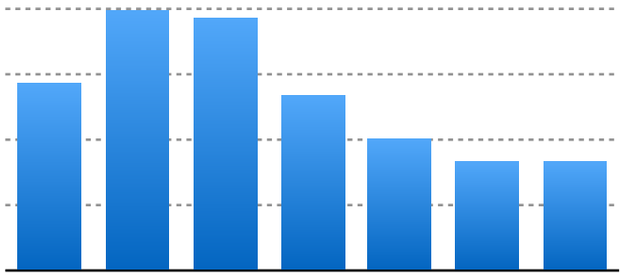




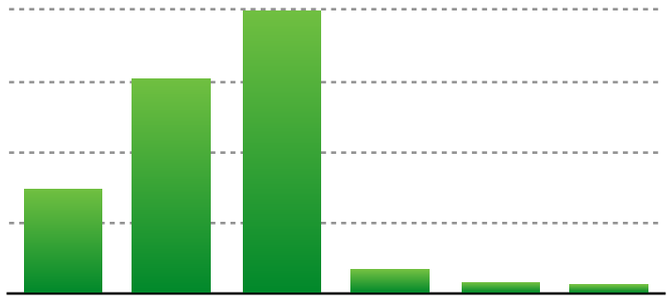
Information-Theoretic Measures

Label distribution Y:

- **Entropy** H(Y)
- **Kurtosis** k(Y) (tailedness)



higher entropy

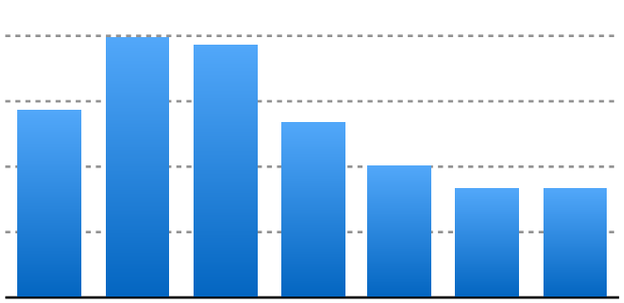


lower entropy

Information-Theoretic Measures

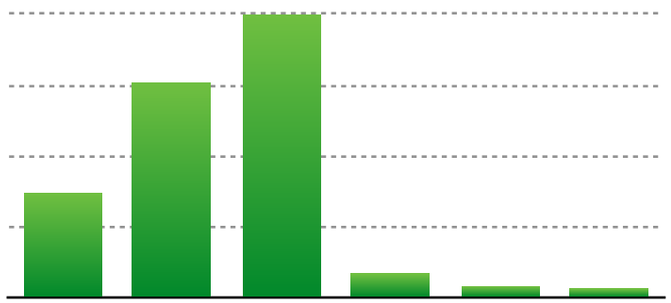
Label distribution Y:

- **Entropy** H(Y)
- **Kurtosis** k(Y) (tailedness)



