Taking a bow

Héctor Martínez Alonso

Post-doc at INRIA

ALMAnaCH (previously Alpage): natural language processing and digital humanities

Project VerDi: identification of omission in newswire

hector.martinez-alonso@inria.fr

When is multitask learning effective?

Semantic sequence prediction under varying data conditions

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

When is multitask learning effective?

Semantic sequence prediction under varying data conditions

(A negative-ish result)

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

But first, raised hands

- What is multitask learning?
- What is a neural network?
- What is sequence prediction?

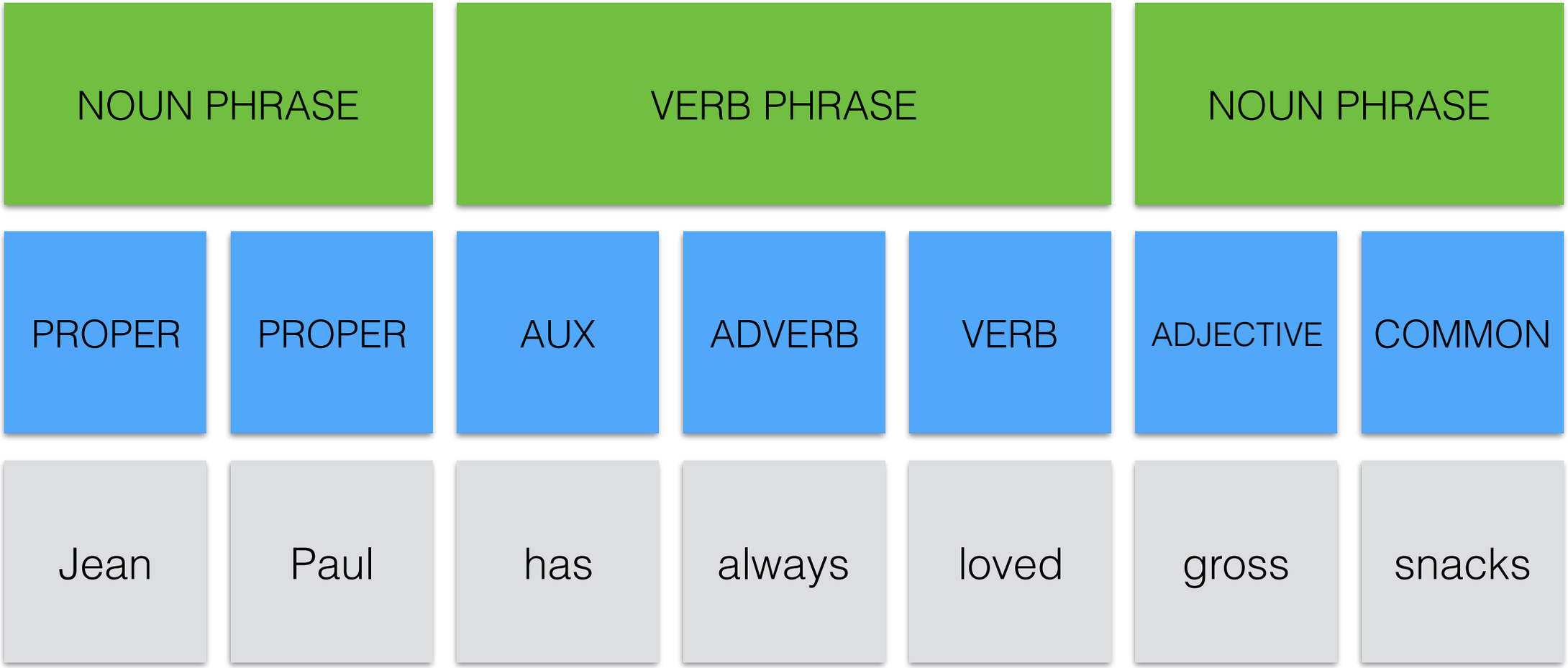
Linguistic sequences

Jean	Paul	has	always	loved	gross	snacks

Linguistic sequences

PROPER	PROPER	AUX	ADVERB	VERB	ADJECTIVE	COMMON
Jean	Paul	has	always	loved	gross	snacks

