One Model to Rule them all

Multitask and Multilingual Modelling for Lexical Analysis

Johannes Bjerva, University of Copenhagen
bjerva@di.ku.dk
http://bjerva.github.io

Outline

- Part I - Multitask Learning
  - Multitask Semantic Tagging
  - Multitask Learning and Information Theory

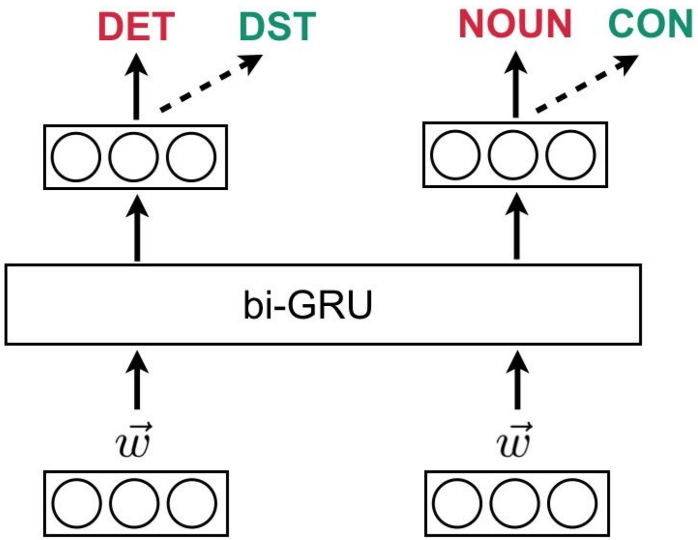
- Part II - Multilingual Approaches
  - Multilingual Effectivity and Language Similarities

- Part III - Combining MTL and Multilinguality
  - Massively Joint Learning

Part I - Multitask Learning

(Neural) Multitask Learning

- Joint learning of several tasks
- Exploiting task relatedness
- Shared parameters (*hard* or *soft* sharing)
- Linguistic auxiliary tasks
    - Language modelling
    - Frequency-based tasks
    - Tagging tasks
    - ...
- Non-linguistic auxiliary tasks
    - Gaze information
    - Typing patterns



Multitask Semantic Tagging with Deep Residual Networks

RQ1. Are semantic tags informative for NLP tasks other than semantic parsing?

RQ2. Are Deep Residual Networks suitable for NLP sequence labelling tasks?

Bjerva, J., Plank, B., Bos, J. (2016). Semantic Tagging with Deep Residual Networks. In COLING.

Deep Residual Networks (ResNets)

- Facilitates error propagation
  - Deeper networks
  - Easier training
- Ensembles of shallower networks (Veit et al., 2016)
- Some usage in NLP
  - Text classification (Conneau et al., 2016)
  - Morphological re-inflection (Östling, 2016)
  - Language identification (Bjerva, 2016)

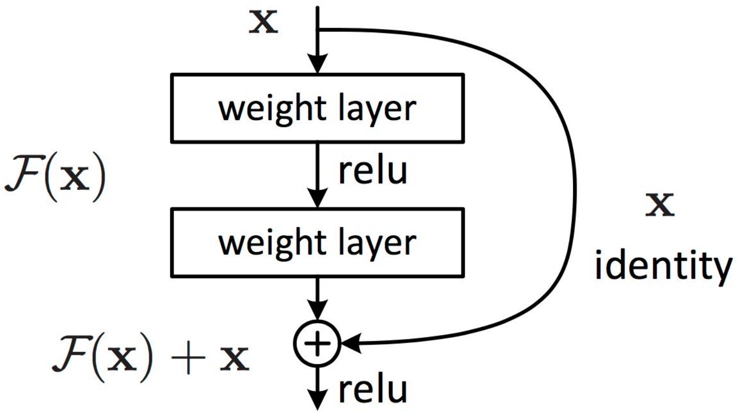
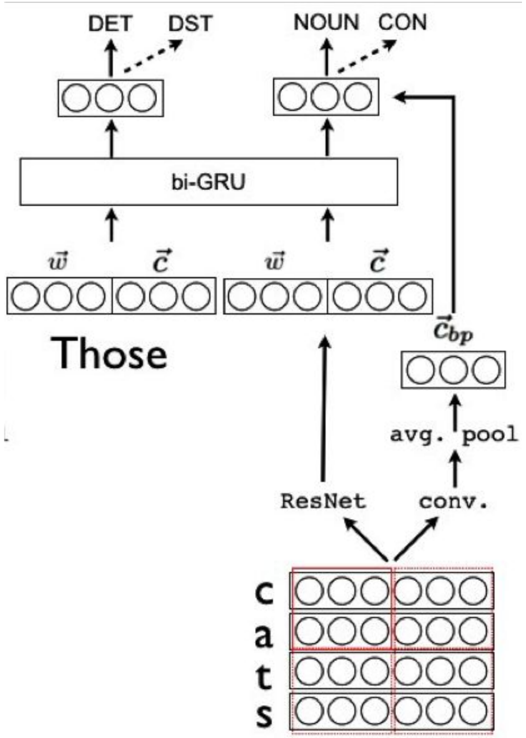


Figure 2. Residual learning: a building block.

He et al. (2015)

System architecture

- Bidirectional Gated Recurrent Units
  - ResNet for character-based word representations **(RQ2)**
  - Pre-trained word representations
- Semtags as auxiliary task for Part-of-Speech tagging **(RQ1)**
- Coarse-grained semtags as auxiliary task for semantic tagging



Results

	Baselines				Basic CNN			ResNet		
	MFC	TNT	Bi-LSTM	Bi-GRU	$\vec{c}$	$\vec{c} \wedge \vec{w}$	+AUX	$\vec{c}$	$\vec{c} \wedge \vec{w}$	+AUX
Semantic tagging	84.64	92.09	94.98	94.26	91.39	94.63	94.53	94.39	**95.14**	94.23
PoS tagging	85.07	92.69	95.04	94.32	77.51	94.89	95.34	92.63	94.88	**95.67**

Bi-LSTM: Plank et al. (2016)

Results

	Baselines				Basic CNN			ResNet		
	MFC	TNT	Bi-LSTM	Bi-GRU	$\vec{c}$	$\vec{c} \wedge \vec{w}$	+AUX	$\vec{c}$	$\vec{c} \wedge \vec{w}$	+AUX
Semantic tagging	84.64	92.09	94.98	94.26	91.39	94.63	94.53	94.39	**95.14**	94.23
PoS tagging	85.07	92.69	$\boxed{95.04}$	94.32	77.51	94.89	95.34	92.63	$\boxed{94.88}$	$\boxed{\mathbf{95.67}}$