higher entropy

lower kurtosis



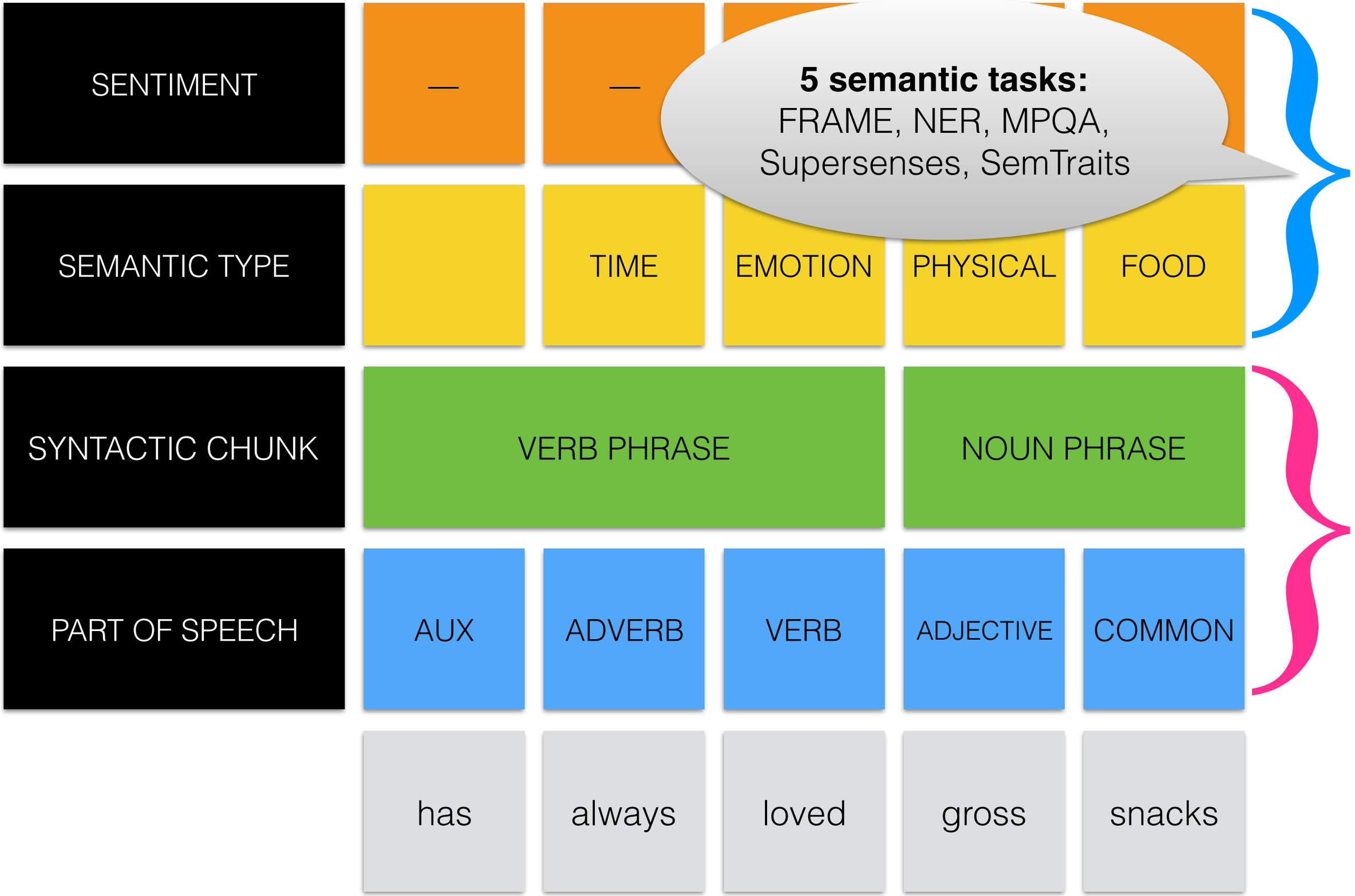
lower entropy

higher kurtosis

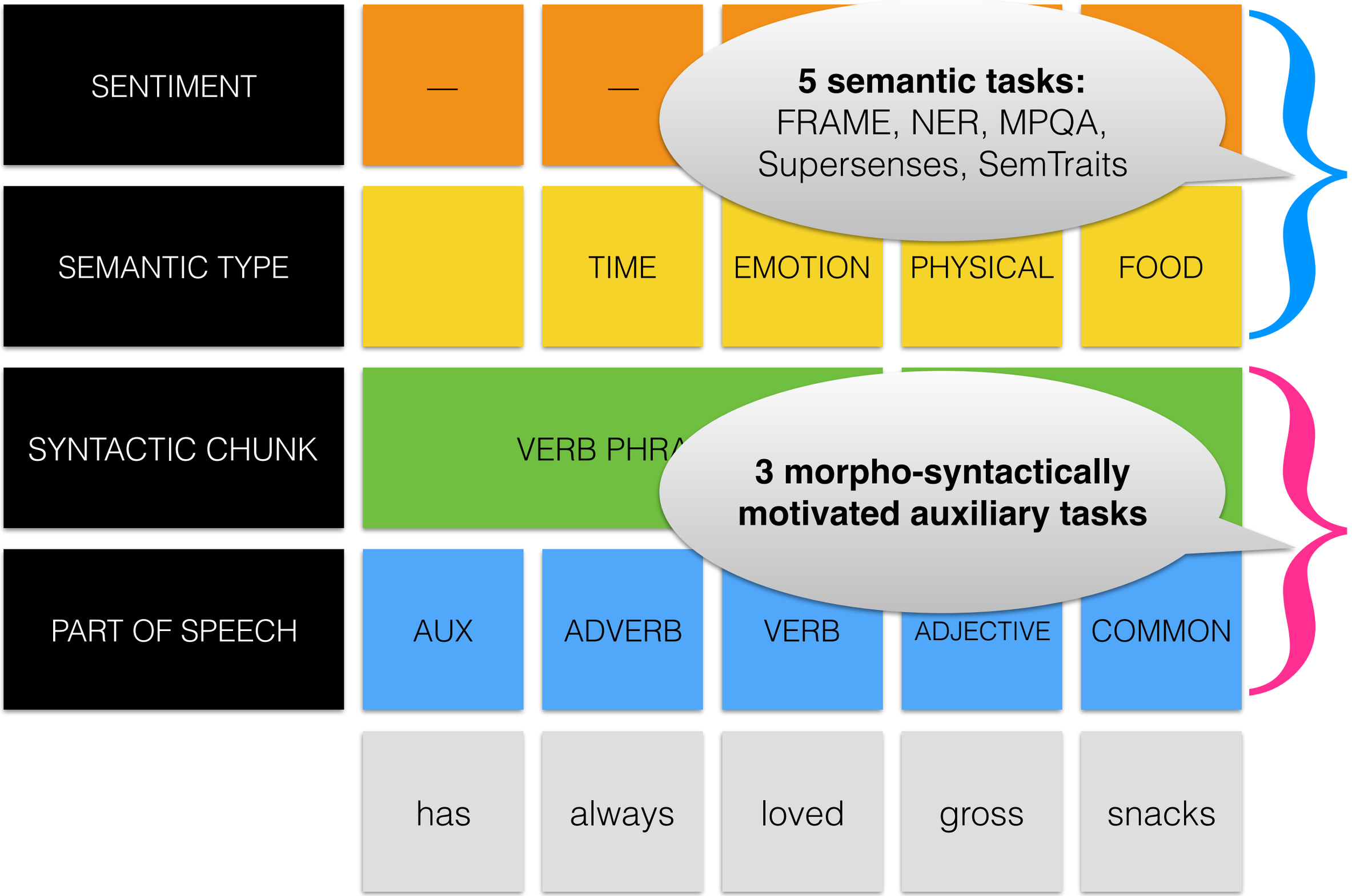
Sequence labeling tasks

SENTIMENT	—	—	POSITIVE	NEGATIVE	—
SEMANTIC TYPE		TIME	EMOTION	PHYSICAL	FOOD
SYNTACTIC CHUNK	VERB PHRASE			NOUN PHRASE	
PART OF SPEECH	AUX	ADVERB	VERB	ADJECTIVE	COMMON
	has	always	loved	gross	snacks

