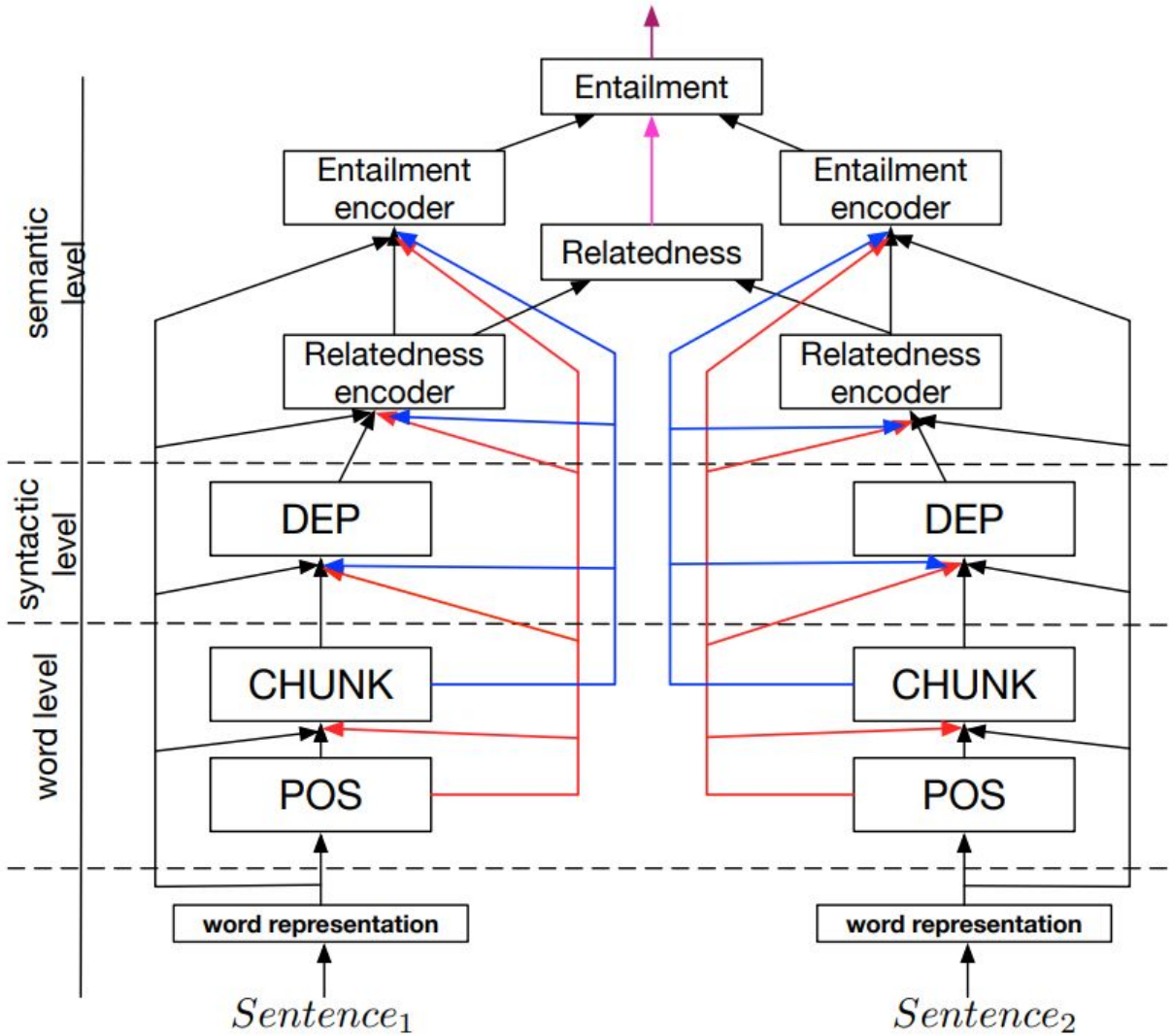
Massive Choice, Ample Tasks (MaChAmp): A Toolkit for Multi-task Learning in NLP