Bi-LSTM: Plank et al. (2016)

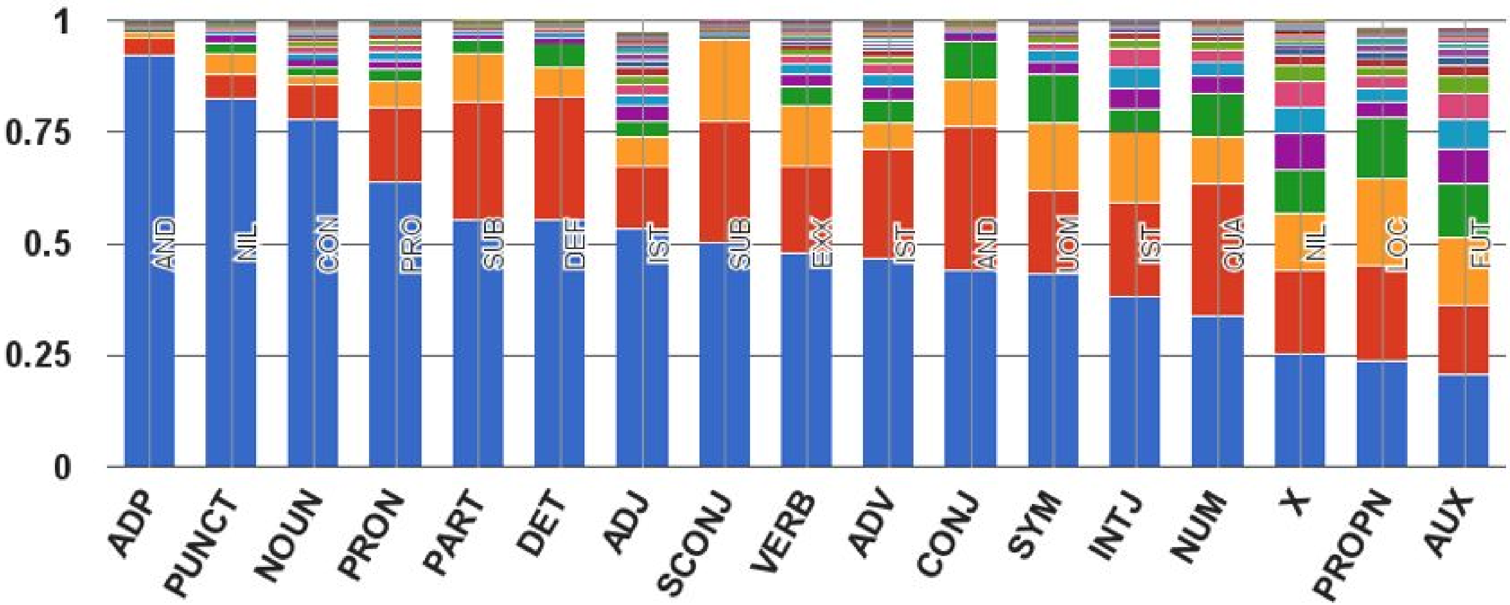
Conclusions and Questions

RQ1. Are semantic tags informative for NLP tasks other than semantic parsing?

RQ2. Are Deep Residual Networks suitable for NLP sequence labelling tasks?

- *Why are the semantic tags useful for PoS?*
- *Why are coarse-grained semantic tags not useful for semtagging?*

Correlations (Semtag and PoS)



Why does MTL work (in NLP sequence labelling)?

- Auxiliary task ***label*** distributions
  (Martínez Alonso and Plank, 2017)
  - Not sufficiently explanatory
    (Bjerva, 2017)
  - Possibly a side-effect
    (Bingel and Søgaard, 2017)

The	*quick*	*brown*	*fox*	*jumps*	*over*	*the*	*lazy*	*dog*	.
DET	ADJ	ADJ	NOUN	VERB	ADP	DET	ADJ	NOUN	PUNCT

The	*quick*	*brown*	*fox*	*jumps*	*over*	*the*	*lazy*	*dog*	.
ADJ	ADJ	DET	DET	NOUN	NOUN	ADJ	ADP	VERB	PUNCT

- Escaping local minima
  (Bingel and Søgaard, 2017)
  - Target task plateau -> Auxiliary task to the rescue

- Regularisation, implicit dataset augmentation,
  globally useful representations
  (cf. Ruder, 2017)

Information-theoretic Perspectives on Multitask Learning

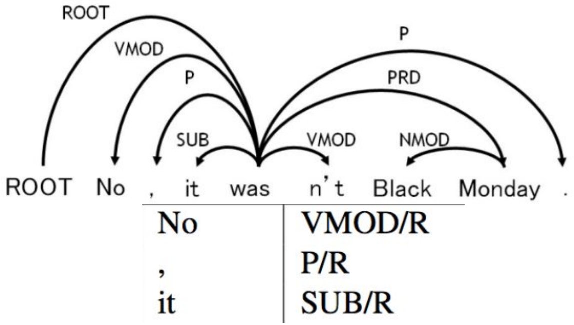
RQ3. Which information-theoretic measures correlate with auxiliary task effectivity?

RQ4. To what extent do such correlations generalise across languages and NLP tasks?

Bjerva, J. (2017). Will my auxiliary tagging task help? Estimating Auxiliary Task Effectivity in Multi-Task Learning. In NoDaLiDa. Best short paper.

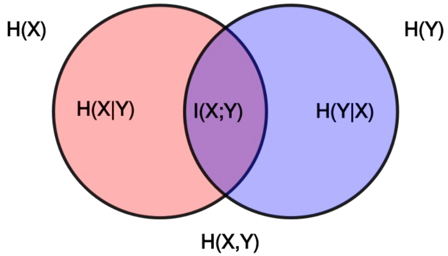
Tasks, Data, and Information-theoretic Measures

- Various semantic tasks (English only)
  - Semantic tagging (Bjerva et al., 2016)
  - Tasks from Martínez Alonso and Plank, 2017
- Universal Dependencies 1.3 (39 languages)
  - Part-of-Speech tagging
  - Dependency relation tagging
    ('Supertagging', Ouchi et al., 2014)