Sequence labeling tasks



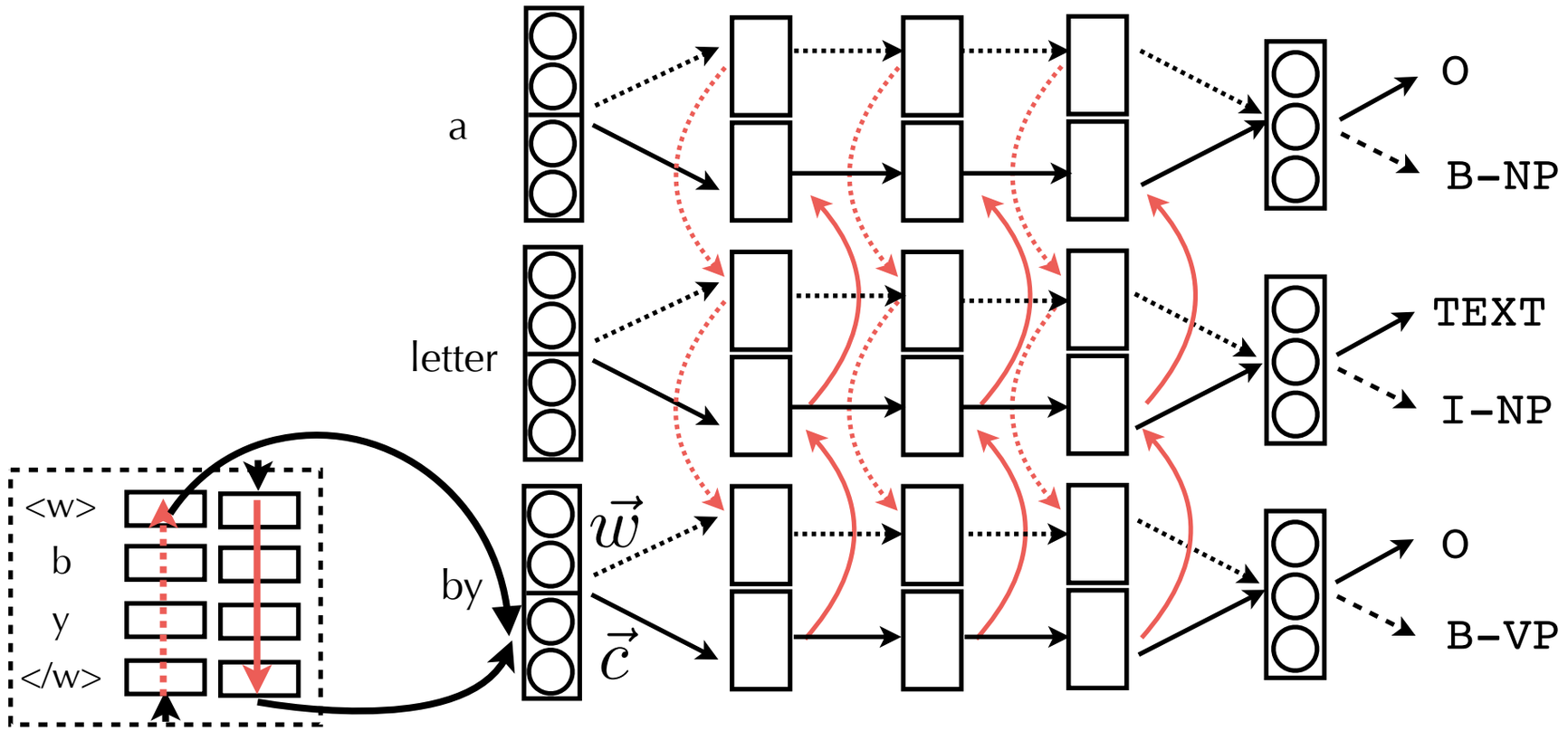
Sequence labeling tasks



Three auxiliary tasks

SYNTACTIC CHUNK	VERB PHRASE			NOUN PHRASE	
PART OF SPEECH	AUX	AVD	VERB	ADJ	NOUN
	has	always	loved	gross	snacks
WORD FREQUENCY (FreqBin)	HIGH	HIGH	MID	LOW	MID

Model: Bi-LSTM with auxiliary loss



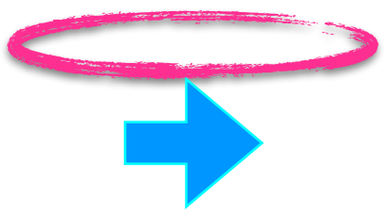
- Baseline: without auxiliary tasks
- Auxiliary tasks: one {Aux} or combination Freqbin+{Aux}
- Goal: delimit behavior of bi-LSTM & interaction main+{Aux} tasks

Analysis 1:

Main task data properties

Data properties of Main Tasks

	$\|Y\|$	prop of O	k(Y)	H(Y_{full})
FRAMES	707	.80	701.41	1.60
MPQA	9	.65	2.79	1.12
NER	9	.83	4.10	0.77
SEMTRAITS	11	.66	5.68	1.29
SUPERSENSES	83	.66	76.73	1.84



high entropy + kurtosis

low entropy + kurtosis

Data properties of Main Tasks

	$\|Y\|$	prop of O	k(Y)	H(Y_{full})
FRAMES	707	.80	701.41	1.60
MPQA	9	.65	2.79	1.12
NER	9	.83	4.10	0.77
SEMTRAITS	11	.66	5.68	1.29
SUPERSENSES	83	.66	76.73	1.84

high entropy + kurtosis

low entropy + kurtosis

Data properties of Main Tasks

	$\|Y\|$	prop of O	k(Y)	H(Y_{full})
FRAMES	707	.80	701.41	1.60
MPQA	9	.65	2.79	1.12
NER	9	.83	4.10	0.77
SEMTRAITS	11	.66	5.68	1.29
SUPERSENSES	83	.66	76.73	1.84

high entropy + kurtosis

low entropy + kurtosis

Overall best Main-Aux combinations

	BL	ΔBest	Description	aux layer	# over
FRAMES	38.93	-8.13	+FREQBIN	outer	0
MPQA	28.26	**0.96**	+POS+FREQBIN	inner	2
NER	90.60	-0.58	+FREQBIN	inner	0
SEMTRAITS	70.42	**1.24**	+FREQBIN	outer	13
SUPERSENSES	62.36	-0.13	+POS+FREQBIN	inner	0

Table 2: Baseline (BL) and best system performance difference (Δ) for all main tasks—

Overall best Main-Aux combinations

	BL	ΔBest	Description	aux layer	# over
FRAMES	38.93	-8.13	+FREQBIN	outer	0
MPQA	28.26	**0.96**	+POS+FREQBIN	inner	2
NER	90.60	-0.58	+FREQBIN	inner	0
SEMTRAITS	70.42	**1.24**	+FREQBIN	outer	13
SUPERSENSES	62.36	-0.13	+POS+FREQBIN	inner	0

Table 2: Baseline (BL) and best system performance difference (Δ) for all main tasks—

- Improvements on two tasks, only one significant

Overall best Main-Aux combinations

	BL	ΔBest	Description	aux layer	# over
FRAMES	38.93	-8.13	+FREQBIN	outer	0
MPQA	28.26	**0.96**	+POS+FREQBIN	inner	2
NER	90.60	-0.58	+FREQBIN	inner	0
SEMTRAITS	70.42	**1.24**	+FREQBIN	outer	13
SUPERSENSES	62.36	-0.13	+POS+FREQBIN	inner	0

Table 2: Baseline (BL) and best system performance difference (Δ) for all main tasks—

- Improvements on two tasks, only one significant
- Bi-LSTM will not benefit from auxiliary loss if there are many labels and entropy is high

Analysis 2:

Aux task data properties

Frequency as an auxiliary task for part-of-speech prediction

- Original frequency binning from Plank et al (2016) :

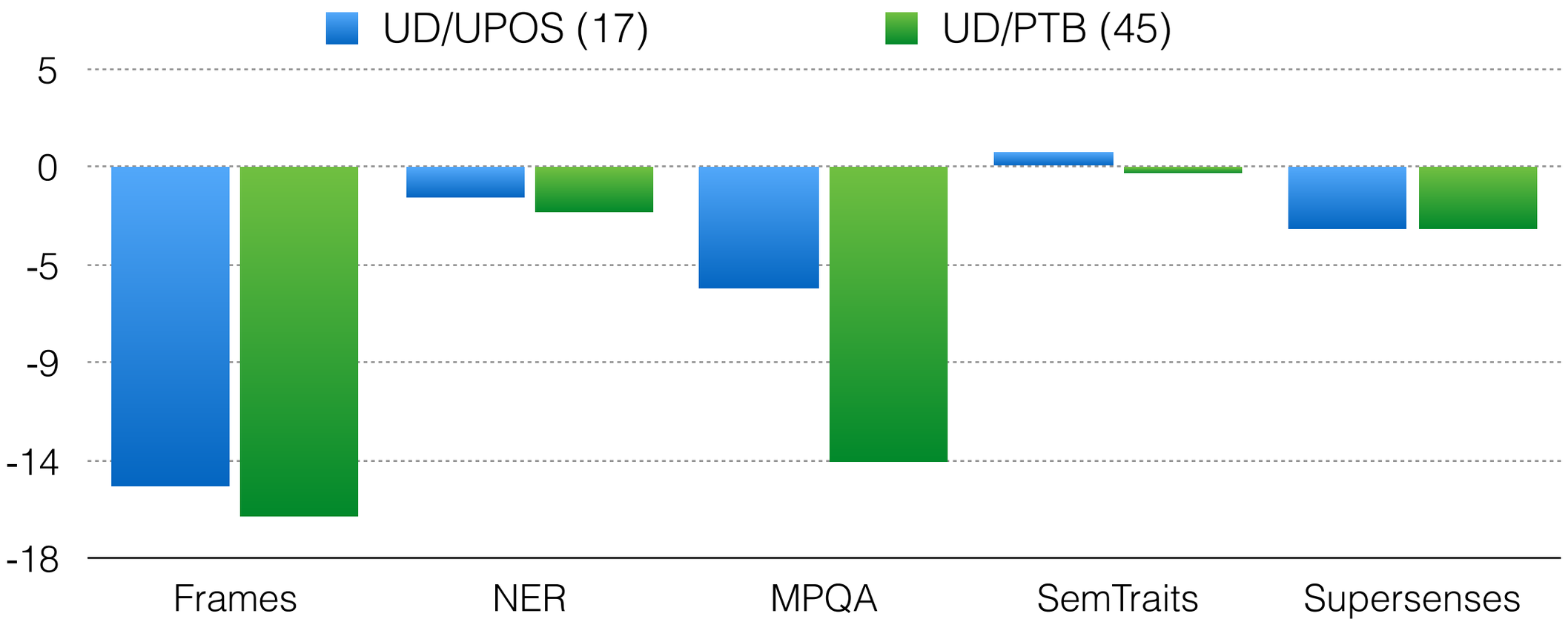
```
int(log10(freq(w)))
```

- We propose instead a uniform distribution calculated from the cumulative word frequencies.

