
Rob and the mysteries in multi-task learning
and input representations for language models

Today

Multi-task learning:

- Auxiliary tasks
- Intermediate fine-tuning

Inputs to language models:

- Dataset embeddings
- Segment embeddings

Framework

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

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

xSID: Cross-lingual Slot and Intent Detection

Rob van der Goot, Ibrahim Sharaf, Aizhan Imankulova, Ahmet Üstün, Marija Stepanović, Alan Ramponi, Siti Oryza Khairunnisa, Mamoru Komachi and Barbara Plank









Slot and Intent Detection

I'd like to see the showtimes for Silly Movie 2.0 at the movie house
Intent: SearchScreeningEvent

xSID

ar	أود أن أرى مواعيد عرض فيلم Silly Movie 2.0 في دار السينما
da	Jeg vil gerne se spilletiderne for Silly Movie 2.0 i biografen
de	Ich würde gerne den Vorstellungsbeginn für Silly Movie 2.0 im Kino sehen
de-st	I mecht es Programm fir Silly Movie 2.0 in Film Haus sechn
en	I'd like to see the showtimes for Silly Movie 2.0 at the movie house
id	Saya ingin melihat jam tayang untuk Silly Movie 2.0 di gedung bioskop
it	Mi piacerebbe vedere gli orari degli spettacoli per Silly Movie 2.0 al cinema
ja	映画館のSilly Movie 2.0の上映時間を見せて。
kk	Мен Silly Movie 2.0 бағдарламасының кинотеатрда көрсетілім уақытын көргім келеді
nl	Ik wil graag de speeltijden van Silly Movie 2.0 in het filmhuis zien
sr	Želela bih da vidim raspored prikazivanja za Silly Movie 2.0 u bioskopu
tr	Silly Movie 2.0'ın sinema salonundaki seanslarını görmek istiyorum
zh	我想看Silly Movie 2.0 在影院 的放映

Experiments

Baselines

- Baseline: contextualized embeddings with joint intent+slots
- Stronger baseline: translate training data to target language and map slot labels with attention (NMT-TRANSFER)

Experiments

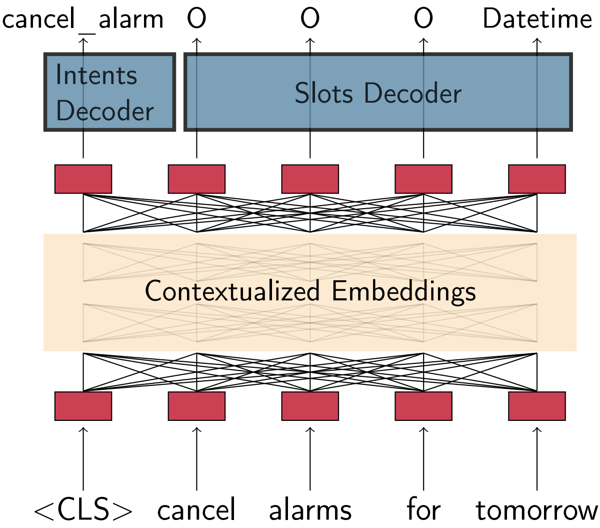
Baselines

- Baseline: contextualized embeddings with joint intent+slots
- Stronger baseline: translate training data to target language and map slot labels with attention (NMT-TRANSFER)

New models:

- Train on auxiliary task in target language:
  - Masked language modeling (AUX-MLM)
  - Neural machine translation (AUX-NMT)
  - UD-parsing (AUX-UD)

Baseline


cancel_alarm
O
O
O
Datetime
Intents Decoder
Slots Decoder
Contextualized Embeddings
<CLS>
cancel
alarms
for
tomorrow

