Sem-mmmBERT: Multi-task Learning with a Pre-defined set of Tasks and no Tuning

Rob van der Goot

September 29, 2022

Problem

"Recently, there has been a flurry of papers that show not only that multi-task learning helps pre-trained models, but that gains are larger when more tasks are used. Such massive multi-task learning settings cover up to around 100 tasks, going beyond earlier work that covered around 50 tasks (Aghajanyan et al., 2021)."

Problem

"Recently, there has been a flurry of papers that show not only that multi-task learning helps pre-trained models, but that gains are larger when more tasks are used. Such massive multi-task learning settings cover up to around 100 tasks, going beyond earlier work that covered around 50 tasks (Aghajanyan et al., 2021)."

`https://newsletter.ruder.io/issues/pre-training-massive-multi-tasking-709680/05e59718-2554-4a0c-84d2-4e1572a020a2`

Problem

"The newly proposed approaches differ in terms of how and when multi-task learning is applied. One choice is fine-tuning an existing pre-trained model on a collection of multiple tasks, i.e. behavioural fine-tuning. This is done by T0 (Sanh et al., 2021), one of the first outcomes of the BigScience workshop, using T5 and FLAN (Wei et al., 2021) using a GPT-3-like pre-trained model. "

Problem

multi-task models may soon hold state-of-the-art results on many benchmarks.

Problem

- Can we exploit a pre-selected combination of NLP tasks in a multi-task setup to improve the ability of an autoencoder language model to learn NLP tasks?

Problem

MaChAmp at SemEval-2022 tasks 2, 3, 4, 6, 10, 11, and 12: Multi-task Multi-lingual Learning for a Pre-selected Set of Semantic Datasets

Rob van der Goot
IT University of Copenhagen
robv@itu.dk

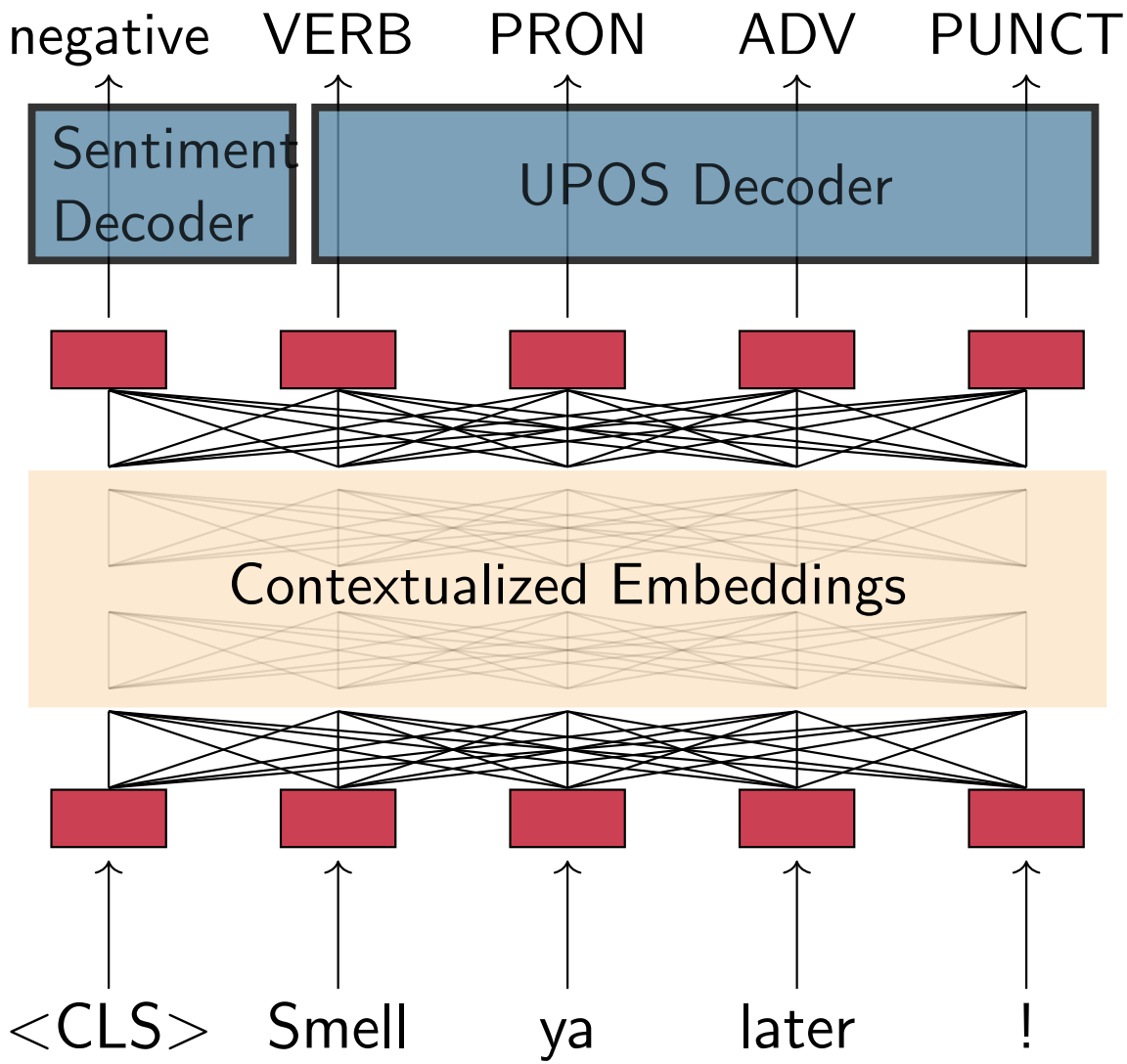
Setup (MaChAmp)