Linguistic sequences



Linguistic sequences

PERSON		TIME	EMOTION	PHYSICAL	FOOD
NOUN PHRASE		VERB PHRASE		NOUN PHRASE	
PROPER	PROPER	AUX	ADVERB	VERB	ADJECTIVE	COMMON
Jean	Paul	has	always	loved	gross	snacks

Linguistic sequences

—	—	—	—	POSITIVE	NEGATIVE	—
PERSON		TIME	EMOTION	PHYSICAL	FOOD	
NOUN PHRASE		VERB PHRASE		NOUN PHRASE		
PROPER	PROPER	AUX	ADVERB	VERB	ADJECTIVE	COMMON
Jean	Paul	has	always	loved	gross	snacks

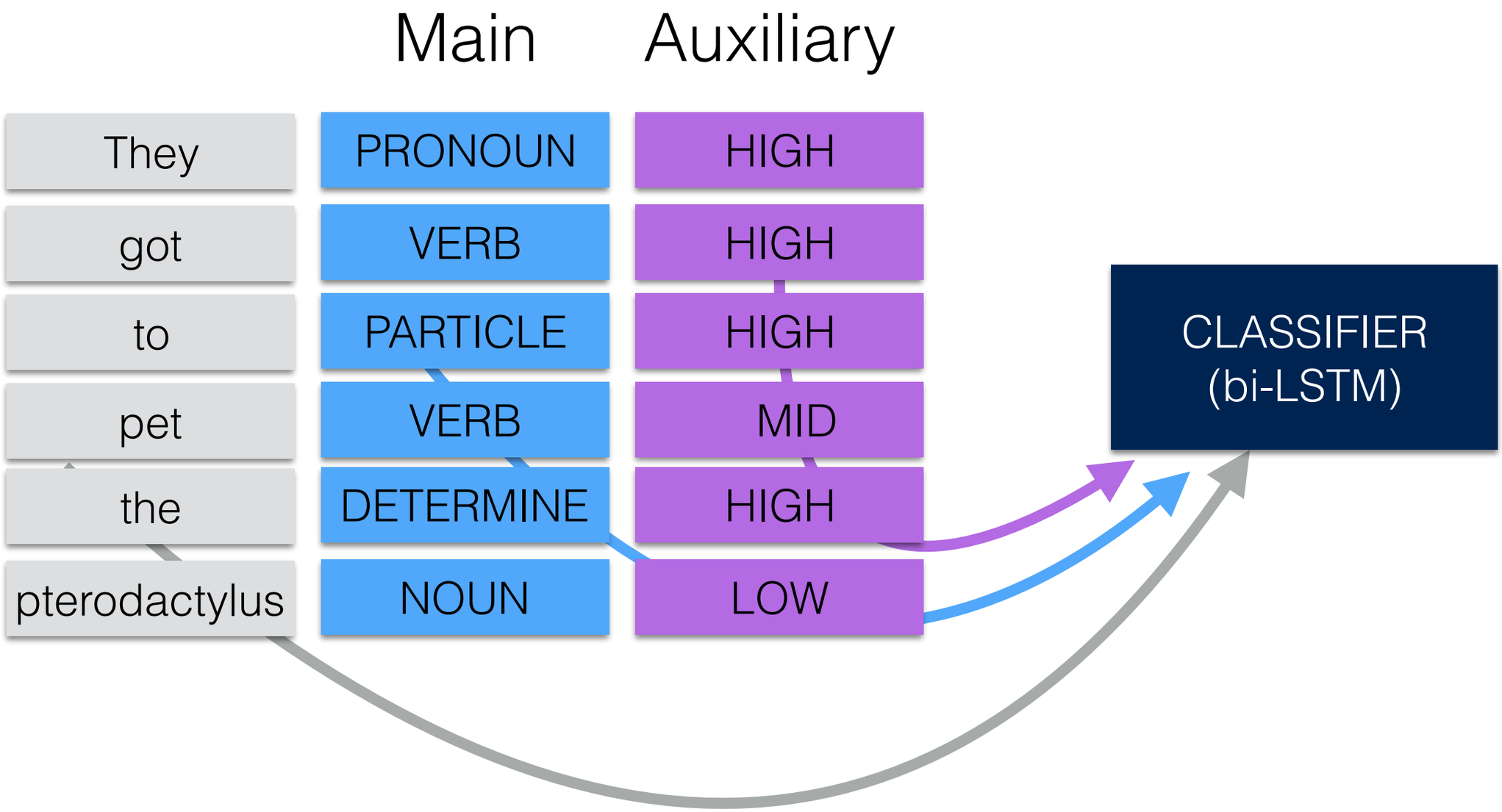
Linguistic sequences

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

Linguistic sequences

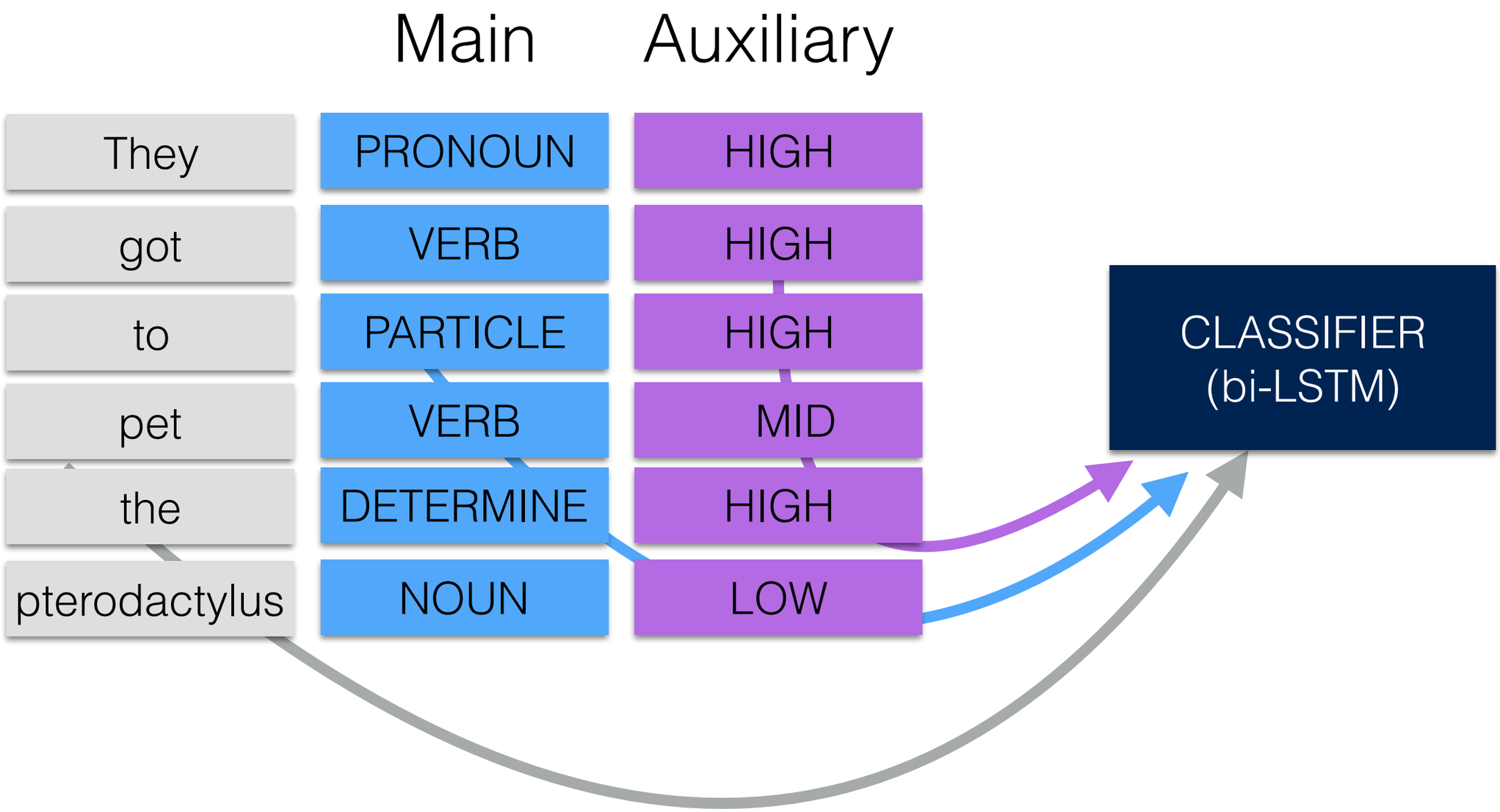
	…			…	
PART OF SPEECH	AUX	ADVERB	VERB	ADJECTIVE	COMMON
	has	always	loved	gross	snacks
WORD FREQUENCY	HIGH	HIGH	MID	LOW	MID