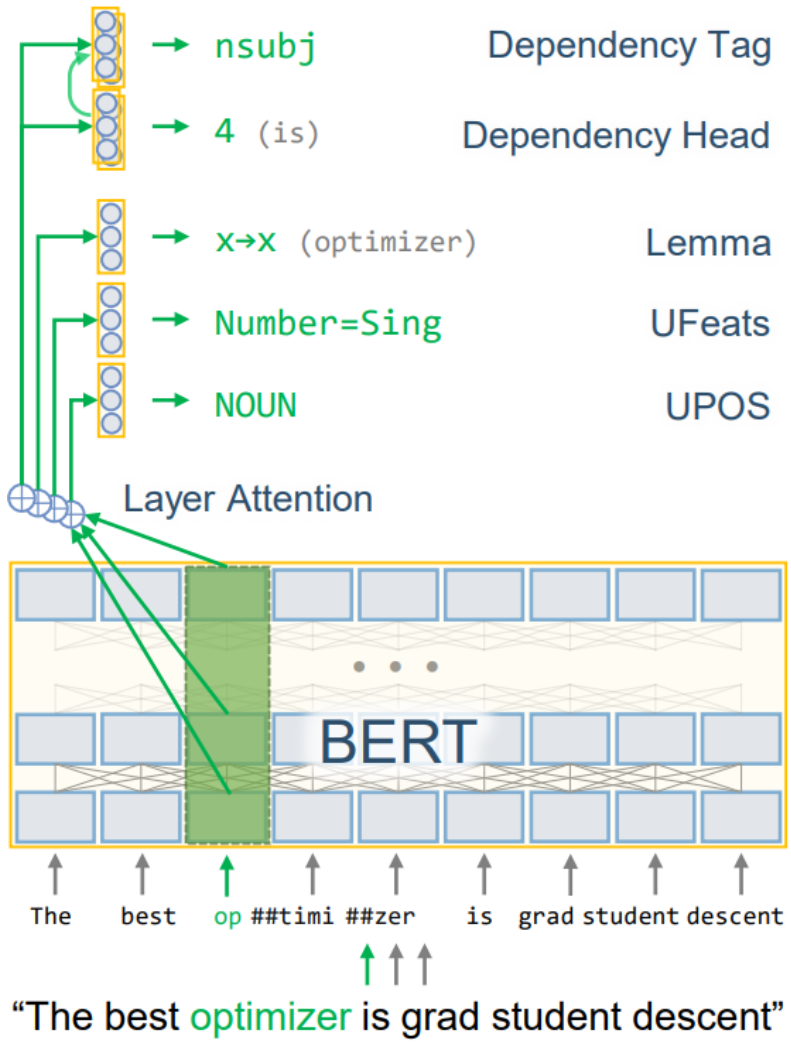
Rob van der Goot * **Ahmet Üstün** * **Alan Ramponi** * **Barbara Plank**
IT University of Copenhagen University of Groningen University of Trento
Fondazione the Microsoft Research - University of Trento COSBI
robv@itu.dk, a.ustun@rug.nl, alan.ramponi@unitn.it, bapl@itu.dk

Multitask Learning



Taken from: A Joint Many-Task Model: Growing a Neural Network for Multiple NLP Tasks. Kazuma Hashimoto, Caiming Xiong, Yoshimasa Tsuruoka, Richard Socher (EMNLP 2017)