What is MaChAmp?

Setup (MaChAmp)

Also: a multi-task learning toolkit for NLP!

Setup (MaChAmp)



MaChAmp #068

Attacks:

seq	classification
seq-bio	dependency
seq-multi	mlm
string2string	seq2seq

MaChAmp is a multi-task NLP toolkit, it can seemingly effortlessly handle multiple NLP tasks simultaneously. It has functionality for joint training, continuous training, dataset smoothing, loss weights and dataset embeddings.

More information on:
machamp-nlp.github.io

Notes

How to use MaChAmp.

This is what the dataset configuration file looks like:

```
{"UD": {,
    "train_data_path": "data/ewt.train",
    "validation_data_path": "data/ewt.dev",
    "word_idx": 1,
    "tasks": {
      "lemma": {
        "task_type": "string2string",
        "column_idx": 2,
      }
    }
}
```

Then I can train with the following command:

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

And predict with:

```
python3 predict.py logs/ewt/model.tar.gz
   data/ewt.dev preds/ewt.dev.out
```

Setup (MaChAmp)

“ multi-task learning is much easier with recent models, even across many tasks. This is due to the fact that many recent models such as T5 and GPT-3 use a text-to-text format.”

Setup (MaChAmp)

“ multi-task learning is much easier with recent models, even across many tasks. This is due to the fact that many recent models such as T5 and GPT-3 use a text-to-text format.”

- Let’s do this in a non-easy way!

Setup (MaChAmp)

MULTITASK PROMPTED TRAINING ENABLES ZERO-SHOT TASK GENERALIZATION

Victor Sanh* Hugging Face
Albert Webson* Brown University
Colin Raffel* Hugging Face
Stephen H. Bach* Brown University

Lintang Sutawika BigScience
Zaid Alyafeai KFUPM
Antoine Chaffin IRISA & IMATAG
Arnaud Stiegler Hyperscience
Teven Le Scao Hugging Face

Arun Raja I^2R, Singapore
Manan Dey SAP
M Saiful Bari NTU, Singapore
Canwen Xu UCSD & Hugging Face
Urmish Thakker SambaNova Systems

Shanya Sharma Walmart Labs
Eliza Szczechla BigScience
Taewoon Kim VU Amsterdam
Gunjan Chhablani BigScience
Nihal V. Nayak Brown University

Debajyoti Datta University of Virginia
Jonathan Chang ASUS
Mike Tian-Jian Jiang ZEALS, Japan
Han Wang NYU
Matteo Manica IBM Research

Sheng Shen UC Berkeley
Zheng-Xin Yong Brown University
Harshit Pandey BigScience
Michael McKenna Parity
Rachel Bawden Inria, France

Thomas Wang Inria, France
Trishala Neeraj BigScience
Jos Rozen Naver Labs Europe
Abheesht Sharma BITS Pilani, India
Andrea Santilli University of Rome

Thibault Fevry BigScience
Jason Alan Fries Stanford University
Ryan Teehan Charles River Analytics
Tali Bers Brown University

Stella Biderman EleutherAI & Booz Allen
Leo Gao EleutherAI
Thomas Wolf Hugging Face
Alexander M. Rush Hugging Face

Setup (MaChAmp)

MaChAmp at SemEval-2022 tasks 2, 3, 4, 6, 10, 11, and 12: Multi-task Multi-lingual Learning for a Pre-selected Set of Semantic Datasets

Rob van der Goot
IT University of Copenhagen
robv@itu.dk

- Note that (almost) no tuning is done!
  - Is this a bad thing?