Experiments

Evaluate 2 embeddings

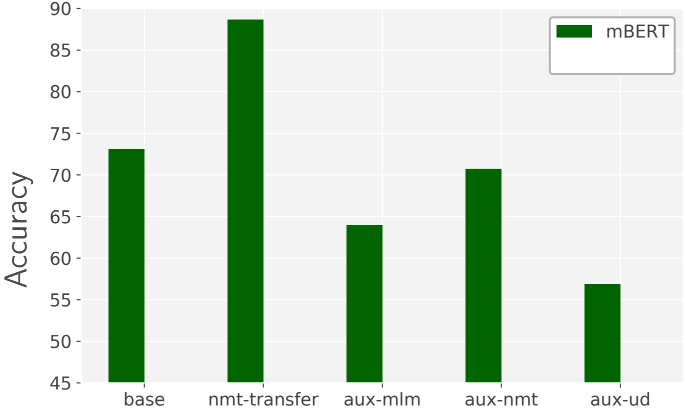
- mBERT: trained on 104 languages (12/13)
- XLM15: trained on 15 languages (5/13)

Results

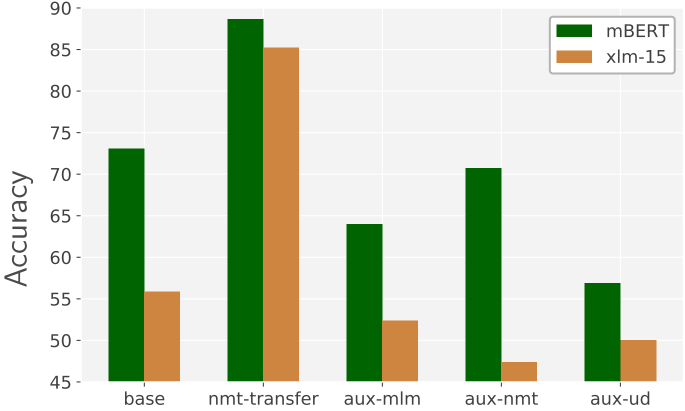
model	Time (minutes)
base	46
nmt-transfer	5,213
aux-mlm	193
aux-nmt	373
aux-ud	79

Table: Average minutes to train a model, averaged over all languages and both embeddings. For nmt-transfer we include the training of the NMT model.

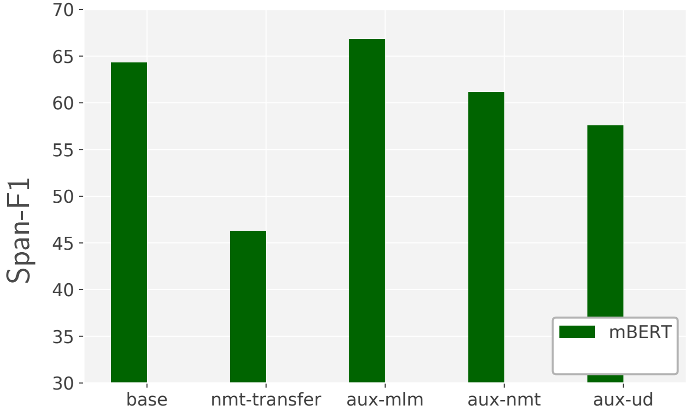
Results (intents)



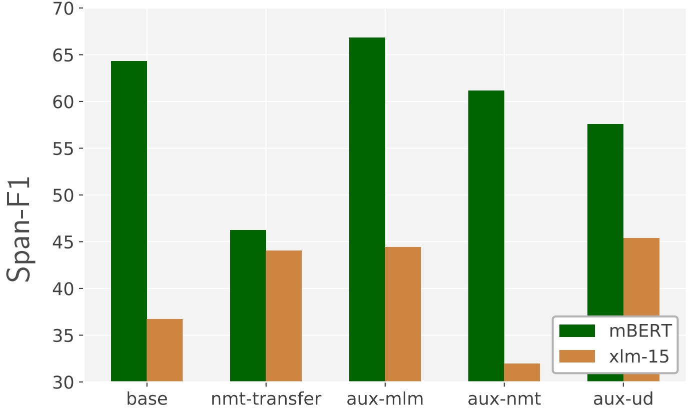
Results (intents)



