Multi-lingual and Multi-task Learning: from Dataset Creation to Modeling.

Rob van der Goot

MilaNLP 22-10-2021

- MultiLexNorm
- xSID
- MaChAmp

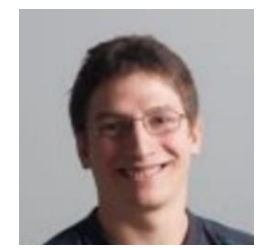
- MultiLexNorm
- xSID
- MaChAmp
  - Newest features!

MultiLexNorm: A Shared Task on Multilingual Lexical Normalization

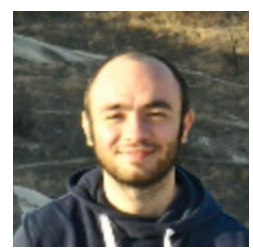
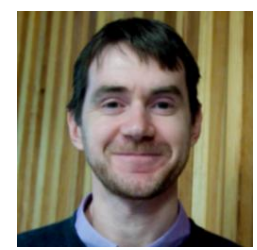
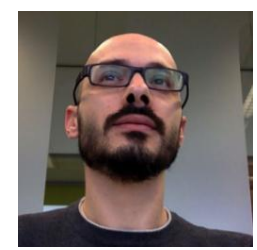
Rob van der Goot, Alan Ramponi, Arkaitz Zubiaga, Barbara Plank, Benjamin Muller, Iñaki San Vicente Roncal, Nikola Ljubešić, Özlem Çetinoğlu, Rahmad Mahendra, Talha Çolakoğlu, Timothy Baldwin, Tommaso Caselli and Wladimir Sidorenko







Lexical Normalization

Han & Baldwin (2011):
"a mapping from 'ill-formed' out-of-vocabulary (OOV) lexical items to their standard lexical forms."

Lexical Normalization

NLPProgress.com:
Lexical normalization is the task of translating/transforming a non standard text to a standard register.
Example:

new	pix	comming	tomoroe
new	pictures	coming	tomorrow

For lexical normalization, only replacements on the word-level are annotated. Some corpora include annotation for 1-N and N-1 replacements. However, word insertion/deletion and reordering is not part of the task.

Lexical Normalization

Lexical normalization is the task of transforming an utterance into its standard form, word by word, including both one-to-many (1-n) and many-to-one (n-1) replacements.

Lexical Normalization

State before shared task:

- Most work on English
- Also work on single other languages
- Varieties in task definitions, guidelines and metrics
- No common evaluation benchmark

Lexical Normalization

State before shared task:

- Most work on English
- Also work on single other languages
- Varieties in task definitions, guidelines and metrics
- No common evaluation benchmark
- Only multi-lingual work (>2) evaluated on 7 languages (van der Goot, 2019)

MultiLexNorm

- Combination of existing datasets
- Annotation style and file format converged
- “new” evaluation metric
- External evaluation (UD)

MultiLexNorm

Lang.	Language name	Normalization example
DA	Danish	De skarpe lamper gjorde destromindre ek bedre . De skarpe lamper gjorde destro mindre ikke bedre .
DE	German	ogäj isch hätts auch dwiddern könn Okay ich hätte es auch twittern können
EN	English	u hve to let ppl decide what dey want to do you have to let people decide what they want to do
ES	Spanish	@username cuuxamee sii peroo veen yaa eem @username escúchame sí pero ven ya eh
HR	Croatian	svi frendovi mi nešto rade , veceras san osta sam . svi frendovi mi nešto rade , večeras sam ostao sam .
ID-EN	Indonesian-English	pdhal not fully bcs those ppl jg sih . padahal not fully because those people juga sih .
IT	Italian	a Roma è cosí primavera che sembra gia giov a Roma è così primavera che sembra già giovedì
NL	Dutch	Kga me wss trg rolle vant lachn Ik ga me waarschijnlijk terug rollen van het lachen
SL	Slovenian	jst bi tud najdu kovanec vreden veliko denarja . jaz bi tudi našel kovanec vreden veliko denarja .
SR	Serbian	komunalci kace pocne kaznjavanje ? komunalci kad počne kažnjavanje ?
TR	Turkish	He o dediyin suala cvb verdim He o dediğin suale cevap verdim
TR-DE	Turkish-German	@username Yerimm senii , damkee schatzymm :-* @username Yerim seni , danke Schatzym :-*

Is the sample of languages biased?