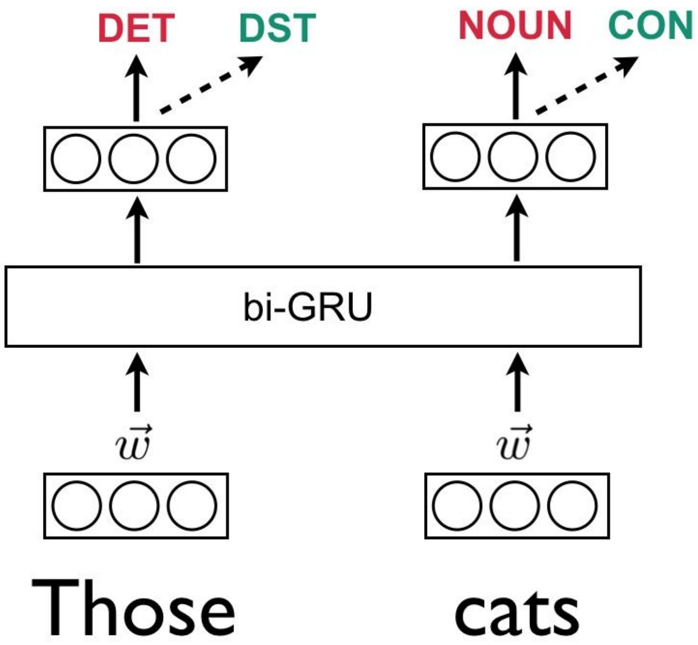
- Main task tagset = X
- Auxiliary task tagset = Y

- Entropy, H(X), H(Y)
- Conditional Entropy, H(X|Y), H(Y|X)
- Mutual Information, I(X;Y)

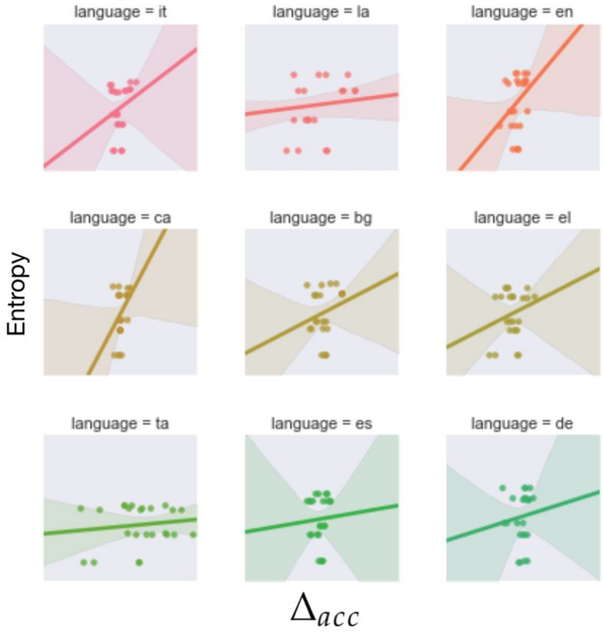


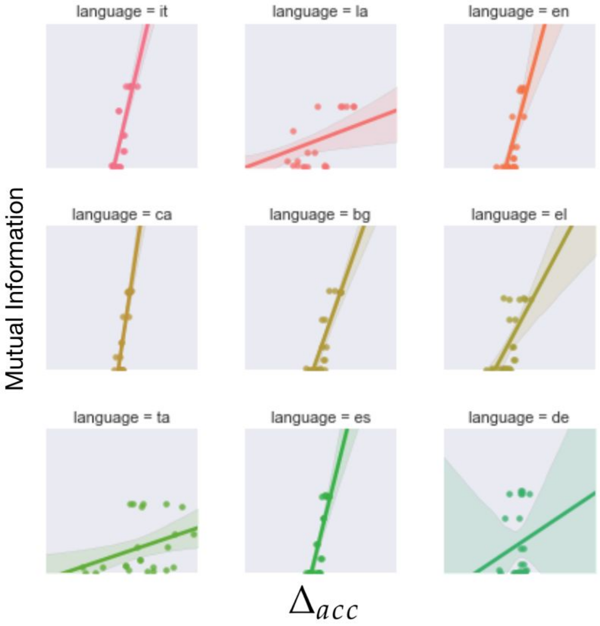
Experiments

- Hard parameter sharing (λ=1)
- Two-layer Bi-GRU (100 units)
- No pre-trained embeddings (64 units)
- No hyperparameter optimisation
- 10 runs per experiment (replicability)
- Eight Dependency Relation granularities
- Three data overlap settings
  - Identical data
  - Some overlap
  - No overlap