Results (slots)



Results (slots)



Resolved mysteries

Sentence level:

- NMT-transfer is hard to outperform, but costly
- Even baseline hard to beat

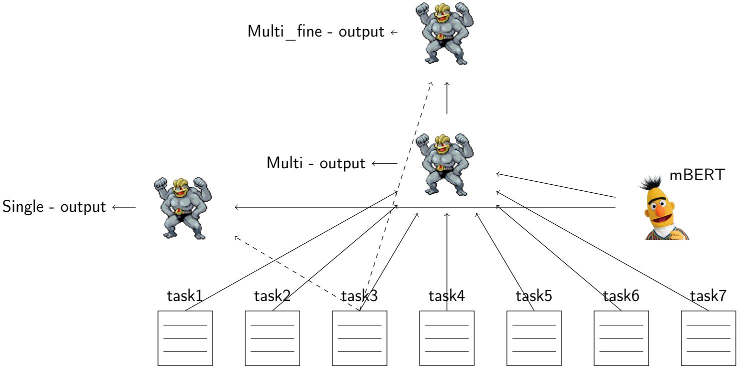
Span level:

- NMT-transfer performs bad (due to alignment)
- In-LM languages: only MLM helps
- Out-LM languages: More explicit tasks (UD) are faster and lead to better performance

Unresolved mysteries

- Can NMT be used as auxiliary task?
- Are there better sentence level auxiliary tasks?
- Can NMT-transfer be improved with better word alignment?
- NMT and MLM hyperparameters
- Modeling jointly versus sequentially

A newer multi-task setup: Intermediate task finetuning



Other names:

- Task Adaptive PreTraining (TAPT)
- Pre-finetune
- Multi-task finetuning
- Multi-task prompted training
- Supplementary training on intermediate labeled data tasks (STILT)
- Intermediate task finetuning
- Intermediate task training
- Intertraining
- ...

Intermediate task finetuning

- STILT
- T0
- Ext5
- MUPPET
- In-BoXBART
- Sem-mmmBERT
- ...

Intermediate task finetuning

- STILT
- T0
- Ext5
- MUPPET
- In-BoXBART
- Sem-mmmBERT
- ...

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

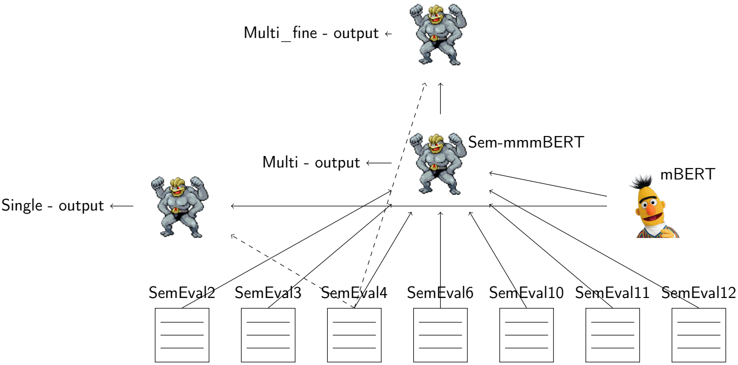
Rob van der Goot
IT University of Copenhagen
robv@itu.dk

Intermediate task finetuning

Research questions:

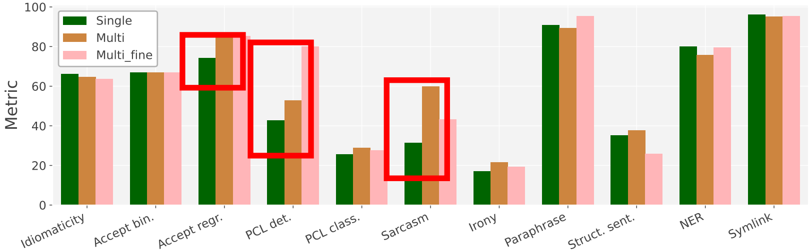
- Can we use this approach in an autoencoder language model?
- Is intermediate task finetuning also beneficial for a somewhat arbitrary set of semantic tasks?