Udify

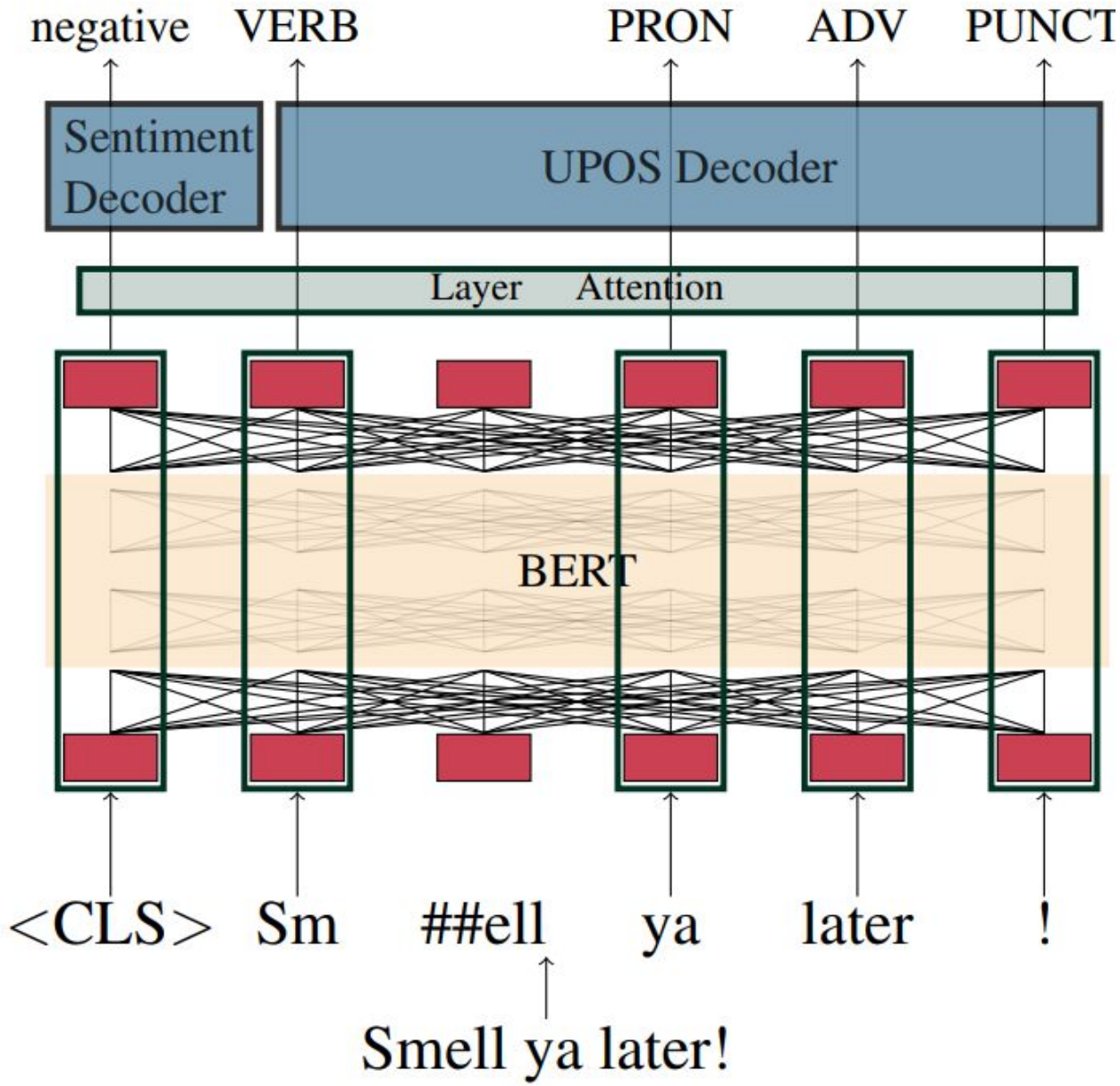


Taken from: 75 Languages, 1 Model: Parsing Universal Dependencies Universally Dan Kondratyuk and Milan Straka (EMNLP 2019)

MaChAmp

One arm alone can move mountains.

MaChAmp


negative
VERB
PRON
ADV
PUNCT
Sentiment Decoder
UPOS Decoder
Layer Attention
BERT
<CLS>
Sm
##ell
ya
later
!
Smell ya later!

Original research questions

- Can it still be beneficial to process tasks sequentially (and embed previous predictions)?
- What would the best order be?

Input to decoder

Normal case:
$\vec{b}$

Input to decoder

Normal case:
$\vec{b}$

Add information from previous task prediction:
$\vec{b} \cdot \vec{l}$

Input to decoder

Normal case:
$\vec{b}$

Add information from previous task prediction:
$\vec{b} \cdot \vec{l}$

Weighted average from previous prediction:
$\vec{b} \cdot \sum_{i=0}^{n} p_i * \vec{l_i}$

Training MaChAmp

```
{
    "UD": {
        "train_data_path": "data/ewt.train",
        "validation_data_path": "data/ewt.dev",
        "word_idx": 1,
        "tasks": {
            "upos": {
                "task_type": "seq",
                "column_idx": 3
            }
        }
    }
}
```