PoS Tagging and Dependency Relations





Multilingual Results

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

Multilingual Results

Group	Language	$\rho(\Delta_{acc}, H(Y))$	$\rho(\Delta_{acc}, H(Y\|X))$	$\rho(\Delta_{acc}, I(X;Y))$
Germanic	Danish	0.27 (p=0.116)	0.42 (p=0.011)	**0.78 (p≪0.001)**
	Dutch	0.31 (p=0.070)	0.16 (p=0.337)	**0.55 (p<0.001)**
	English	0.30 (p=0.076)	0.19 (p=0.280)	**0.58 (p<0.001)**
	German	0.03 (p=0.849)	0.13 (p=0.448)	0.18 (p=0.293)
	Norwegian	-0.03 (p=0.858)	0.23 (p=0.183)	0.23 (p=0.177)
	Swedish	-0.03 (p=0.843)	0.29 (p=0.091)	0.31 (p=0.068)
Romance	Catalan	0.34 (p=0.042)	0.33 (p=0.047)	**0.72 (p≪0.001)**
	French	0.06 (p=0.734)	0.38 (p=0.023)	0.48 (p=0.003)
	Galician	0.10 (p=0.574)	0.18 (p=0.304)	0.28 (p=0.099)
	Italian	0.12 (p=0.503)	0.52 (p=0.001)	**0.67 (p≪0.001)**
	Portuguese	-0.02 (p=0.921)	0.61 (p<0.001)	**0.66 (p<0.001)**
	Romanian	-0.31 (p=0.067)	0.34 (p=0.040)	0.04 (p=0.825)
	Spanish	0.02 (p=0.890)	0.60 (p<0.001)	**0.70 (p≪0.001)**
Slavic	Bulgarian	0.20 (p=0.242)	0.50 (p=0.002)	**0.76 (p≪0.001)**
	Croatian	-0.24 (p=0.159)	0.43 (p=0.009)	0.22 (p=0.189)
	Czech	-0.15 (p=0.376)	0.49 (p=0.002)	0.39 (p=0.017)
	O.C. Slavonic	-0.08 (p=0.634)	0.34 (p=0.044)	0.35 (p=0.038)
	Polish	0.13 (p=0.437)	0.40 (p=0.015)	**0.59 (p<0.001)**
	Russian	0.29 (p=0.086)	0.40 (p=0.015)	**0.81 (p≪0.001)**
	Slovene	-0.24 (p=0.156)	0.41 (p=0.014)	0.19 (p=0.259)
Turkic	Kazakh	0.23 (p=0.172)	0.04 (p=0.817)	0.36 (p=0.030)
	Turkish	0.50 (p=0.002)	-0.17 (p=0.317)	0.43 (p=0.008)
Uralic	Estonian	0.45 (p=0.006)	-0.14 (p=0.430)	0.39 (p=0.017)
	Finnish	0.02 (p=0.924)	0.37 (p=0.025)	**0.50 (p=0.002)**
	Hungarian	0.14 (p=0.413)	0.09 (p=0.594)	0.27 (p=0.116)
Other	Arabic	-0.16 (p=0.362)	0.53 (p<0.001)	0.47 (p=0.004)
	Basque	0.41 (p=0.014)	-0.01 (p=0.952)	**0.49 (p=0.002)**
	Chinese	-0.15 (p=0.399)	0.46 (p=0.005)	0.41 (p=0.012)
	Farsi	0.20 (p=0.244)	0.41 (p=0.012)	**0.75 (p≪0.001)**
	Greek	0.20 (p=0.248)	0.19 (p=0.264)	0.44 (p=0.007)
	Hebrew	0.06 (p=0.724)	0.37 (p=0.028)	**0.52 (p=0.001)**
	Hindi	-0.26 (p=0.121)	0.24 (p=0.161)	0.00 (p=0.979)
	Irish	-0.24 (p=0.150)	0.54 (p<0.001)	0.35 (p=0.034)
	Indonesian	-0.42 (p=0.011)	0.51 (p=0.001)	0.11 (p=0.510)
	Latin	0.19 (p=0.271)	0.16 (p=0.362)	0.47 (p=0.004)
	Latvian	0.64 (p<0.001)	-0.23 (p=0.171)	**0.53 (p<0.001)**
	Tamil	0.16 (p=0.337)	0.12 (p=0.482)	0.31 (p=0.067)

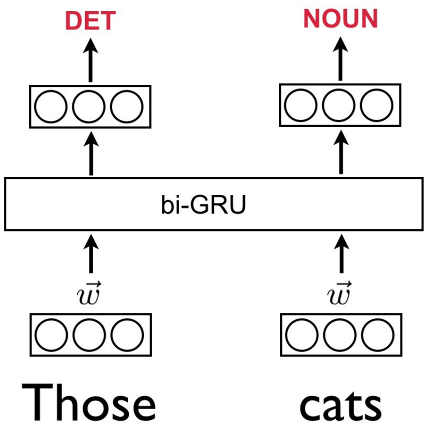
MTL in NLP - A complex situation

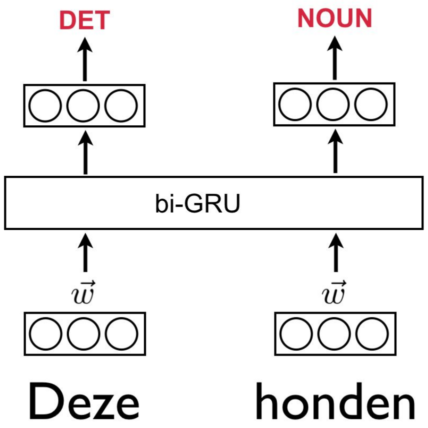
- Correlations between two tasks
  - An auxiliary task will not likely help if it does not correlate with the main task
  - Best auxiliary task is more data for the same task?
- Correlations between tasks and words
  - Multivariate distributions (tasks and words)?
  - Taking sequences into account

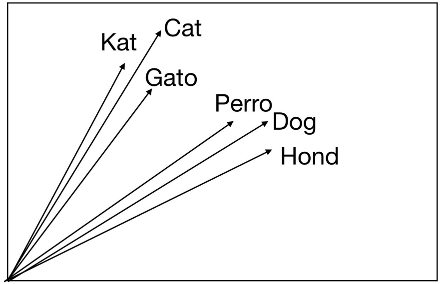
Part II - Multilingual Approaches

(Neural) Multilingual Learning

- Joint learning of several languages
- Exploiting language similarities
    - (Morphological)
    - Lexical
    - Syntactic
- Shared parameters
- Multilingual word embeddings







Language Similarities and Multilingual Learning

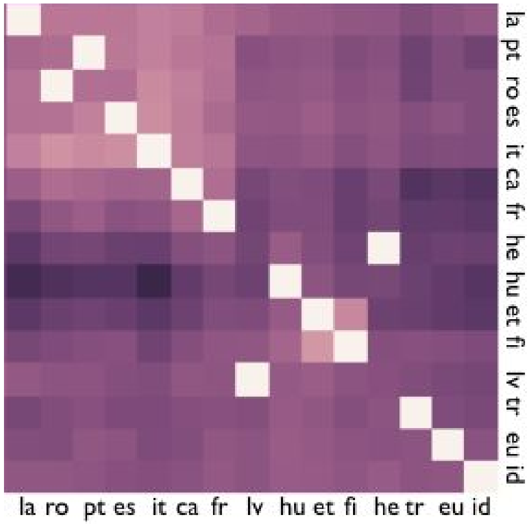
RQ5. What language similarity measures correlate with multilingual effectivity?

RQ6. Do such correlations generalise across language families and NLP tasks?

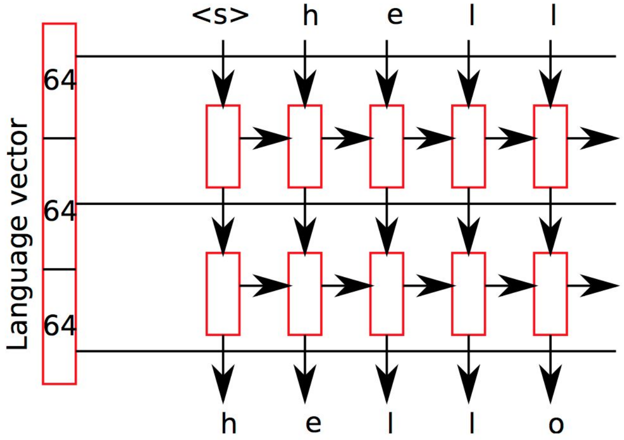
Bjerva, J. (under review). Quantifying the Effects of Multilinguality in NLP Sequence Prediction Tasks.