Intermediate task finetuning



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

Intermediate task finetuning



Resolved mysteries

- Medium performance baseline tgt task
- Largest gains for some sub-tasks (task-relatedness)
- Language

Unresolved mysteries

- How can we do better?
  - Use other LM's
  - Finetune hyperparameters
  - Add/select pre-training tasks
- Can we predict which tasks to select?
  - Hard without many overlapping datasets (task/language dimension)
  - Too many combinations possible

Parsing with Pretrained Language Models, Multiple Datasets, and Dataset Embeddings

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

Miryam de Lhoneux
Uppsala University
KU Leuven
University of Copenhagen
ml@di.ku.dk

Dataset embeddings

- Embed the data source to inform the model
- Allow to learn dataset specific information (about data, annotation, etc.) **and** commonalities
- Usually concatenated to word embedding

Dataset embeddings

How can we inform a BERT-like model of the data source?

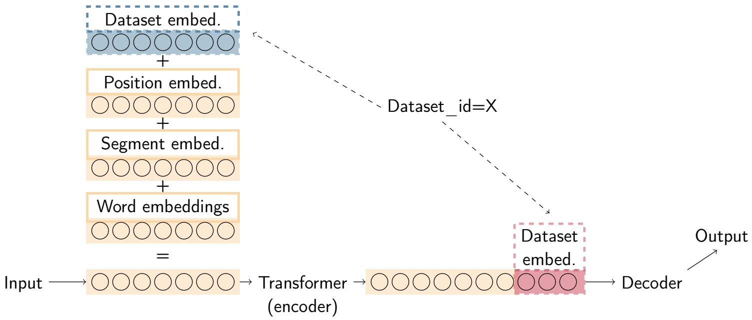
- Concatenate to the output of BERT (before task-specific decoder)

Dataset embeddings

How can we inform a BERT-like model of the data source?

- Concatenate to the output of BERT (before task-specific decoder)
- Sum to input embedding

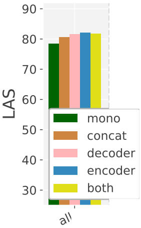
Dataset embeddings



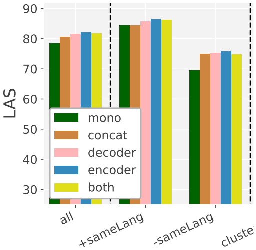
Setup

- Universal Dependencies data
- Language clusters from Smith et al. (2018)
- Baselines:
  - mono: single treebank training
  - concat: concatenate treebanks of cluster

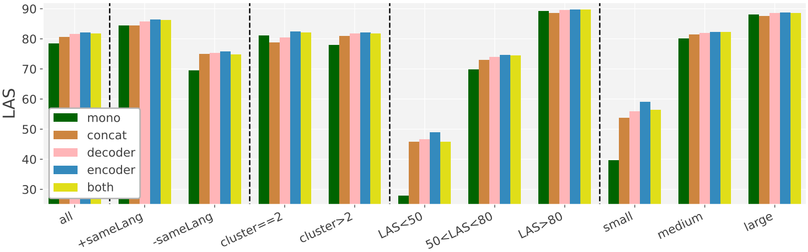
Dataset embeddings



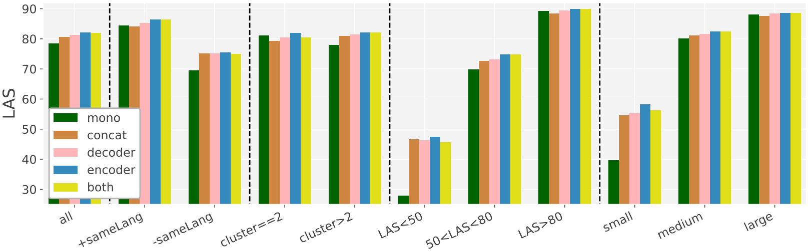
Dataset embeddings



Dataset embeddings



Dataset embeddings (Trained on all)