Training MaChAmp

Training can be as easy as:

```
python3 train.py --dataset_config upos.json
```

Training MaChAmp

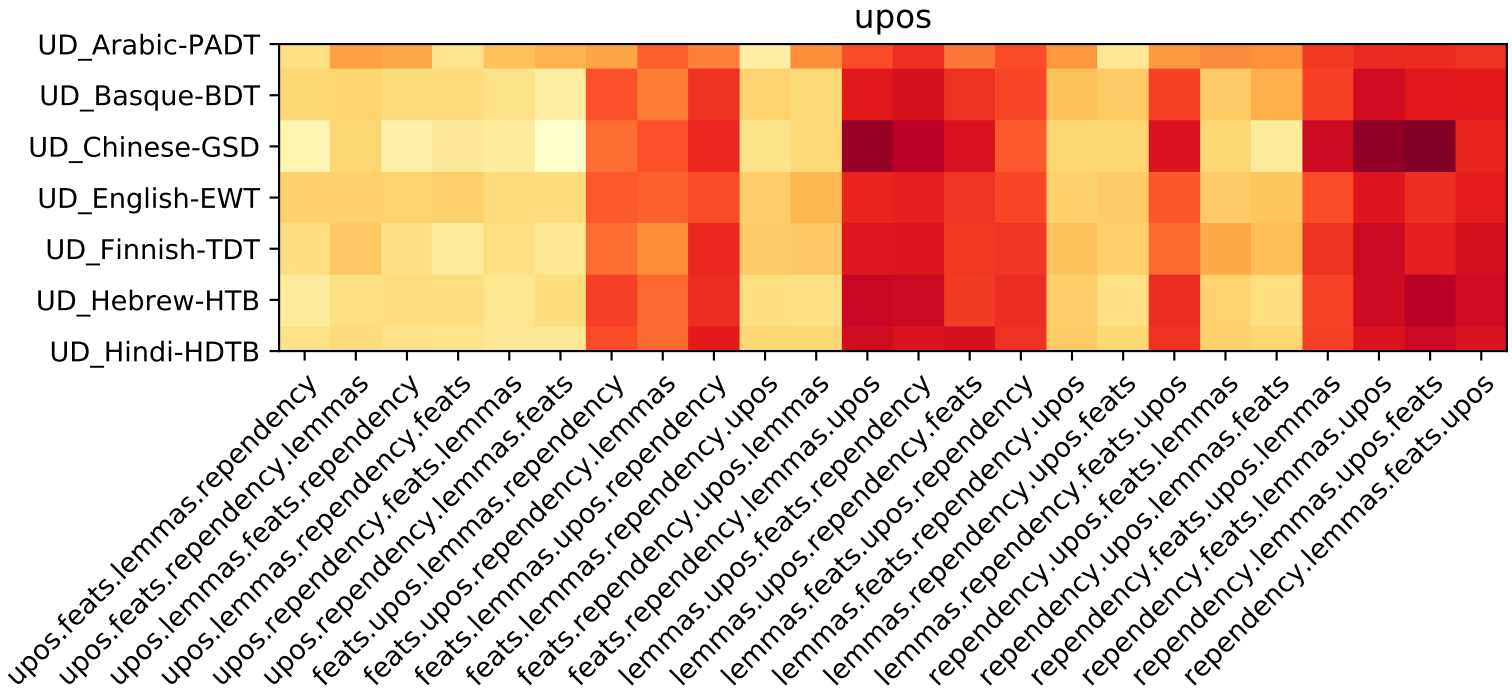
Training can also be done like:

```
python3 train.py --dataset_config upos.json --name \
EWT.upos --device -1 --parameters_config newParams.json
```