Language Similarities

Levenshtein distance

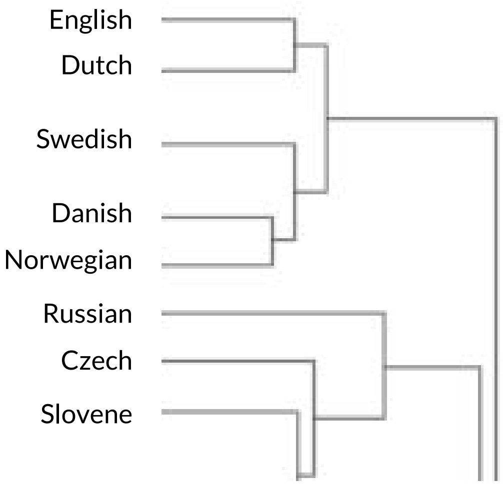


Language vector distance (Östling and Tiedemann, 2017)

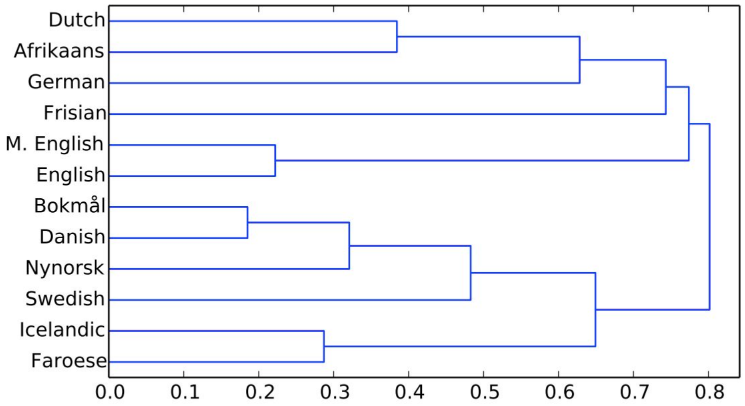


Clustered Language Similarities

Levenshtein distance

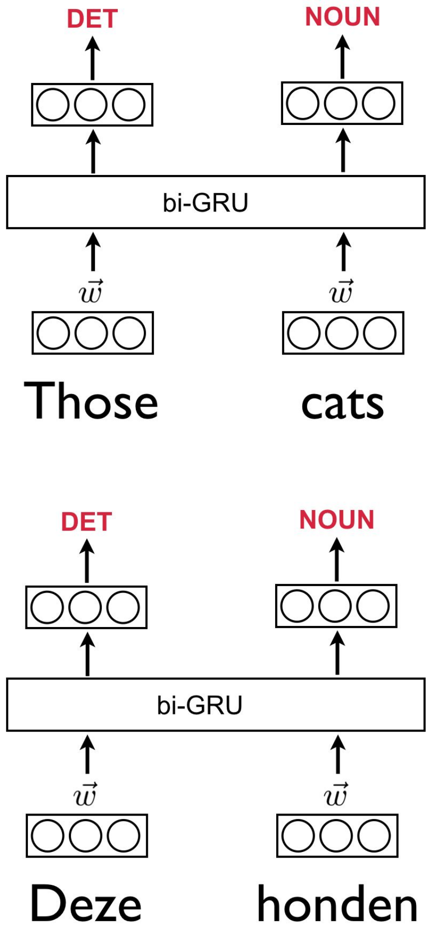


Language vector distance (Östling and Tiedemann, 2017)



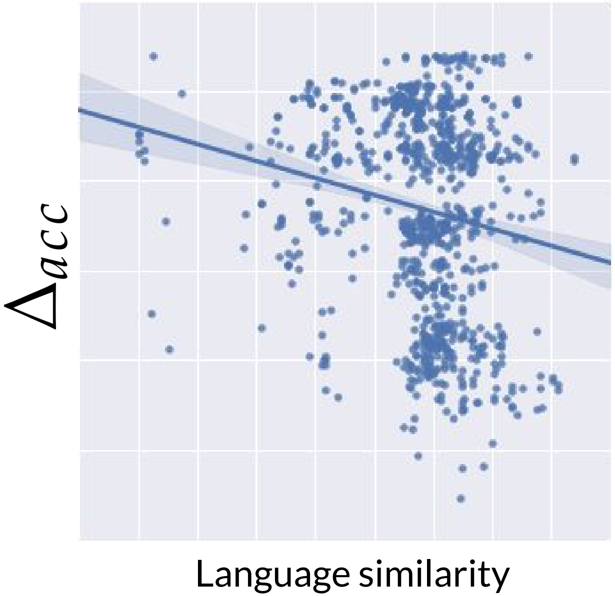
Experiments

- Hard parameter sharing
- Two-layer Bi-GRU (100 units)
- Multilingual embeddings
- 10 runs per experiment (replicability)
- All language pairs in UD 2.0 data