Metric

- Previously: accuracy, accuracy over OOV words, F1 score, BLEU, word error rate, character error rate, etc.
- Now: accuracy normalized for amount of words to be normalized. Error Reduction Rate:

$$ERR = \frac{\%\text{accuracy} - \%\text{words\_not\_normed}}{100 - \%\text{words\_not\_normed}}$$

MultiLexNorm: Results

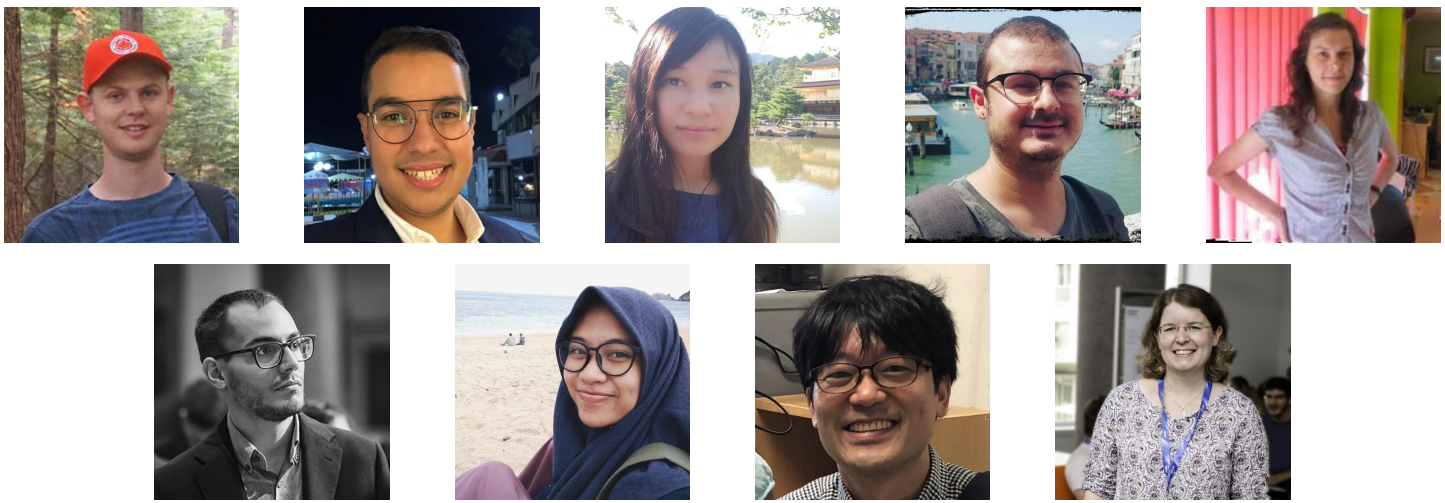
Team	Avg.	DA	DE	EN	ES	HR	ID-EN	IT	NL	SL	SR	TR	TR-DE
ÚFAL-2	**67.30**	**68.67**	**66.22**	**75.60**	59.25	**67.74**	**67.18**	**47.52**	**63.58**	**80.07**	**74.59**	**68.58**	**68.62**
HEL-LJU-2*	53.58	56.65	59.80	62.05	35.55	56.24	55.33	35.64	45.88	66.97	66.44	51.18	51.18
MoNoise	49.02	51.27	46.96	74.35	45.53	52.63	59.79	21.78	49.53	61.91	59.58	28.21	36.72
TrinkaAI-2	43.75	45.89	47.30	65.96	**61.33**	41.28	56.36	15.84	45.74	59.51	44.52	15.54	25.77
thunderml-1	43.44	46.52	46.62	64.07	60.29	40.09	59.11	11.88	44.05	59.33	44.46	15.88	29.01
team-2	40.70	48.10	46.06	63.73	21.00	40.39	59.28	13.86	43.72	60.55	46.11	15.88	29.71
learnML-2	40.30	40.51	43.69	61.57	56.55	38.11	56.19	5.94	42.77	58.25	39.99	14.36	25.68
maet-1	40.05	48.10	46.06	63.90	21.00	40.39	59.28	5.94	43.72	60.55	46.11	15.88	29.71
MFR	38.37	49.68	32.09	64.93	25.57	36.52	61.17	16.83	37.70	56.71	42.62	14.53	22.09
CL-MoNoise-2*	19.27	16.77	33.45	28.44	15.80	14.07	9.11	38.61	20.68	6.96	21.00	7.43	18.93
BLUE-2	6.73	49.68	-1.91	26.81	-9.36	-10.06	-7.22	-31.68	-2.09	-1.04	42.62	9.97	14.99
LAI	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MaChAmp*	-21.25	-88.92	-93.36	50.99	25.36	42.62	39.52	-312.87	1.49	56.80	39.44	-12.67	-3.42