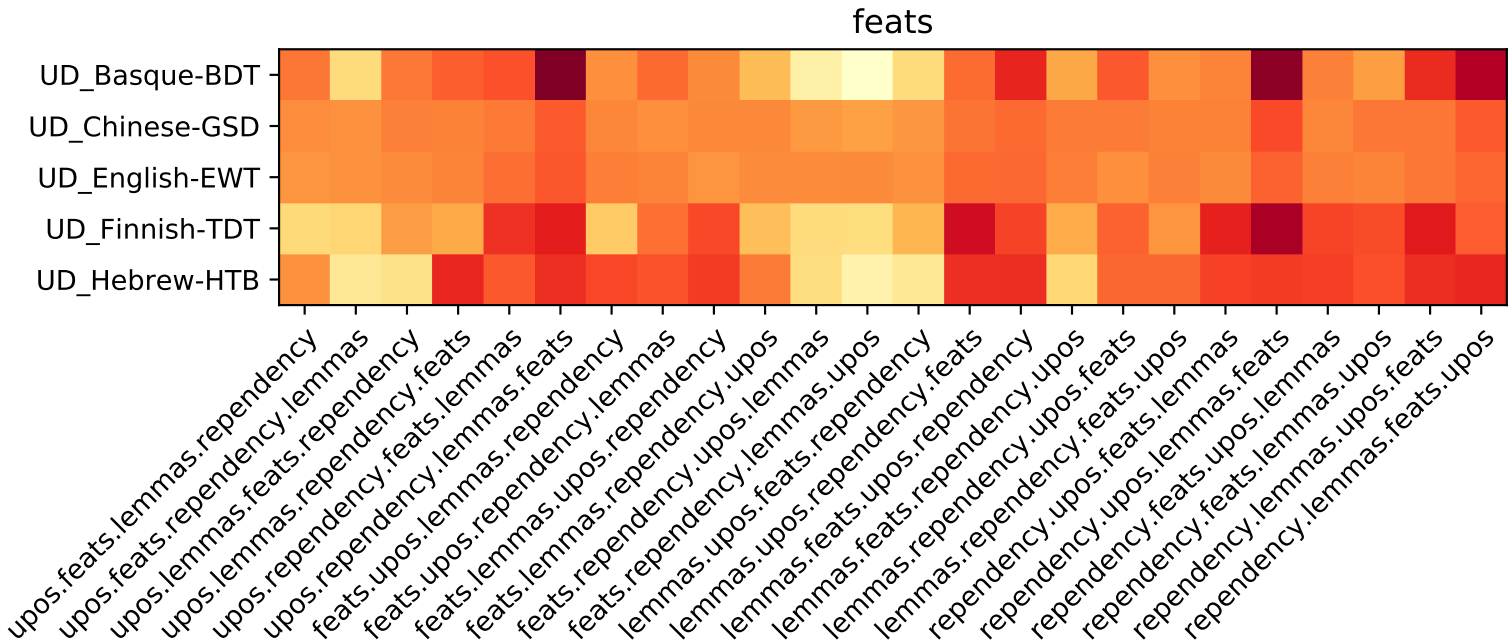
Training MaChAmp

```
{
    "UD": {
        "train_data_path": "data/ewt.train",
        "validation_data_path": "data/ewt.dev",
        "word_idx": 1,
        "tasks": {
            "upos": {
                "task_type": "seq",
                "column_idx": 3
            },
            "xpos": {
                "task_type": "seq",
                "column_idx": 4,
                "prev_task_embed_dim":32,
                "order":2
            }
        }
    }
```

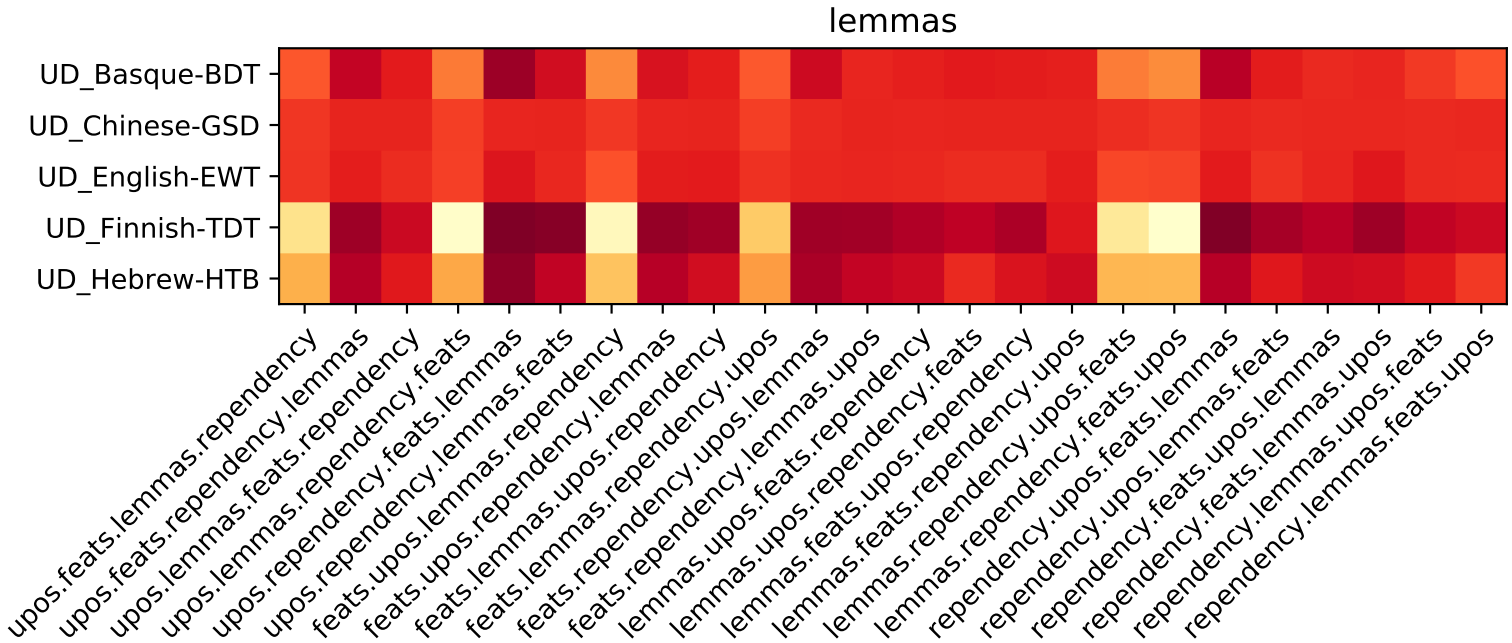
Multiple languages (UPOS)