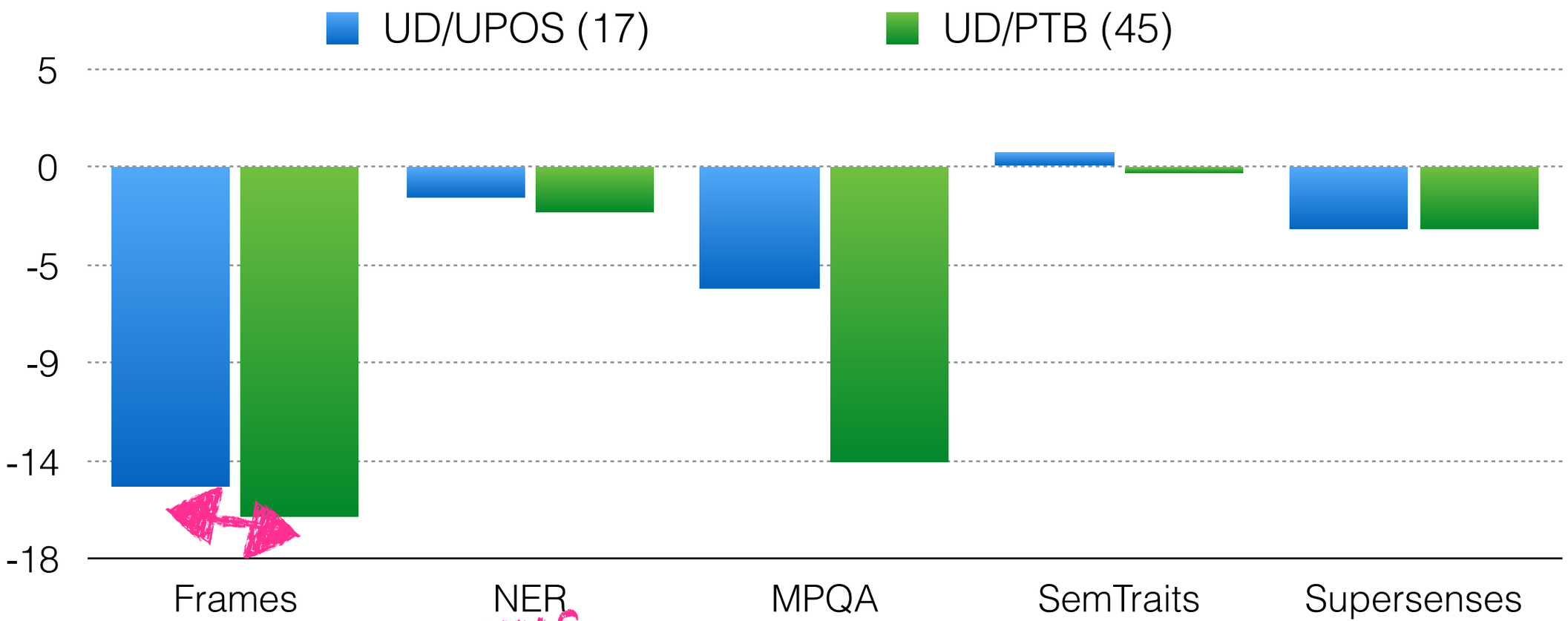
- Resulted in best freqbin instantiation (reported before).

  - Benefits from higher entropy on the aux-label distribution.