Setup (MaChAmp)

SemEval Task	Included sub-tasks	Languages	Citation
2: Multilingual Idiomaticity Detection	Idiomaticity detection (1-shot)	EN, PT, GL	[8, 7]
3: PreTENS	1: Binary acceptability	EN, IT, FR	[9]
	2: Regression acceptability	EN, IT, FR	
4: Patronizing and Condescending Language Detection	1: Binary PCL detection	EN	[6, 5]
	2: Multi-label PCL classification	EN	
6: iSarcasmEval	1: Sarcasm detection	EN, AR	[1]
	2: Irony-labeling	EN	
	3: Paraphrase sarcasm detection	EN, AR	
10: Structured Sentiment Analysis	Expressions, entities and relations	CA, EN, ES, EU, NO	[2]
11: MultiCoNER - Multilingual Complex Named Entity Recognition	Named Entity Recognition	BN, DE, EN, ES, FA, HI, KO, MI, NL, RU, TR, ZH	[4]
12: Symlink	Entities and relations	EN	[3]

Setup (MaChAmp)

Task	MaChAmp task-type	#words	#sents	#sents smoothed
2-a1	classification	10,199	139	2,742
3-1	classification	99,044	11,669	25,131
3-2	regression	4,761	785	6,518
4-1	classification	399,376	8,369	21,283
4-2	classification	135,750	2,202	10,917
6-a	classification	83,266	5,254	16,863
6-b	classification*6	12,183	691	6,115
6-c	classification	29,242	1,287	8,346
10	seq seq seq	1,109,260	58,799	56,413
11	seq_bio	2,768,898	171,300	96,288
12	seq seq	944,176	3,120	12,994

Table: The task-types used within MaChAmp for each of the (sub-)tasks, and the data size before and after smoothing.

Setup (MaChAmp)

Task	MaChAmp task-type	#words	#sents	#sents smoothed
2-a1	**classification**	10,199	139	2,742
3-1	classification	99,044	11,669	25,131
3-2	**regression**	4,761	785	6,518
4-1	classification	399,376	8,369	21,283
4-2	classification	135,750	2,202	10,917
6-a	classification	83,266	5,254	16,863
6-b	**classification*6**	12,183	691	6,115
6-c	classification	29,242	1,287	8,346
10	**seq seq seq**	1,109,260	58,799	56,413
11	seq_bio	2,768,898	171,300	96,288
12	**seq seq**	944,176	3,120	12,994

Table: The task-types used within MaChAmp for each of the (sub-)tasks, and the data size before and after smoothing.

2a: Multilingual Idiomaticity Detection

[CLS] bad hat [SEP] The disapproval is literally of the hats. [SEP] The moral character of a bad hat is secondary. [SEP] Shocking, used as a quasi-adverb like this, was thought a vulgarism [SEP]

2a: Multilingual Idiomaticity Detection

[CLS] bad hat [SEP] The disapproval is literally of the hats. [SEP] The moral character of a bad hat is secondary. [SEP] Shocking, used as a quasi-adverb like this, was thought a vulgarism [SEP]

Segment ID's 11100000000011111100000000000000

2a: Multilingual Idiomaticity Detection

```
{
    "SEMEVAL2-A1": {
        "train_data_path": "data/task2/train.all.conll",
        "validation_data_path": "data/task2/dev.all.conll",
        "sent_idxs": [2,3,4,5],
        "tasks": {
            "idiomaticy-1": {
                "column_idx": 6,
                "task_type": "classification",
                "metric": "macro-f1"
            }
        }
    }
}
```

3-2: PreTENS: acceptability regression

- Regression not supported!
- Added now!
- linear layer and mean square error loss