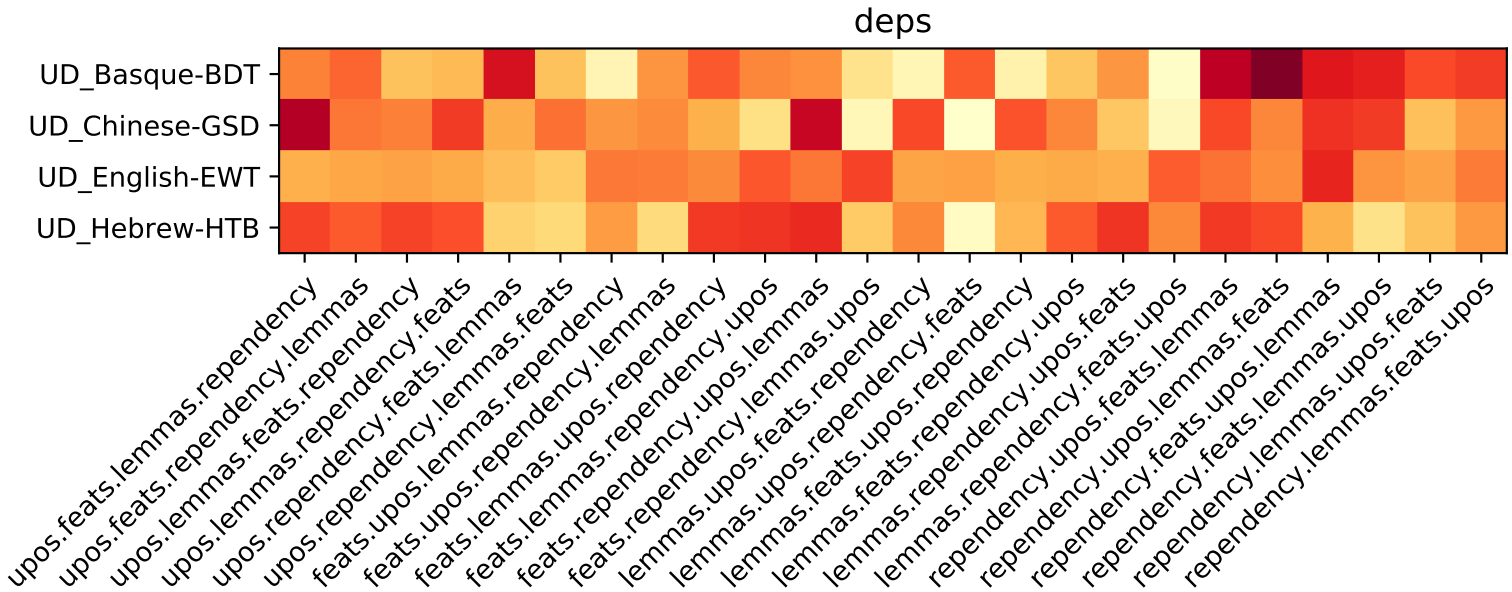
Multiple languages (Morphological tagging)



Multiple languages (Lemmas)



Multiple languages (Dependency parsing)



Comparison to previous work

	JMT_{all}	Random
POS	97.88	97.83
Chunking	97.59	97.71
Dependency UAS	94.51	94.66
Dependency LAS	92.60	92.80
Relatedness	0.236	0.298
Entailment	84.6	83.2

Taken from: A Joint Many-Task Model: Growing a Neural Network for Multiple NLP Tasks. Kazuma Hashimoto, Caiming Xiong, Yoshimasa Tsuruoka, Richard Socher (EMNLP 2017)

Now what?