MultiLexNorm: Results

treebank	avg.	de-tweede	en-aae	en-monoise	en-tweebank2	it-postwita	it-twittiro	tr-iwt151
ÚFAL-2	**64.17**-37	**73.58**-5	**62.73**-53	58.57-33	59.08-66	**68.28**-13	72.22-5	54.74-90
HEL-LJU-1*	63.73-14	73.49-3	60.48-18	56.29-9	60.27-18	68.15-3	**72.46**-0	54.94-48
MoNoise	63.44-21	73.20-5	62.27-40	56.83-18	58.90-46	67.55-3	70.69-0	54.61-35
MFR	63.31-16	72.86-5	60.32-32	56.74-15	60.31-37	67.34-3	70.72-0	54.89-25
TrinkaAI-2	63.12-33	72.86-7	60.16-40	56.64-19	59.87-39	66.98-7	71.14-0	54.20-119
maet-1	63.09-27	72.80-3	59.44-40	56.64-24	59.80-44	67.41-10	71.07-0	54.45-74
team-2	63.03-27	72.80-3	59.44-40	56.64-24	59.80-44	67.19-4	70.86-3	54.45-74
thunderml-2	63.02-33	72.67-3	59.57-42	56.74-28	59.25-44	67.34-4	71.35-1	54.24-112
learnML-2	62.88-30	72.31-8	58.98-44	56.16-31	59.98-45	66.99-6	71.24-0	54.48-79
CL-MoNoise*	62.71-24	72.65-0	60.90-0	55.26-0	58.53-0	66.53-99	70.10-50	54.98-20
BLUE-2	62.53-0	72.57-0	59.57-0	54.20-0	59.81-0	66.74-0	69.99-0	54.84-0
LAI	62.45-0	72.71-0	59.21-0	53.65-0	59.99-0	66.49-0	70.06-0	55.00-0
MaChAmp*	61.89-43	71.28-7	60.77-37	54.61-43	57.97-52	64.65-2	69.82-0	54.08-162
gold	—	—	—	**60.84**-18	**60.35**-0	—	—	—

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

Cross-lingual Slot and Intent Detection (xSID):

- 250 test and 150 dev from Facebook (Schuster et al., 2019) data
- 250 test and 150 dev from Snips (Coucke et al., 2018) data
- Concatenate them and converge annotation style
- Colleague-sourced to get higher quality/consistency as crowd-sourced

xSID

Cross-lingual Slot and Intent Detection (xSID):

- 250 test and 150 dev from Facebook (Schuster et al., 2019) data
- 250 test and 150 dev from Snips (Coucke et al., 2018) data
- Concatenate them and converge annotation style
- Colleague-sourced to get higher quality/consistency as crowd-sourced
- Kappa agreement on Dutch slots (3 annotators): 0.92

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	Silly Movie 2.0 の上映時間を見せて。映画館
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