Multitask learning

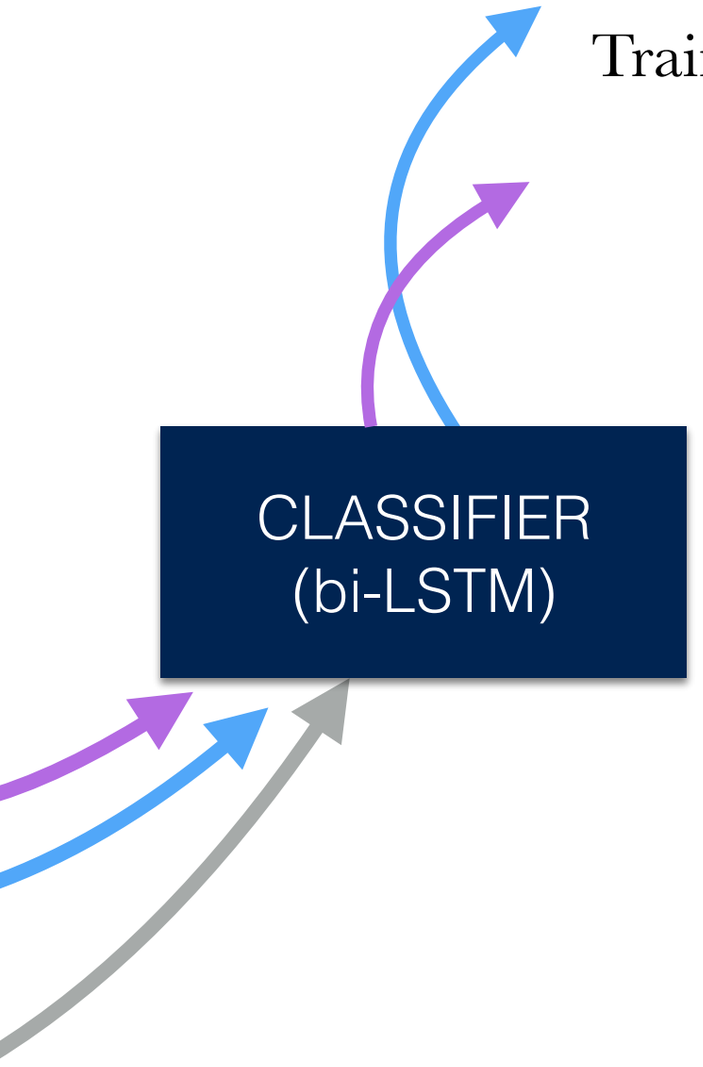


Multitask learning

Original Main(POS)+Aux(Freq) of Plank et al (2016)