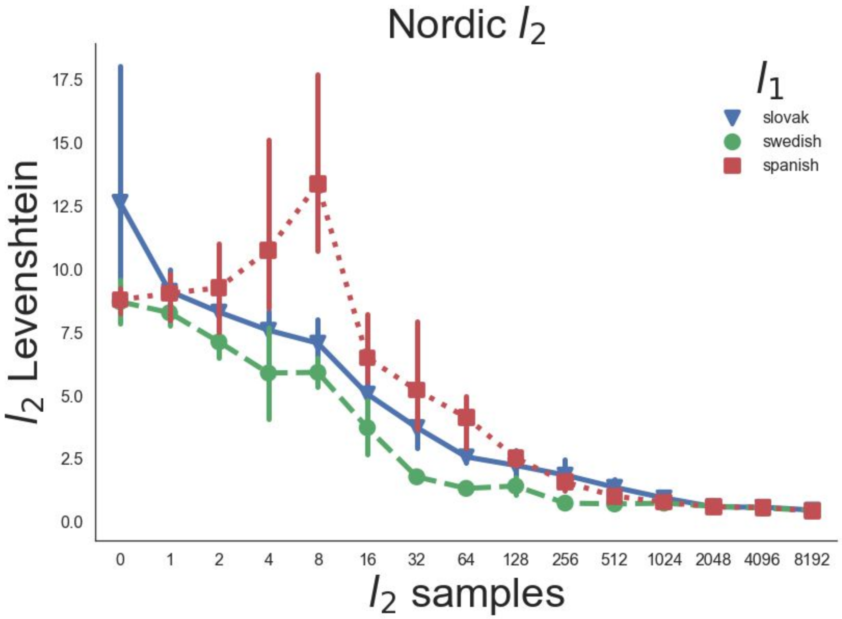


Results

PoS tagging



Morphological inflection



Conclusions

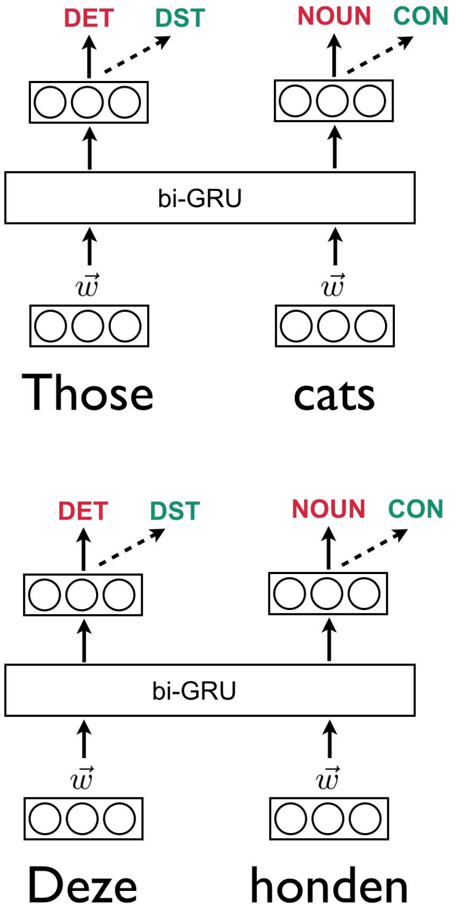
RQ5. What language similarity measures correlate with multilingual effectivity?

RQ6. Do such correlations generalise across language families and NLP tasks?

Part III - Combining MTL and Multilinguality

Joint Multitask and Multilingual Learning

- Joint learning of several languages and tasks
- Exploiting language and task similarities
- Shared parameters
- Multilingual word embeddings



Massively Joint Learning

RQ7. Can MTL and Multilinguality be exploited jointly to improve model transfer?

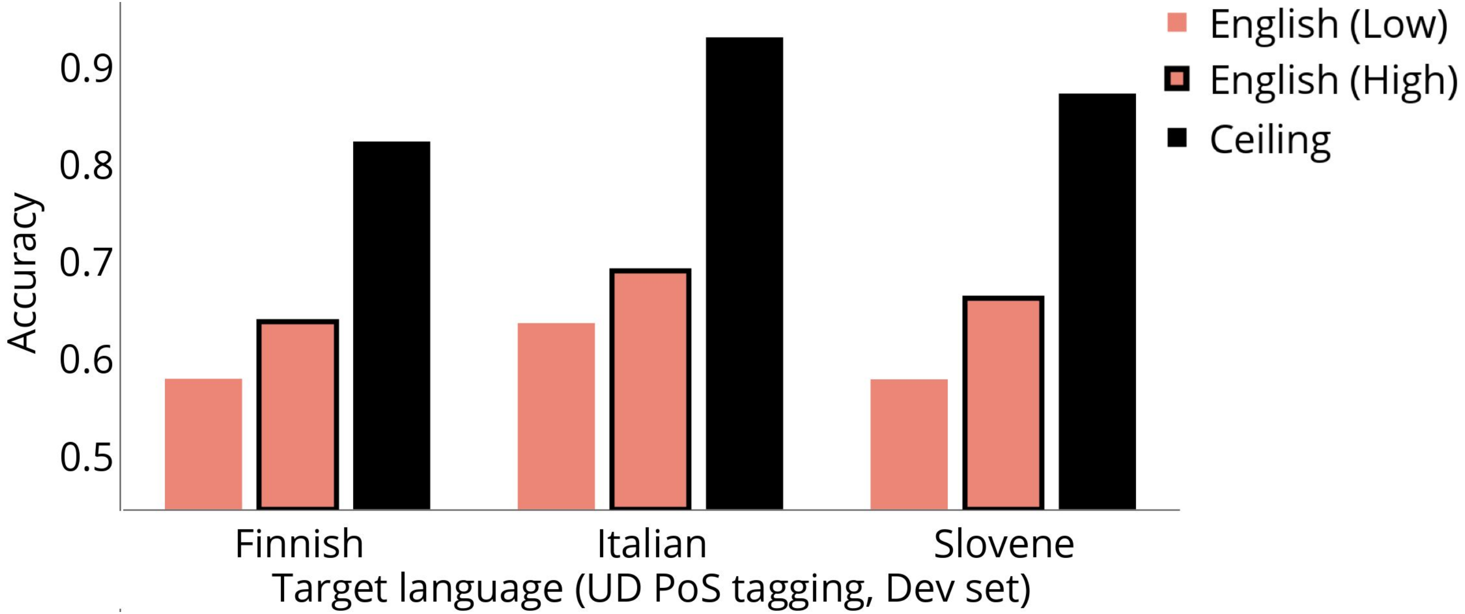
Pilot experiment

- Bi-GRU tagger with multilingual word embeddings (Guo et al., 2016)
    - No hyperparameter optimisation
    - Few training epochs
- Two embedding settings:
    - High resource: trained on Europarl
    - Low resource: trained on Bible texts
- **Main task**: Part-of-Speech tagging
- **Aux. task**: Dependency Relation Tagging
- **Evaluation**: PoS accuracy on target language (Finnish, Italian, Slovene)
- **Ceiling**: Accuracy when training on target language