3-2: PreTENS: acceptability regression

```
{
    "SEMEVAL3-2": {
        "train_data_path": "data/task3/2.train.all.conll",
        "validation_data_path": "data/task3/2.dev.all.conll
        "sent_idxs": [1],
        "tasks": {
            "sts": {
                "task_type": "regression",
                "column_idx": 2,
                "metric": "spearman"
            }
        }
    }
}
```

6-2: iSarcasmEval: Irony labeling

- Classification task, but multi-label
- Each label as separate task

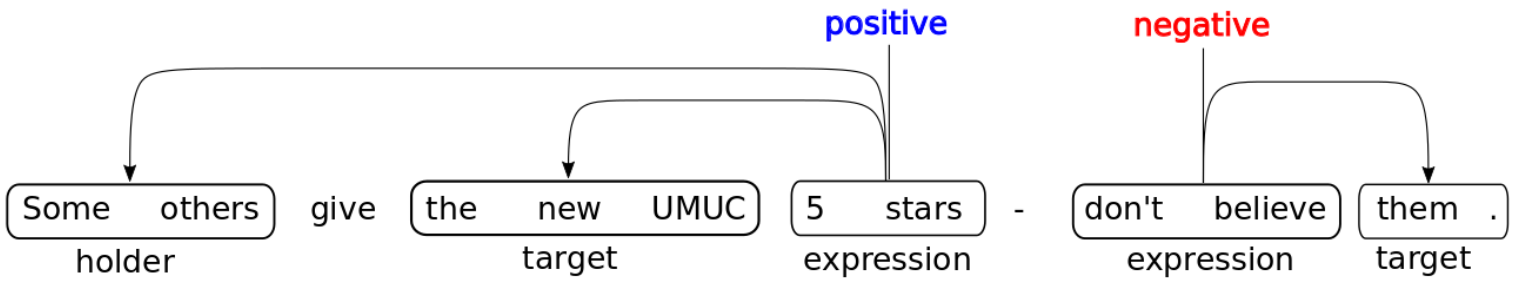
6-2: iSarcasmEval: Irony labeling

```
{
    "SEMEVAL6-b": {
        "train_data_path": "data/task6/2.train.en.conll",
        "validation_data_path": "data/task6/2.dev.en.conll"
        "sent_idxs": [1],
        "tasks": {
            "sarcasm": {
                "task_type": "classification",
                "column_idx": 4,
                "metric": "macro-f1"
            },
            "irony": {
                "task_type": "classification",
                "column_idx": 5,
                "metric": "macro-f1"
            },
            "satire": {
                "task_type": "classification",
```

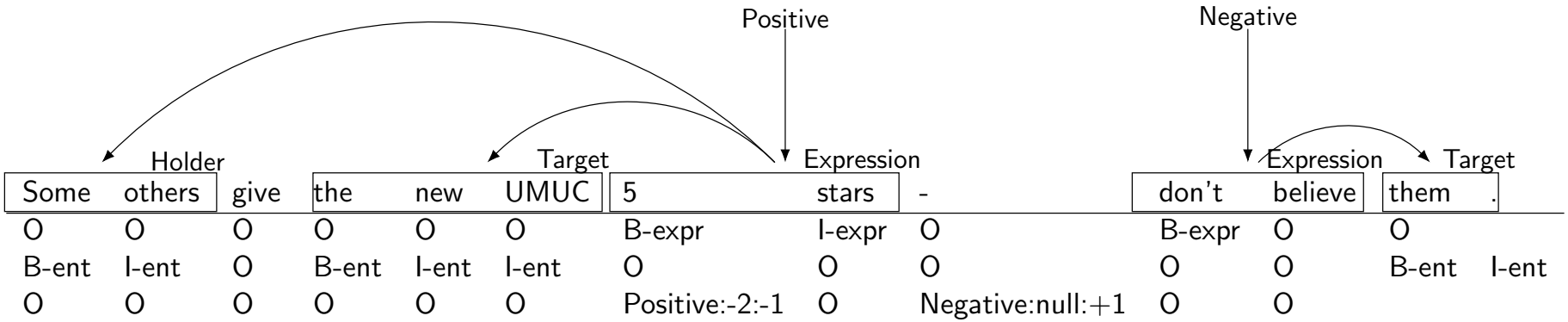
6-2: iSarcasmEval: Irony labeling


```
        "column_idx": 6,
        "metric": "macro-f1"
    },
    "understatement": {
        "task_type": "classification",
        "column_idx": 7,
        "metric": "macro-f1"
    },
    "overstatement": {
        "task_type": "classification",
        "column_idx": 8,
        "metric": "macro-f1"
    },
    "rhetorical_question": {
        "task_type": "classification",
        "column_idx": 9,
        "metric": "macro-f1"
    }
```