MaChAmp

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



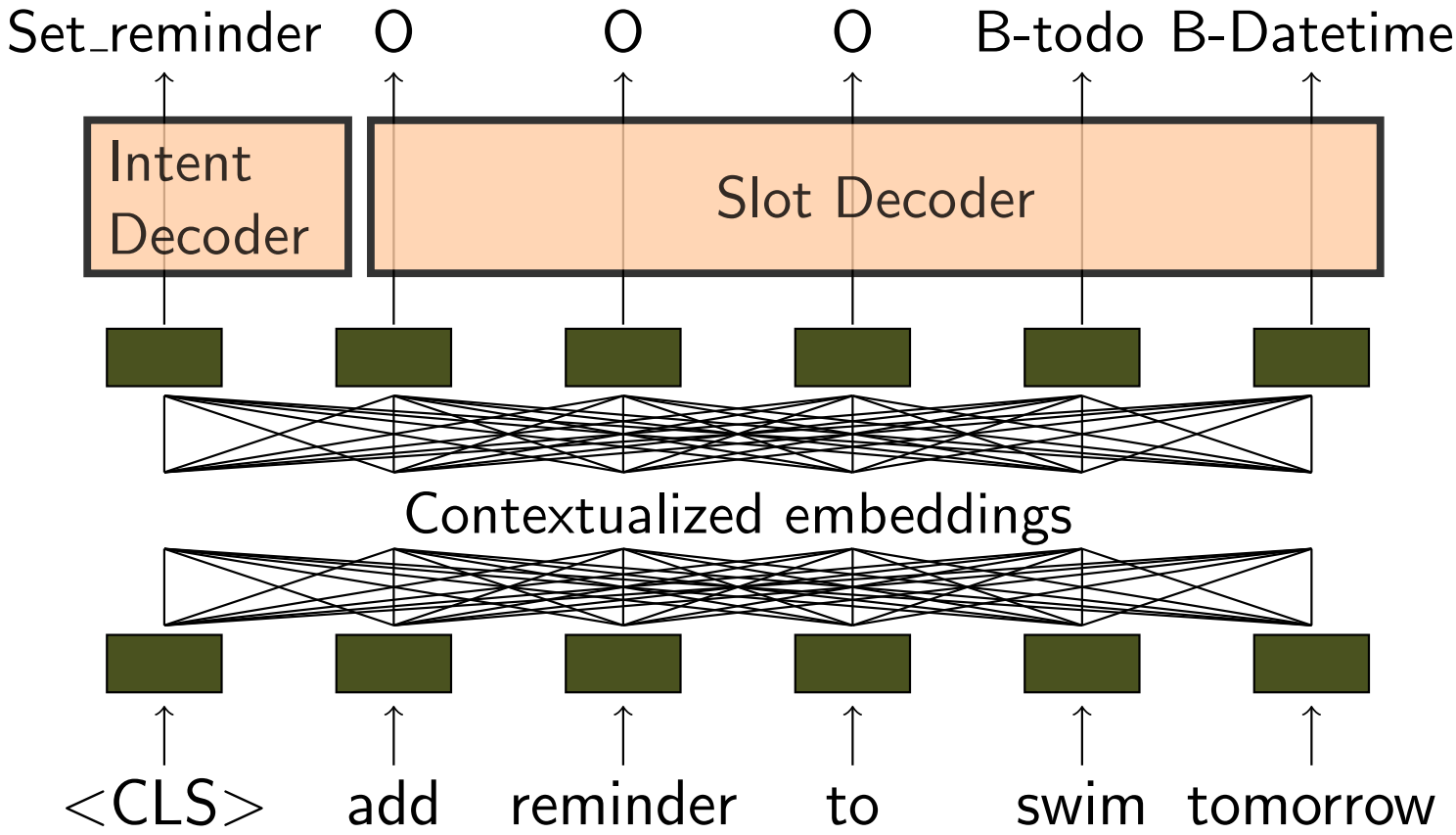



MaChAmp

Task-types:

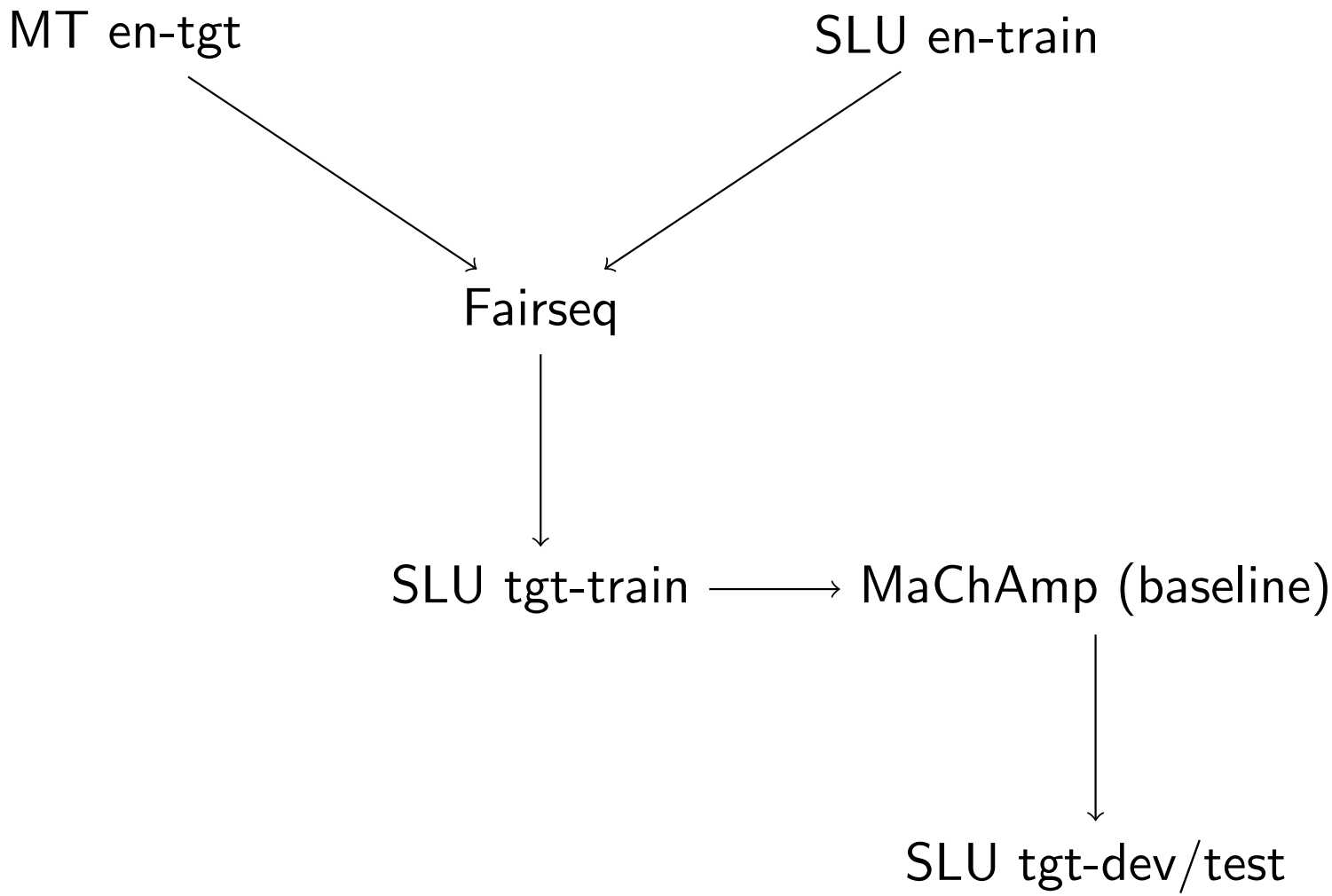
- seq
- string2string
- seq_bio
- multiseq
- dependency
- classification
- mlm
- seq2seq
- ...