Resolved mysteries

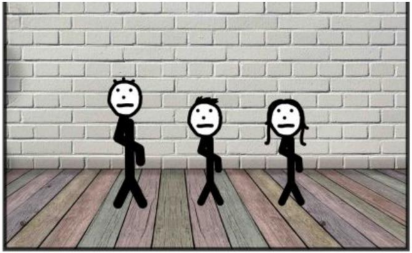
- Cheap method for performance improvement
- Encoder > decoder
- Training on all treebanks outperforms training on clusters
- Note that you can also embed other things

Unresolved mysteries

- What if we do not know the data source of our input?
- `https://aclanthology.org/2021.adaptnlp-1.19.pdf`
- `https://aclanthology.org/2020.acl-main.778.pdf`

Frustratingly Easy Performance Improvements for Low-resource Setups: A Tale on BERT and Segment Embeddings

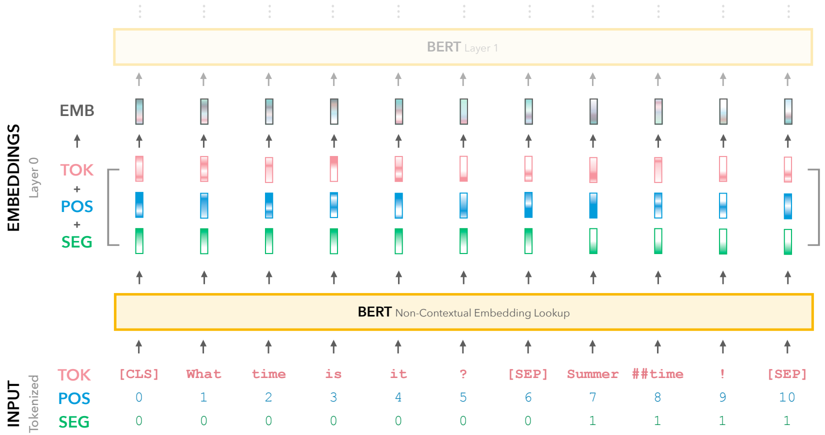
Rob van der Goot,[♣] Max Müller-Eberstein,[♣] Barbara Plank[♣◇]
[♣]Computer Science Department, IT University of Copenhagen
[◇]Center for Information and Language Processing (CIS), LMU Munich, Germany
robv@itu.dk, mamy@itu.dk, bapl@itu.dk



Segment embeddings

- Under explored feature of BERT
- Interesting?

Inputs to modern language models



Inputs to modern language models

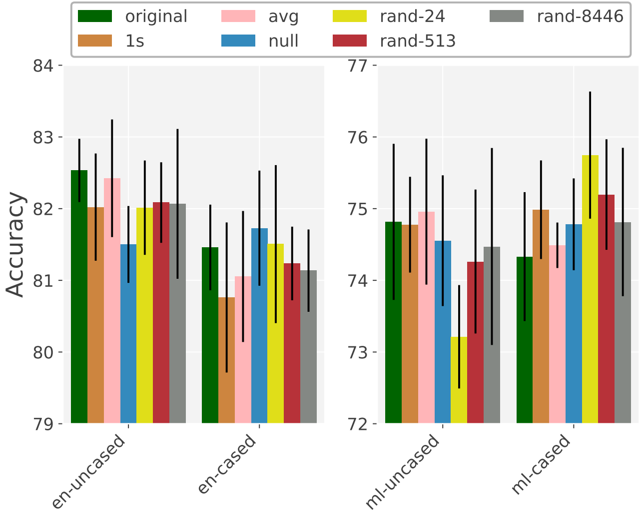
	TOK	[CLS]	first	?	[SEP]	second	!	[SEP]
	POS	0	1	2	3	4	5	6
		+	+	+	+	+	+	+
SEG	ORIGINAL	0	0	0	0	1	1	1
	1S	1	1	1	1	1	1	1
	AVG	0	0	0	0	0	0	0
	NULL	0	0	0	0	0	0	0
	RAND	0	0	0	0	0	0	0
	0S	0	0	0	0	0	0	0