10: Structured Sentiment Analysis

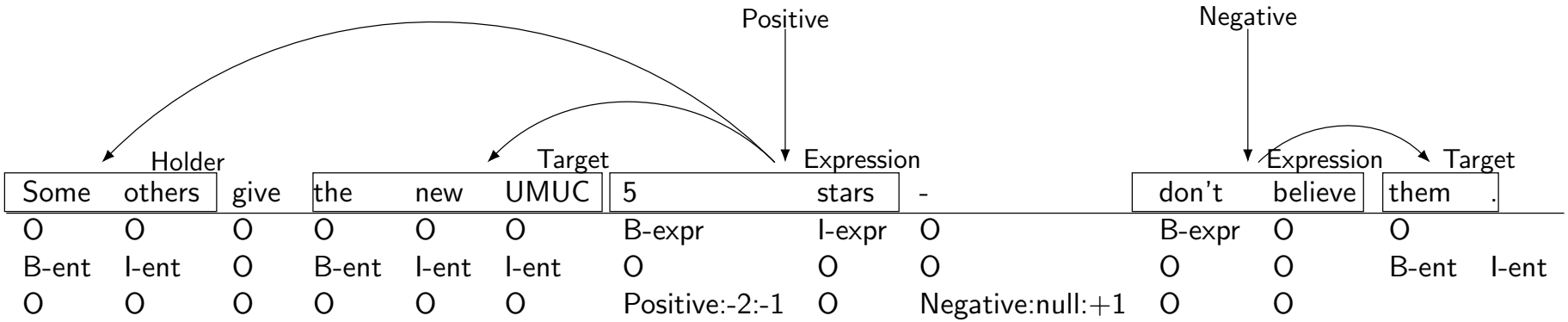


10: Structured Sentiment Analysis



- Inspired by Biomedical Event Extraction as Sequence Labeling (Ramponi et al, 2020)

10: Structured Sentiment Analysis



- Inspired by Biomedical Event Extraction as Sequence Labeling (Ramponi et al, 2020)
- Note that items can be overlapping, and are BIO-encoded
- However, the `seq` task-type outperformed `seq_bio` and `multiseq`.