PTB- and UD-PoS as Aux task

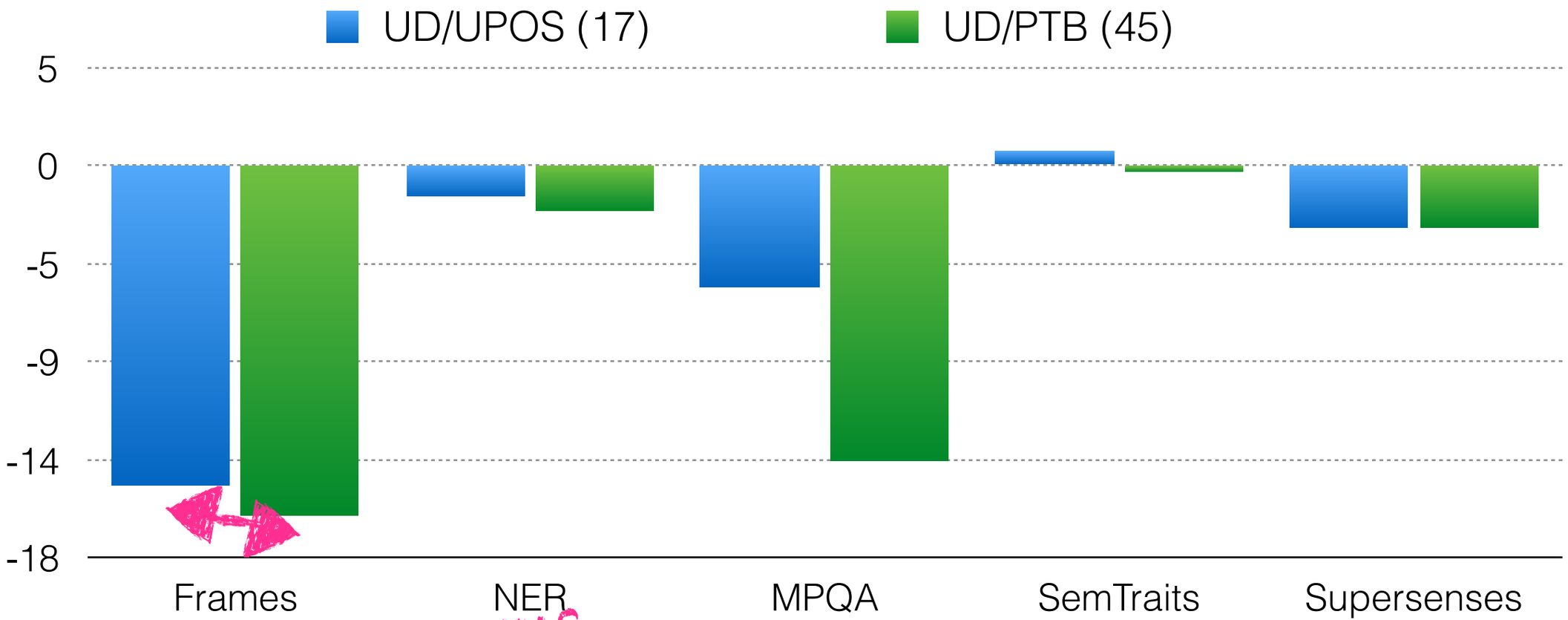


PTB- and UD-PoS as Aux task



Obs! same corpus, tagset changes!

PTB- and UD-PoS as Aux task



Obs! same corpus, tagset changes!

- Preference of architecture to more compact auxiliary task distributions (fewer labels, lower kurtosis)

One small step

One small step

- This study is necessarily incomplete.

One small step

- This study is necessarily incomplete.
- One architecture, only 5 main tasks, 3 auxiliary tasks, and limited parameter space exploration.

One small step

- This study is necessarily incomplete.
- One architecture, only 5 main tasks, 3 auxiliary tasks, and limited parameter space exploration.
- However, it is—to the best of our knowledge— **the first to relate data-dependent conditions with performance measures in MTL.**

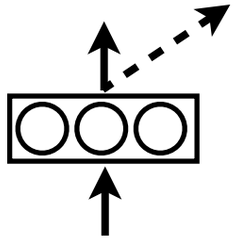
One small step

- This study is necessarily incomplete.
- One architecture, only 5 main tasks, 3 auxiliary tasks, and limited parameter space exploration.
- However, it is—to the best of our knowledge— **the first to relate data-dependent conditions with performance measures in MTL.**
- We're happy to see follow-up work, e.g., Bingel & Søgaard (2017, this EACL!) and Bjerva (2017, NoDaLiDa).

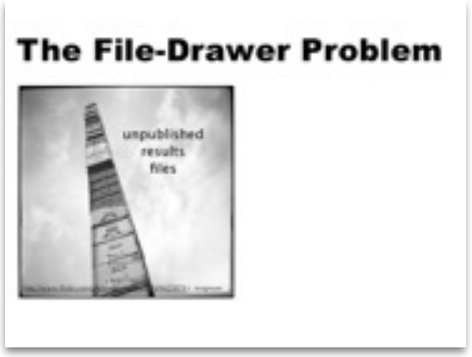
Take-home message

@barbara_plank

Take-home message

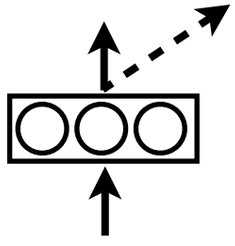


- Multi-task learning is promising, but not always effective.

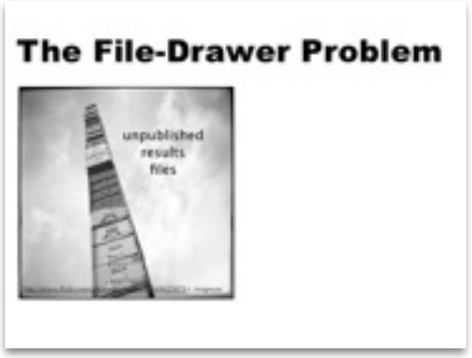


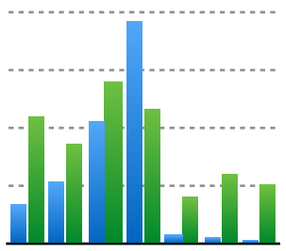
@barbara_plank

Take-home message



- Multi-task learning is promising, but not always effective.

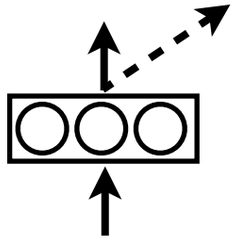




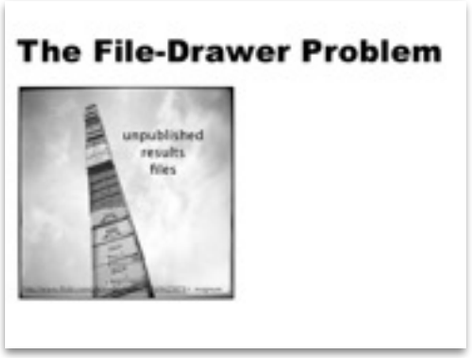
- Information-theoretic properties of main and auxiliary tasks give some clues a priori.
- Current architecture prefers compact distributions.