So, how about using word-level tasks to improve phrase-level tasks?

Now what?

So, how about using word-level tasks to improve phrase-level tasks?

- Add sentence level tasks (and make it possible to use multiple datasets simultaneously)
- Make challenging: cross-lingual settings!

```
"UD": {
    "train_data_path": "data/ewt.train",
    "validation_data_path": "data/ewt.dev",
    "word_idx": 1,
    "tasks": {
        "upos": {
            "task_type": "seq",
            "column_idx": 3
        },
    }
},"RTE" {
    "train_data_path": "data/ewt.train",
    "validation_data_path": "data/ewt.dev",
    "sent_idxs": [0,1],
    "tasks": {
        "rte": {
            "task_type": "classification",
            "column_idx": 2
        },
    }
```

Task Types

- seq
- string2string
- dependency
- multiseq
- masked_crf
- classification

Task Types

- seq
- string2string
- dependency
- multiseq
- masked_crf
- classification

Performance

- EWT
- PMB
- GLUE

	EWT v2.3					PMB v3.0				
Task	dep	feats	lemma	upos	xpos	lemma	semtag	supertag	verbnet	wordnet
Task type	dep	seq	s2s	seq	seq	s2s	seq	seq	seq	s2s
Train size			205k					43k		
MaChAmp$_{(ST)}$	**89.90**	**97.18**	**98.21**	**97.01**	96.64	**97.52**	**98.32**	94.87	94.37	89.15
MaChAmp$_{(MT)}$	89.61	97.15	97.79	**97.01**	**96.79**	97.33	98.23	**94.91**	**94.54**	**89.32**
UDify	89.67	97.15	97.80	96.90	–	–	–	–	–	–

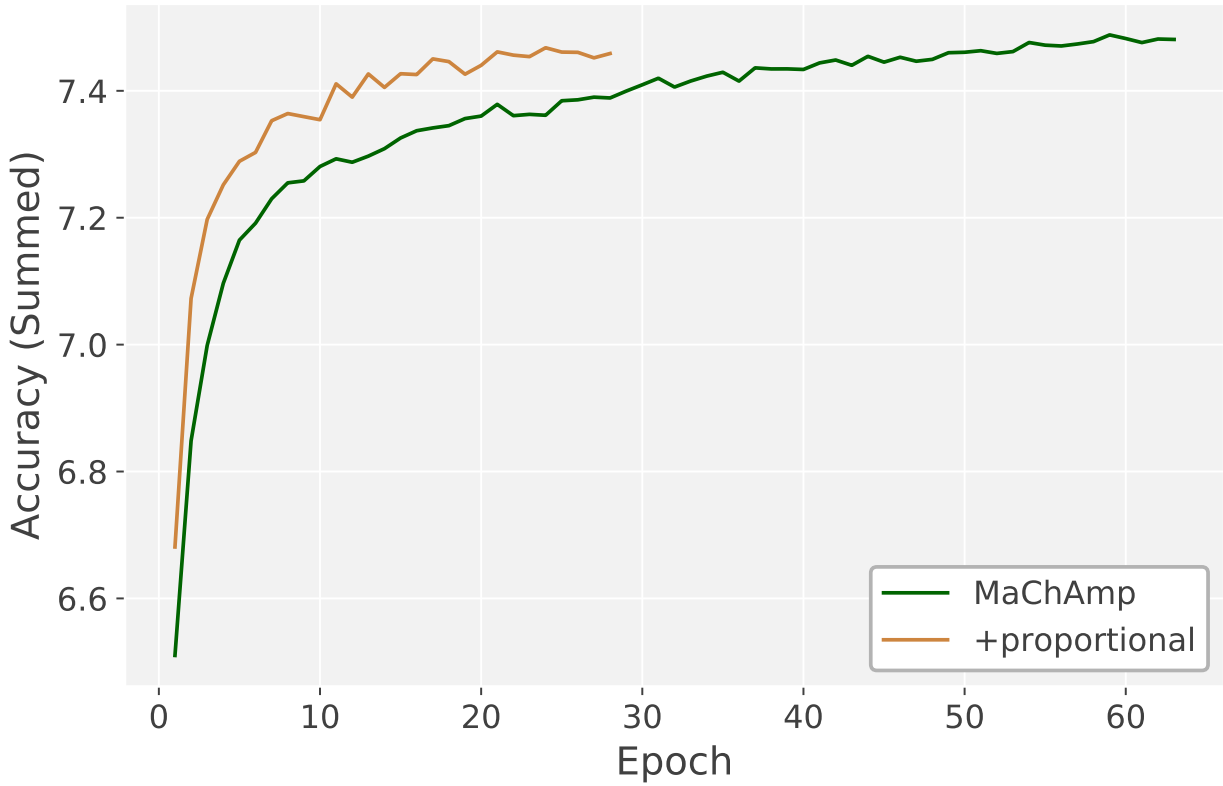
	GLUE								
Task	cola	mnli	mnli-mis	mrpc	qnli	qqp	rte	snli	sst-2
Task type	c	c	c	c	c	c	c	c	c
Train size	8.5k	392k	392k	3.6k	108k	363k	2.5k	549k	67k
MaChAmp$_{(ST)}$	**78.04**	81.99	82.15	**86.03**	**88.31**	**89.75**	72.20	89.58	**90.71**
MaChAmp$_{(MT)}$	72.20	82.35	**82.80**	82.11	86.58	89.27	**73.65**	89.61	90.25
BERT-base	–	**84.4**	**86.7**	–	–	–	–	**93.3**	–