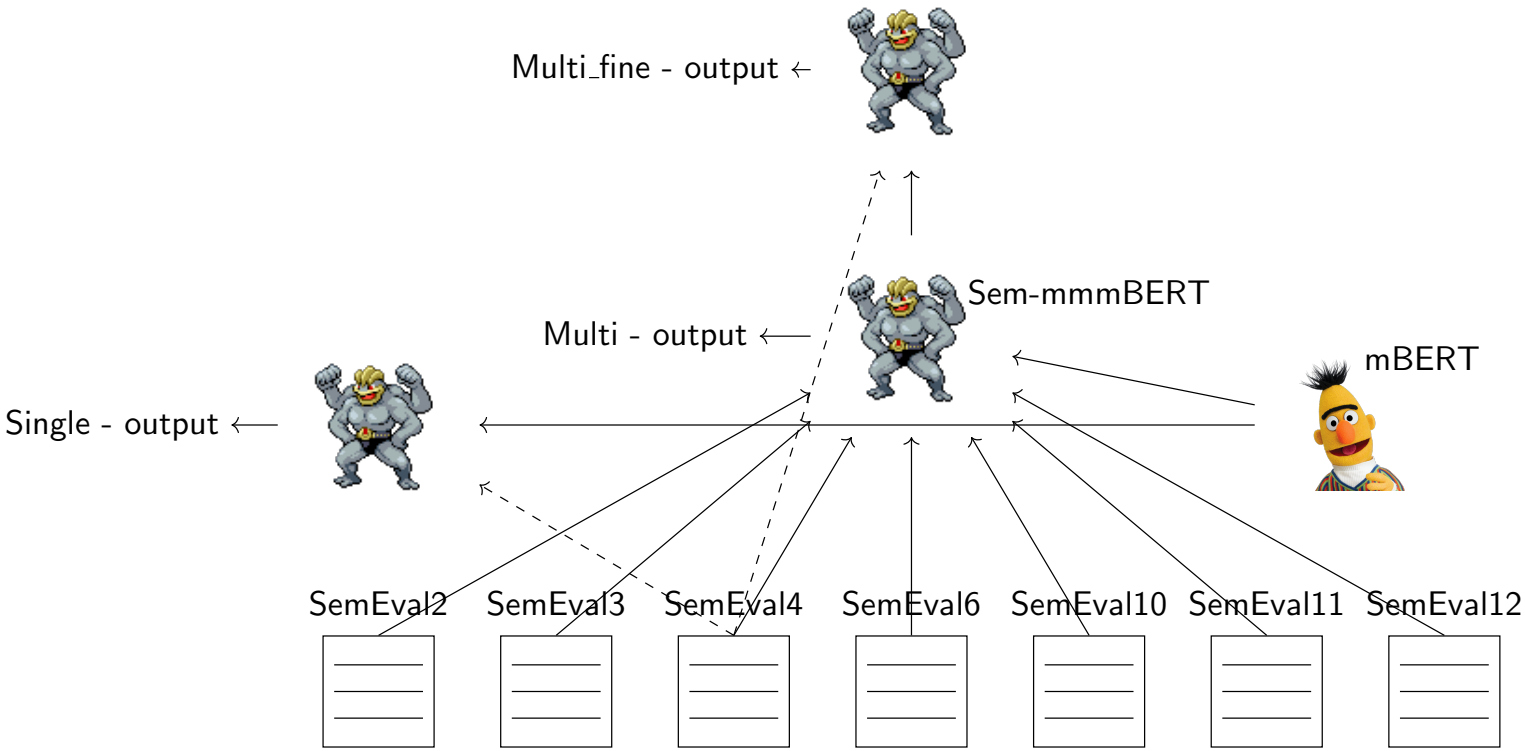
12: Symlink

- Similar as task 10, but linking mathematical symbols
- And non-tokenized input!
- Used `_is_punctuation` from huggingface, and save location of split
- Rest of procedure remains the same

12: Symlink

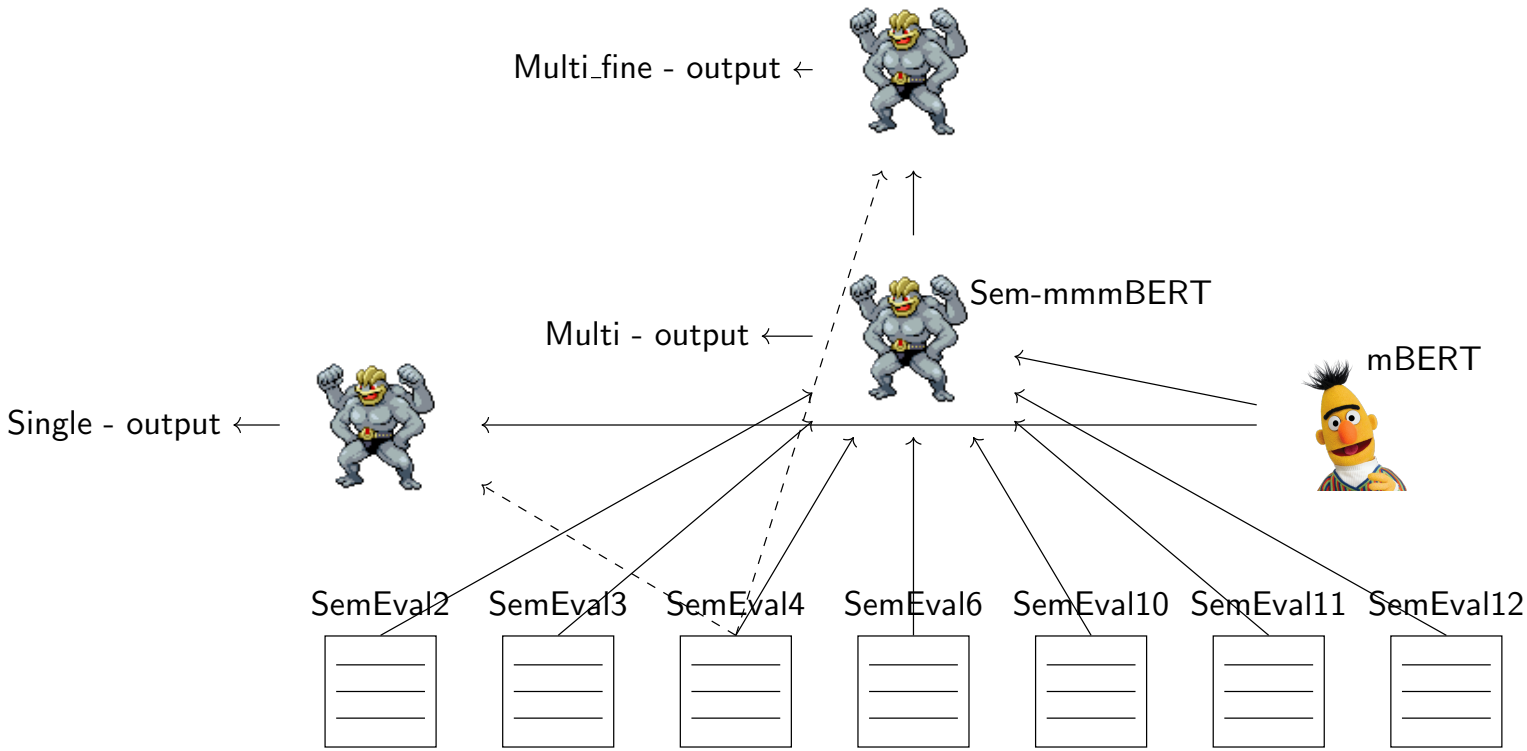
```
{
    "SEMEVAL12": {
        "train_data_path": "data/task12/train.all.conll",
        "validation_data_path": "data/task12/dev.all.conll"
        "word_idx": 1,
        "tasks": {
            "entities12": {
                "task_type": "seq",
                "column_idx": 2
            },
            "relations12": {
                "task_type": "seq",
                "column_idx": 3
            }
        }
    }
}
```