Model transfer scenarios

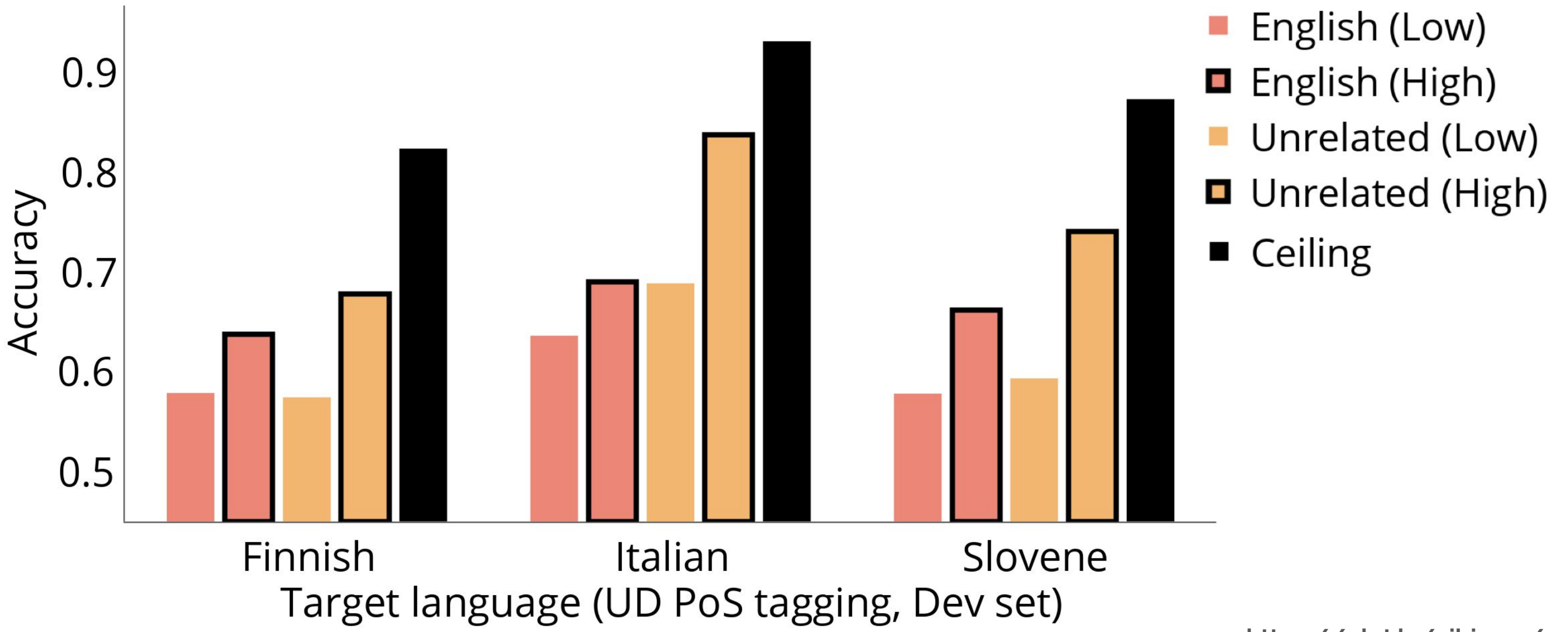
Setting	Source PoS	Source Supertag	Target PoS	Target Supertag
i	English	-	-	-
ii	+ unrelated	-	-	-
iii	+ unrelated	+ unrelated	-	-
iv	+ unrelated	+ unrelated	-	Target language
group-iii	+ related	+ related	-	-
group-iv	+ related	+ related	-	Target language