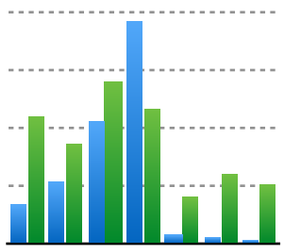
@barbara_plank

Take-home message



- Multi-task learning is promising, but not always effective.



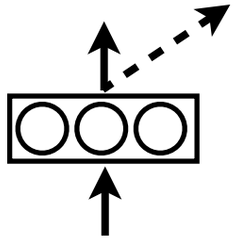


- Information-theoretic properties of main and auxiliary tasks give some clues a priori.
- Current architecture prefers compact distributions.

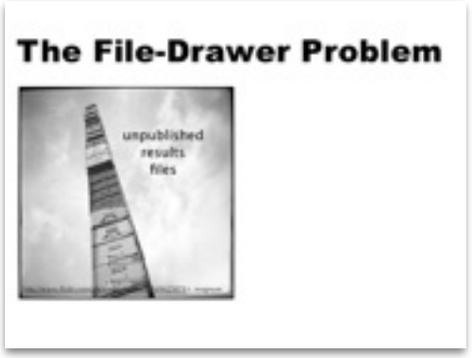
- More work needed!

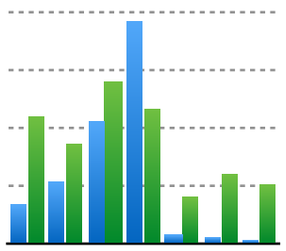
@barbara_plank

Take-home message



- Multi-task learning is promising, but not always effective.





- Information-theoretic properties of main and auxiliary tasks give some clues a priori.
- Current architecture prefers compact distributions.

- More work needed!

Thanks!

@barbara_plank

Appendix

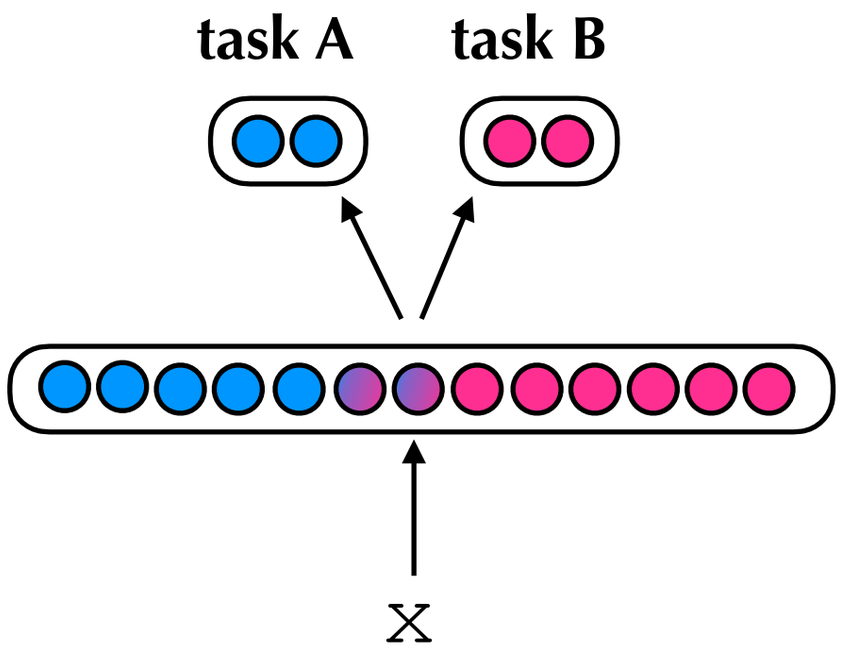
Identifying co-informativeness

	$\|Y\|$	BL	ΔU	R^2
FRAMES	707	38.93	-8.13	.00
MPQA	9	28.26	**0.44**	.09
NER	9	90.60	-1.31	.26
SEMTRAITS	11	70.42	**1.12**	.44
SUPERSENSES	83	62.36	-0.69	.47
CHUNK	22	94.76	-0.14	.49
POS	17	94.35	**0.21**	.68
DEPRELS	47	88.70	-0.16	.64

Table 4: Label inventory size ($|Y|$), FREQBIN-baseline absolute difference in performance (Δ)–improvements are in bold, significant improvements are underlined—and coefficient of determination for label-to-frequency regression (R^2).

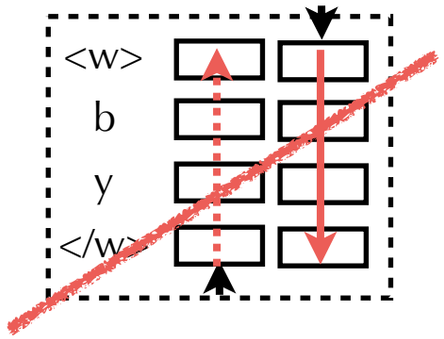
Analysis: Network width

- Hypothesis: Enlargening the hidden layers decreases MTL effectiveness; related to ***reduced capacity*** (Caruana, 1997)
- Tested the net *k*-times wides {k=num aux tasks}
- Result confirms: generalization performance drops



Analysis: Importance of characters

- Hypothesis: Characters are not equally important for all tasks.
- Evaluate bi-LSTM without character Bi-LSTM