Setup (MaChAmp)



Setup (MaChAmp)

Setup (MaChAmp)



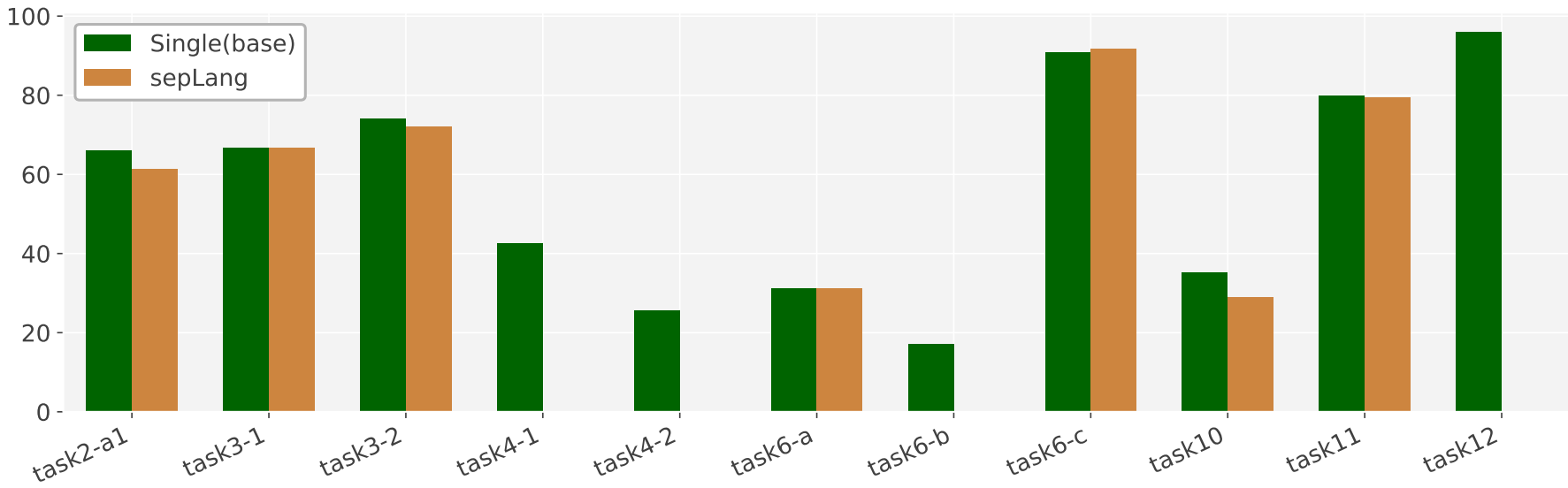
Setup (MaChAmp)

