Johannes Bjerva
University of Groningen, The Netherlands
j.bjerva@rug.nl

Estimating Auxiliary Task Effectivity in Multitask Learning

Multitask Learning (MTL)

- Simultaneously learning several related tasks.
- Common information shared across tasks.
- Proven useful for parsing and POS tagging.

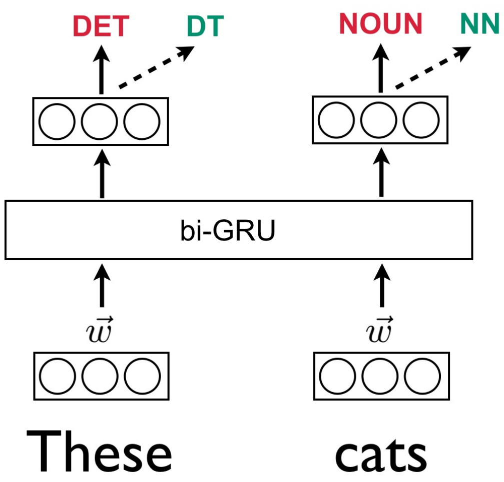
When and why is MTL learning useful in NLP?
We take an information-theoretic perspective.

tokens	These	cats	live	in	that	house	.
UD POS	DET	NOUN	VERB	ADP	DET	NOUN	PUNCT
PTB POS	DT	NNS	VBP	IN	DT	NN	.

tokens	Jim	bought	300	shares	of	Apple	.
UD POS	NOUN	VERB	NUM	NOUN	ADP	NOUN	PUNCT
NE	PER					ORG	

Architecture

- Bi-directional GRU
  - 2 layers, 100d
- Word embeddings
  - no pretraining, 64d
- All parameters shared



Method

- Main task: POS tagging
- Auxiliary task: Dependency Relations (DepRel)

tokens	These	cats	live	in	that	house
UD POS	DET	NOUN	VERB	ADP	DET	NOUN
DepRel	det_R	nsubj_R	root_L_R	case_R	det_R	nmod_L

Category	Directionality	Example	H
Full	Full	nmod:poss/R_L	3.77
Full	Simple	nmod:poss/R	3.35
Simple	Full	nmod/R_L	3.00
Simple	None	nmod	2.03
None	Full	R_L	1.54
None	Simple	R	0.72

Table 1: Granularities of DepRel instantiations

Information Theory

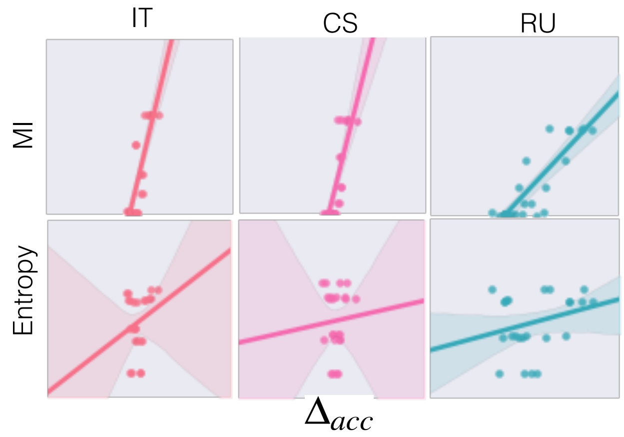
- Entropy (suggested in literature) — H(Y)
- Conditional Entropy — H(Y|X)
- Mutual Information — I(X;Y)

Experiments

- Experiments on 39 languages (most in UD 1.3).
- Varying overlap between Main and AUX task data.
- Comparing Δacc and information-theoretic measures

Results

- Δacc and MI — significant correlation (Table 2)
- Δacc and Entropy — no correlation



Conclusions

- Mutual Information is indicative of auxiliary task effectivity, across a sample of 39 languages.
- Results on semantic tasks in the literature are in line with our findings.

Auxiliary task	$\rho(\Delta_{acc}, H(Y))$	$\rho(\Delta_{acc}, H(Y\|X))$	$\rho(\Delta_{acc}, I(X;Y))$
Dependency Relations (Identity)	−0.06 (p=0.214)	0.12 (p=0.013)	0.08 (p=0.114)
Dependency Relations (Overlap)	0.07 (p=0.127)	0.27 (p<0.001)	**0.43 (p≪0.001)**
Dependency Relations (Disjoint)	0.08 (p=0.101)	0.25 (p<0.001)	**0.41 (p≪0.001)**

Table 2: Correlation scores and associated p-values, between change in accuracy (Δ_{acc}) and entropy ($H(Y)$), conditional entropy ($H(Y|X)$), and mutual information ($I(X;Y)$), calculated with Spearman's ρ, across all languages and label instantiations. Bold indicates the strongest significant correlations.

The 21st Nordic Conference on Computational Linguistics (NoDaLiDa), Gothenburg, Sweden