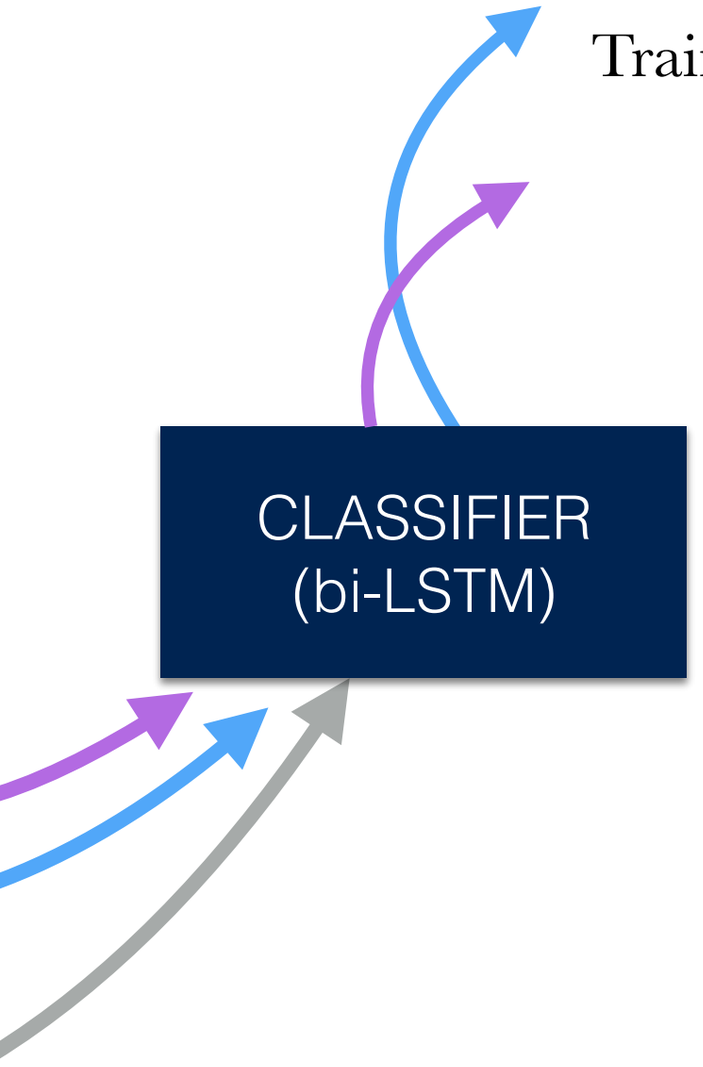
Multitask learning



Training a system with a good main-auxiliary task pair, we can improve the performance of the main task

Provided the auxiliary task is informative

Multitask learning



Training a system with a good main-auxiliary task pair, we can improve the performance of the main task

Provided the auxiliary task is informative

Using a neural network instead of another classification method, we can use heterogeneous data for training instead of requiring a corpus with two parallel annotation layers.

Our work

1. Benchmarking the usage of frequency as auxiliary task

2. Assessing the applicability of MTL to semantic-sequence prediction: supersenses, sentiment, named entities, etc.

3. Establishing information-theoretic criteria for dataset selection, and determining the contribution of different types of data representation

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

- The work in Plank et al (2016) uses the truncated logarithm of a word's frequency to calculate it's frequency label

- We compare this method i.a. with a uniform distribution calculated from the cumulative word frequencies.

2) Semantic sequences

	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 in bold, significant improvements underlined—plus number of systems over baseline for each main task.

2) Semantic sequences

- Frames: Event labels — *Arrival*, *Finish*. Very sparse
- NER: Named entities — *Person*, *Organization*.
- MPQA: Sentiment — *Attitude, Subjective*. Very Sparse.
- Semtraits: *Animate*, *Object*, *Property*
- Supersenses: *noun.food, verb.emotion*

3) 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	**<u>1.12</u>**	.44
SUPERSENSES	83	62.36	-0.69	.47
CHUNK	22	94.76	-0.14	.49
POS	17	94.35	**<u>0.21</u>**	.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).

Conclusions

- We have found that few tasks actually benefit from using frequency, or frequency+POS as an auxiliary task.

- This behavior maps to the co-informativeness of the main and auxiliary label distribution, and in general to distributions with fairly high entropy (and low kurtosis)

- We argue strongly that semantic tasks are harder to predict given immediately observable data properties, such as the skewness of the distribution and the power P(label|word).

Thanks!

Questions?