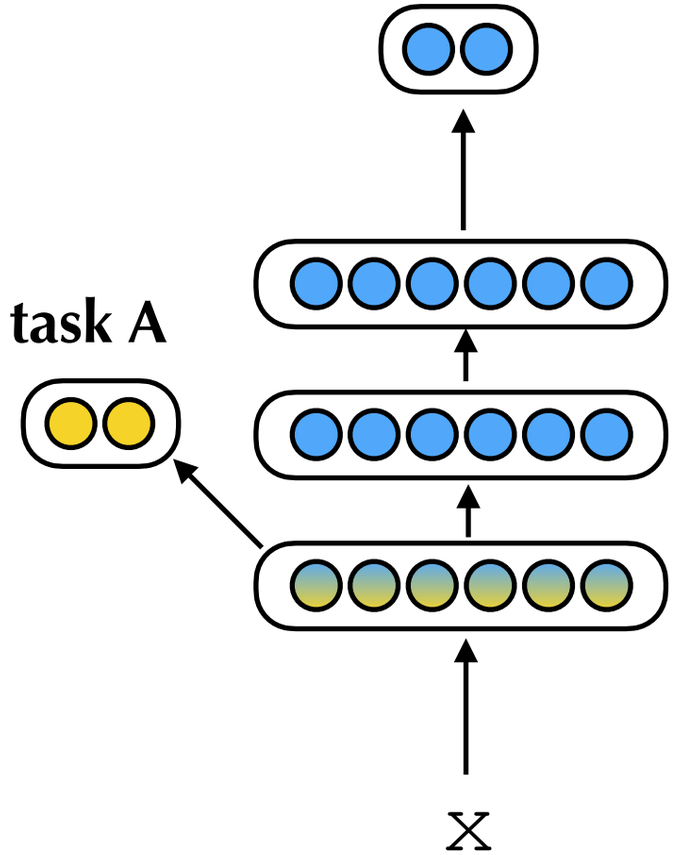
- Results:
  - highest drops for POS and NER (word *form*)
  - less so for semantic tasks, only for FRAMES and MPQA slight improvements (albeit these were tasks with overall lowest performance)

Five main tasks

- FRAMES (FrameNet 1.5) names — *Leadership, Quantity,…*
- NER (CoNLL03) — *Person, Organization..*
- MPQA: Sentiment — *Attitude, Subjective, Objective,…*
- SUPERSENSES (SemCor): *noun.food, verb.emotion,…*
- SEMTRAITS (coarser SemCor): *Animate, Physical, Property,…*

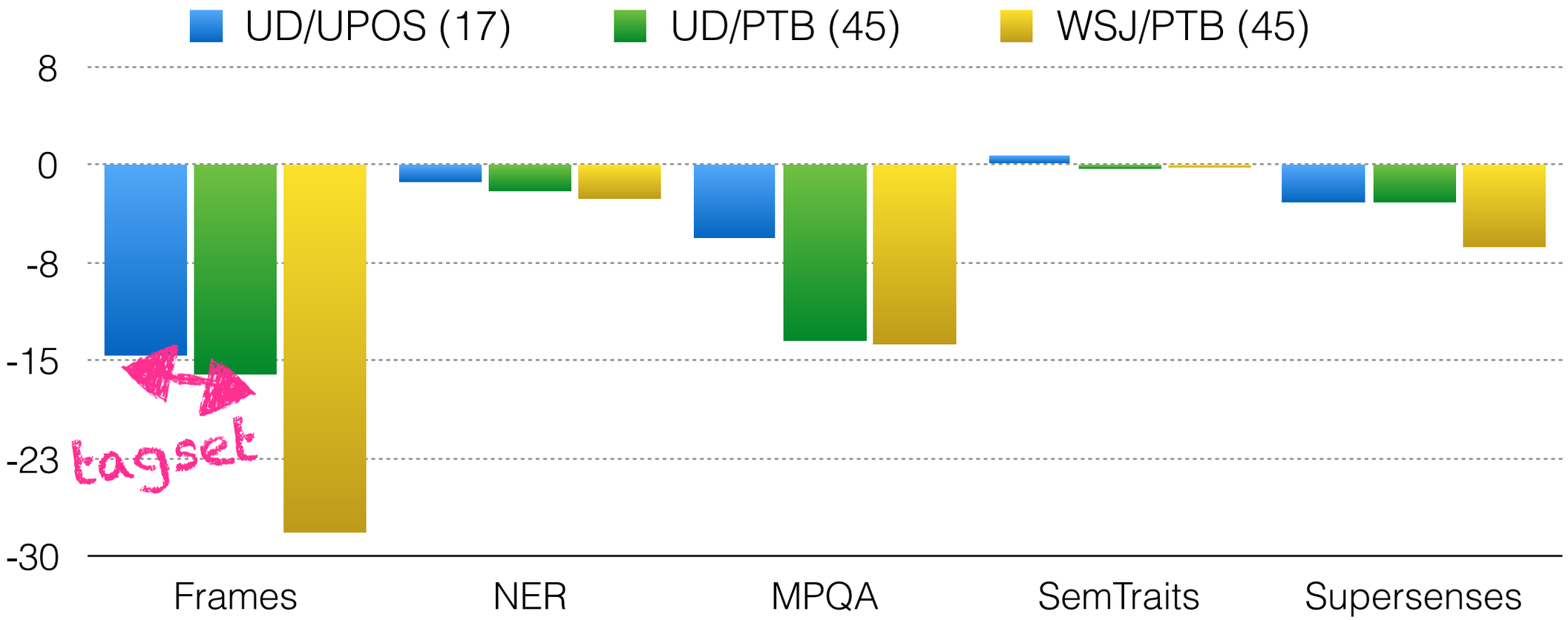
MTL: learning as selective sharing

e.g., Søgaard & Goldberg (2016)



… and many more architectures.

Changing POS corpus



- Preference of architecture to more compact auxiliary task distributions