MaChAmp


Set_reminder
O
O
O
B-todo
B-Datetime
Intent Decoder
Slot Decoder
Contextualized embeddings
<CLS>
add
reminder
to
swim
tomorrow

Experiments

NMT-TRANSFER



MT training data: Ted Talks + Opensubtitles

Experiments

Proposed models:

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

How easy is this?

```
{
    "train_data_path": "data/xSID/en.train",
    "validation_data_path": "data/xSID/en.dev",
    "word_idx": 0,
    "tasks": {
        "slots":
        {
            "task_type": "seq",
            "column_idx": 1
        }
        "intents":
        {
            "task_type": "classification",
            "column_idx": -1
        }
    }
}
```

How easy is this?

```
{
    "train_data_path": "data/opensubtitles/en-it.train",
    "validation_data_path": "data/opensubtitles/en-it.dev",
    "sent_idxs": [0],
    "tasks": {
        "nmt":
        {
            "task_type": "seq2seq",
            "column_idx": 1
        }
    }
}
```

How easy is this?

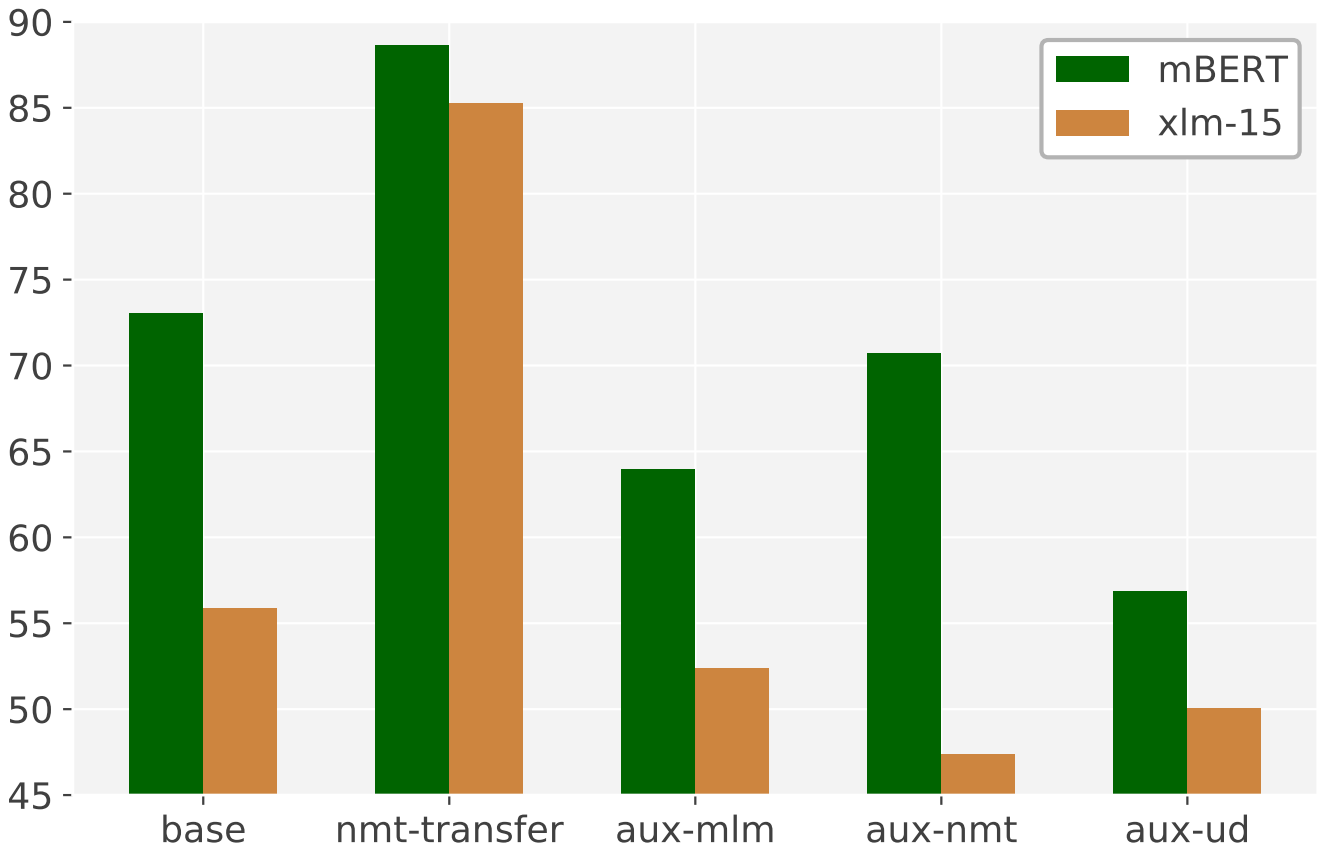
```
python3 train.py --dataset_configs configs/xsid.json \
configs/nmt-it.json
```