Source: English PoS



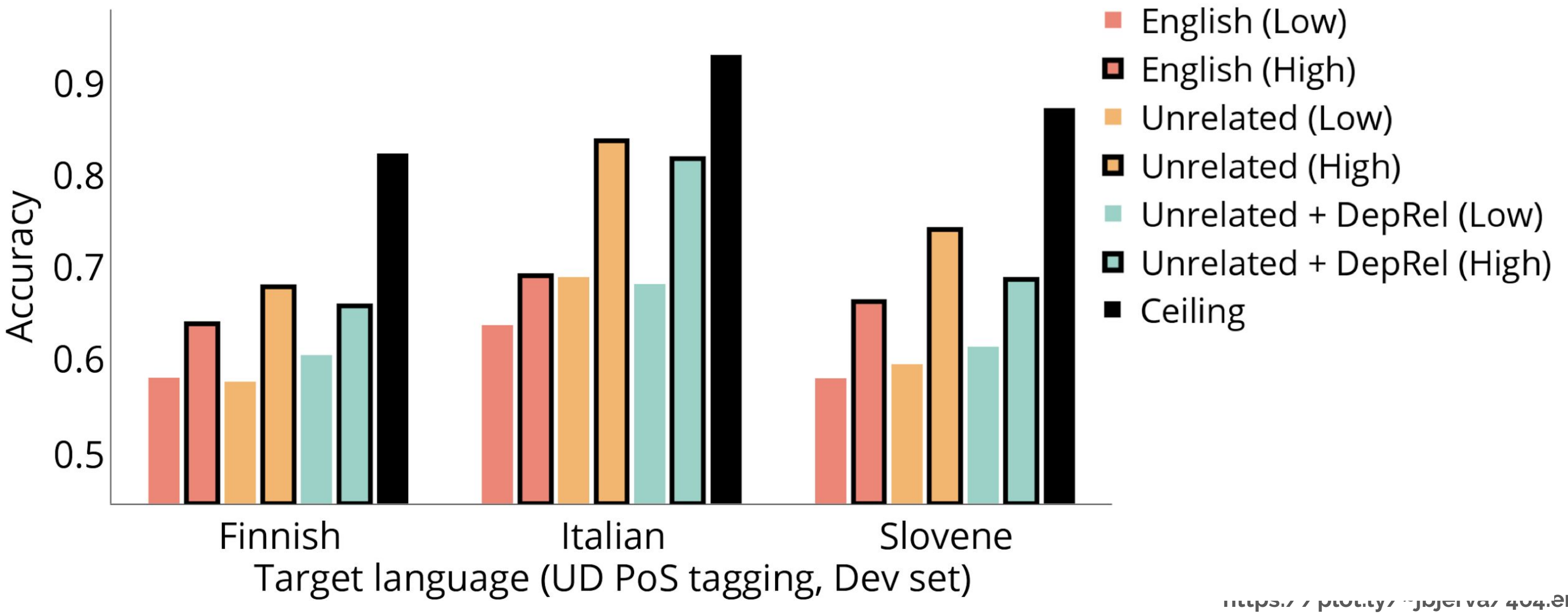
https://plot.ly/~jbjerva/464.embed

Source: English/Unrelated PoS



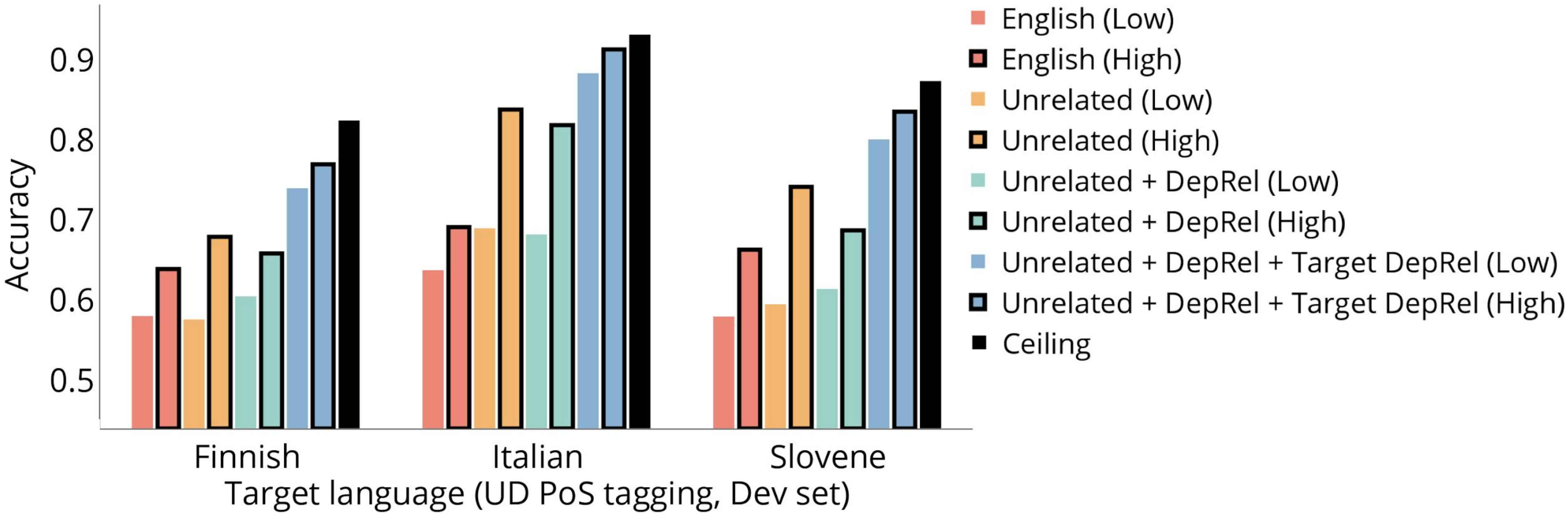
https://plot.ly/~jbjerva/464.embed

Source: English/Unrelated PoS + DepRel



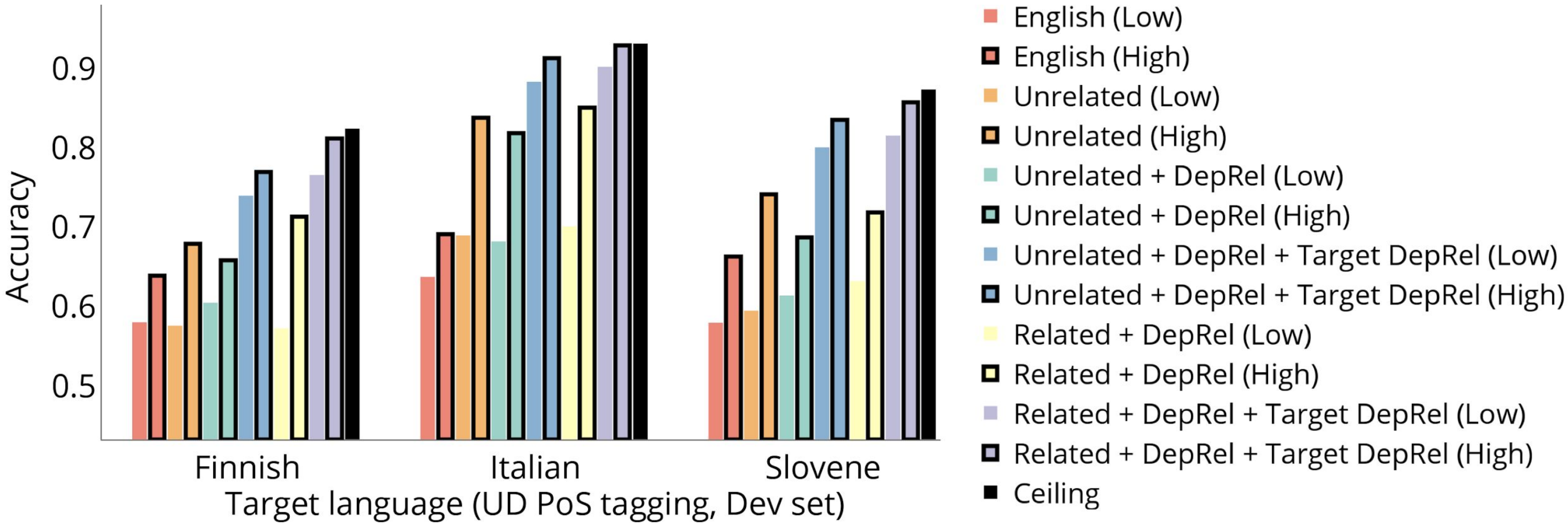
https://plot.ly/~jbjerva/404.embed

Source: English/Unrelated PoS + DepRel + Target DepRel



https://plot.ly/~jbjerva/464.embed

Source: + in-group PoS/DepRel



https://plot.ly/~jbjerva/464.embed

Preliminary results

RQ7. Can MTL and Multilinguality be exploited jointly to improve model transfer?

- Positive preliminary results
- Unrealistic setting

Summary

- Information-theoretic measures taking joint probabilities into account offer some explanatory value for MTL effectivity.
- Measures of language similarity exhibit some correlation with multilingual effectivity.
- Preliminary experiments in multitask multilingual learning show promise.

Future work

- Jointly learning tasks and languages
  - Sluice networks
- Estimating MTL effectivity with multivariate mutual information
  - Taking words / sequences into consideration
  - Taking several tasks into consideration

Semantic Tags for Multilingual Semantic Parsing

- POS tags: insufficient and irrelevant information
- Insufficient:
  - *every* (DT / univ. quant.)
  - *no* (DT / neg.)
  - *some* (DT / exist. quant.)
- Irrelevant:
  - *walks* (VBZ / pres. simpl.)
  - *walk* (VBP / pres. simpl.)

(1.1)

We	*must*	*draw*	*attention*	*to*	*the*	*distribution*	*of*	*this*
PRON	AUX	VERB	NOUN	ADP	DET	NOUN	ADP	DET
form	*in*	*those*	*dialects*	.				
NOUN	ADP	DET	NOUN	PUNCT				

(1.2)

We	*must*	*draw*	*attention*	*to*	*the*	*distribution*	*of*	*this*
PRO	NEC	EXS	CON	REL	DEF	CON	AND	PRX
form	*in*	*those*	*dialects*	.				
CON	REL	DST	CON	NIL				

Parallel Meaning Bank p01/d3421
(Original source: Tatoeba)

Semantic Task Results