Proportional sampling

- Downscale each task to the smallest task
- Upscale each task to the largest task
- Take the average amount of batches over all tasks for all tasks

Proportional sampling

- Downscale each task to the smallest task
- Upscale each task to the largest task
- Take the average amount of batches over all tasks for all tasks



Zero-shot learning

Preliminary Results

	English	Arabic	Turkish	Urdu	Vietnamese
En-XNLI	81.84	65.93	61.52	58.02	69.08
En-XNLI +En-UD +X-UD	-	63.17 WO: 0.64/0.30	60.56 WO: 0.52/0.19	59.72 WO: 0.76/0.31	67.65 WO: 0.70/0.22
En-XNLI +X-UD (4 Tasks)		64.57 WO: 0.64/0.30	59.39 WO: 0.49/0.17	58.75 WO: 0.76/0.27	70.02 WO: 0.68/0.15

Natural language understanding

set	a	reminder	to	tell	my	wife	i	love	her
O	O	B-Rem./NN	O	B-Rem.	I-R	I-R	I-R	I-R	I-R

Intent: setReminder

Natural language understanding

```
# text: Set alarm every minute for next hour
# intent: alarm/set_alarm
# slots: 10:36:datetime
1       set     alarm/set_alarm NoLabel
2       alarm   alarm/set_alarm NoLabel
3       every   alarm/set_alarm B-datetime
4       minute  alarm/set_alarm I-datetime
5       for     alarm/set_alarm I-datetime
6       next    alarm/set_alarm I-datetime
7       hour    alarm/set_alarm I-datetime
```

Natural language understanding

```
{
    "NLU": {
        "train_data_path": "data/nlu/train-en.conllu",
        "validation_data_path": "data/nlu/eval-en.conllu",
        "word_idx": 1,
        "tasks": {
            "slot": {
                "task_type": "seq",
                "metric": "span-f1",
                "column_idx": 3
            },
            "intent": {
                "task_type": "classification",
                "column_idx": -1
            }
        }
    }
}
```

NLU (zero-shot ES)

model	intents	slots	exact
schuster-translate	85.39	72.87	54.95
schuster-cove	53.34	22.50	10.56

NLU (zero-shot ES)

model	intents	slots	exact
schuster-translate	85.39	72.87	54.95
schuster-cove	53.34	22.50	10.56
MaChAmp-NLU	84.21	75.01	56.02
MaChAmp-NLU-UD-sep	84.77	70.79	48.41
MaChAmp-NLU-UD-mixed	87.79	71.01	48.10

Results from: Cross-lingual Transfer Learning for Multilingual Task Oriented Dialog. Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis (NAACL 2019)

Other things this framework has been used for:

- POS tagging for code-switched data
- Biomedical event extraction
- Rik: effect of PMB predictions on Boxer performance
- Inspect BERT layer performances for UD tasks
- Nested named entity tagging for Danish
- ~~Lexical normalization~~

Smell you later!

Multiple languages (UPOS)