Setup

- Compare three levels of annotation: sentence-pair, sentence, word
- Used GLUE tasks and UD subset from Smith et al. (2018)

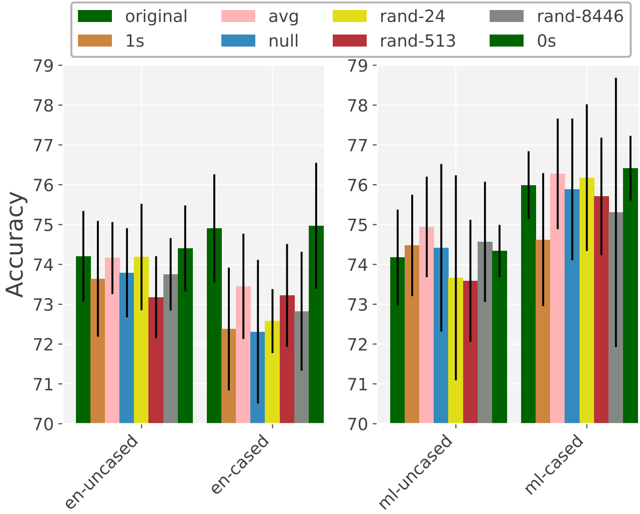
Inputs to modern language models

Glue single sentence tasks:



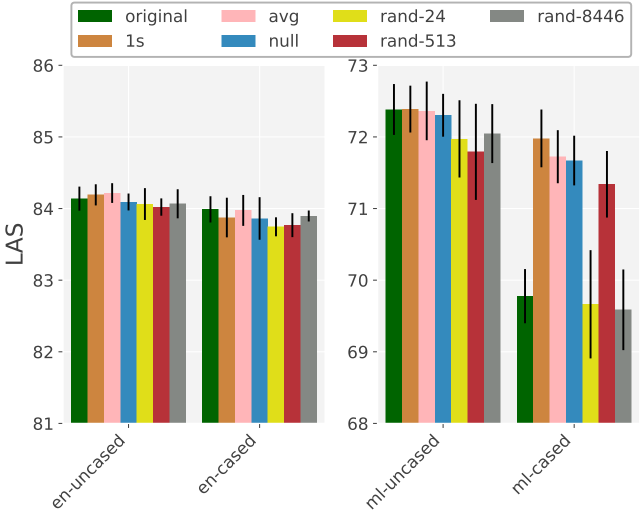
Inputs to modern language models

Glue sentence-pair tasks:



Inputs to modern language models

UD:



Resolved mysteries

- We can gain substantial performance improvements in multi-lingual bert models on word-level tasks by setting the segment id to 1
- Area code of Rob's parents address is a good alternative random seed

Unresolved mysteries

- Why?
- What is stored in the segment embeddings?
- Was it just luck?
- How is mBERT exactly trained?
- What is the difference between cased and uncased mBERT?

Other things I did:

- MultiLexNorm (Including DA/NL)
- Frisian-Dutch UD
- Cross-domain dialogue act classification social media
- NER: Danish, Classical Arabic
- Experimental setup (Tune set)
- Unsupervised code-switch detection
- Biomedical event extraction
- Cross-domain dependency parsing (Max), Job postings (Mike), Relation Extraction (Elisa)
- Tokenization with PLM
- How large should my dev/test be


Thanks!
Questions?