STILT: Supplementary Training on Intermediate Labeled-data Tasks (Phang et al. 2018)

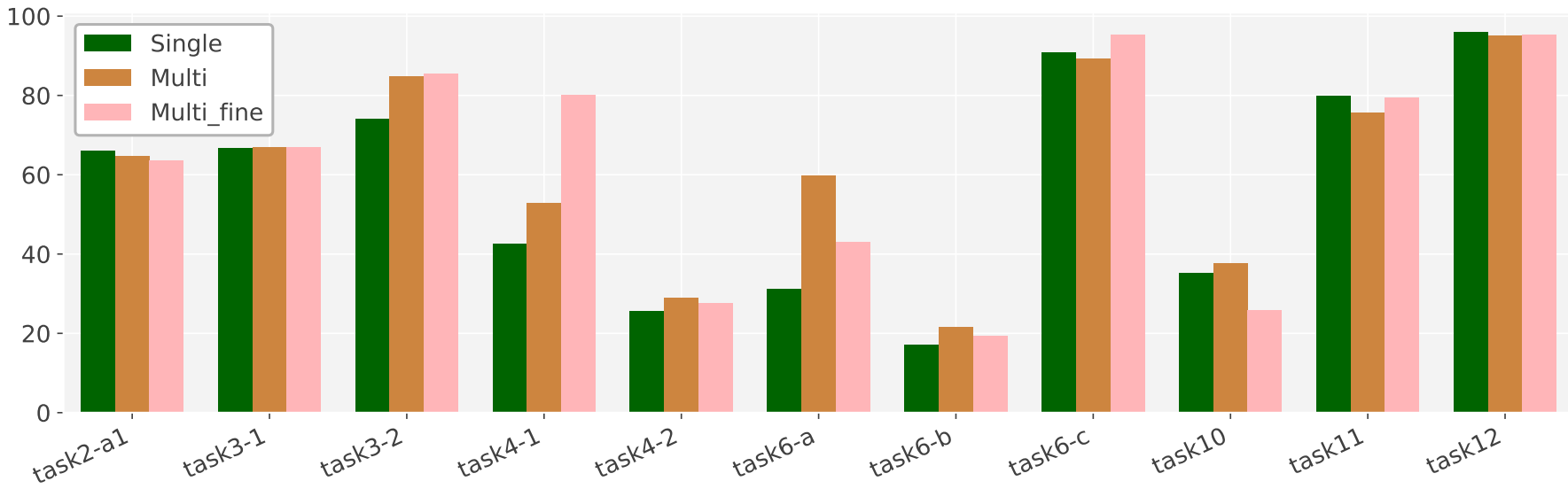
Setup (MaChAmp)

```
train.py --dataset_config config/task4.json
train.py --dataset_config config/*.json --name multi
train.py --finetune logs/multi/*/model.tar.gz \
 --dataset_config config/task4.json --name multi.task4
```

Setup (MaChAmp)



Setup (MaChAmp)



Setup (MaChAmp)

Task	Single mBERT	Multi_fine RemBERT	Ranking
task2-a1	—	66.07	NA
task3-1	78.78	86.42	11/21
task3-2	0.6792	-0.164	17/17 (3/17)
task4-1	0.4172	0.4211	56/78
task4-2	0.0772	0.1546	34/49
task6-a	0.3639	0.3187	31/43 & 12/32
task6-b	0.0919	0.0851	3/22
task6-c	0.2400	0.2250	16/16 & 13/13
task10	0.472	0.501	13/22
task11	0.6027	0.6768	18/26
task12	2.67	7.42	—

Setup (MaChAmp)

We will release:

- Sem-mmmBERT: Semeval-Machamp-Multitask-Multilingual BERT
- Sem-RemmmBERT: Semeval-Machamp-Multitask-Multilingual RemBERT

Setup (MaChAmp)

multi-task models may soon hold state-of-the-art results on many benchmarks.

- Can we do better?

Setup (MaChAmp)

multi-task models may soon hold state-of-the-art results on many benchmarks.

- Can we do better?
  - Use other LM's
  - Finetune hyperparameters
  - Add/select pre-training tasks

Setup (MaChAmp)

All code available at:
`https://bitbucket.org/robvanderg/semeval2022`
paper is on the way

MaChAmp: `https://machamp-nlp.github.io/`