Experiments

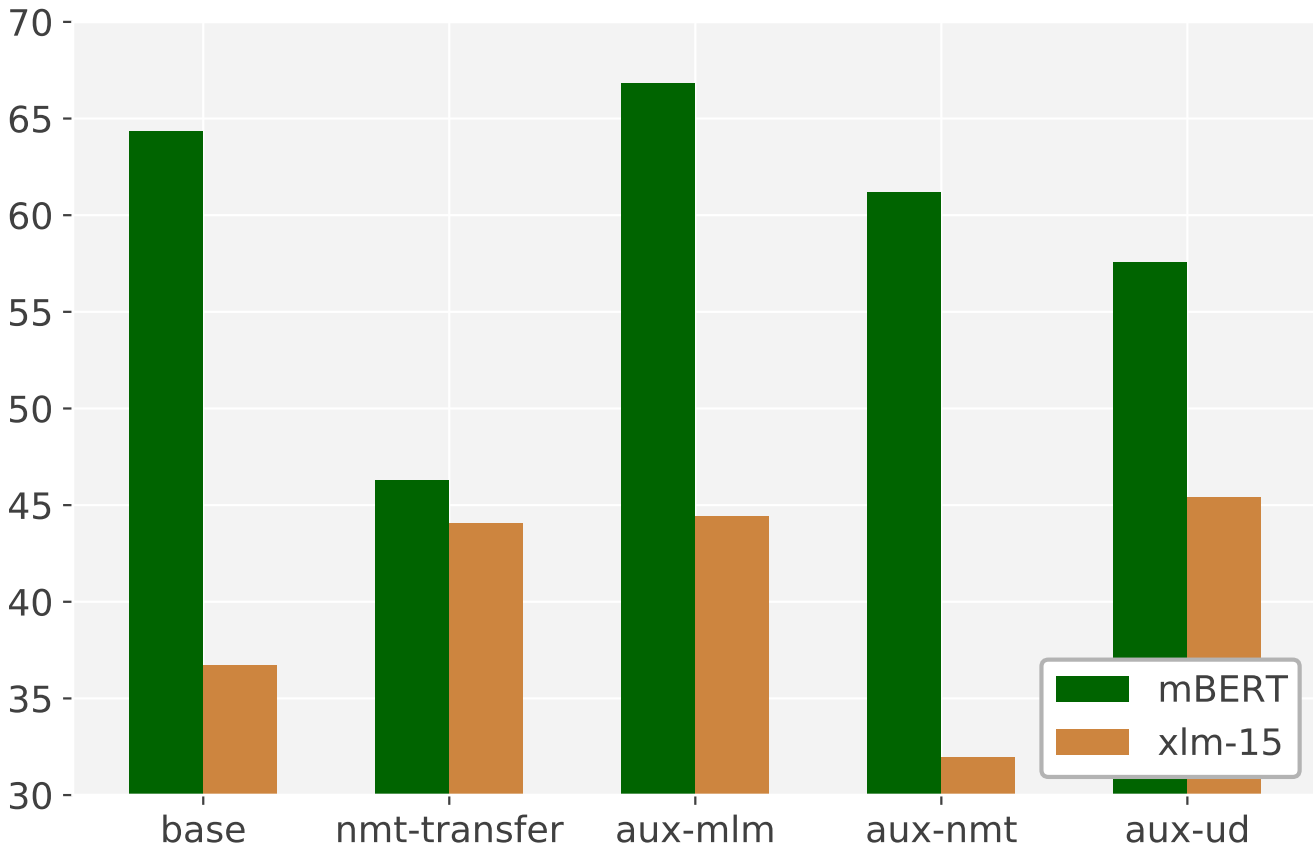
Evaluate 2 embeddings

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

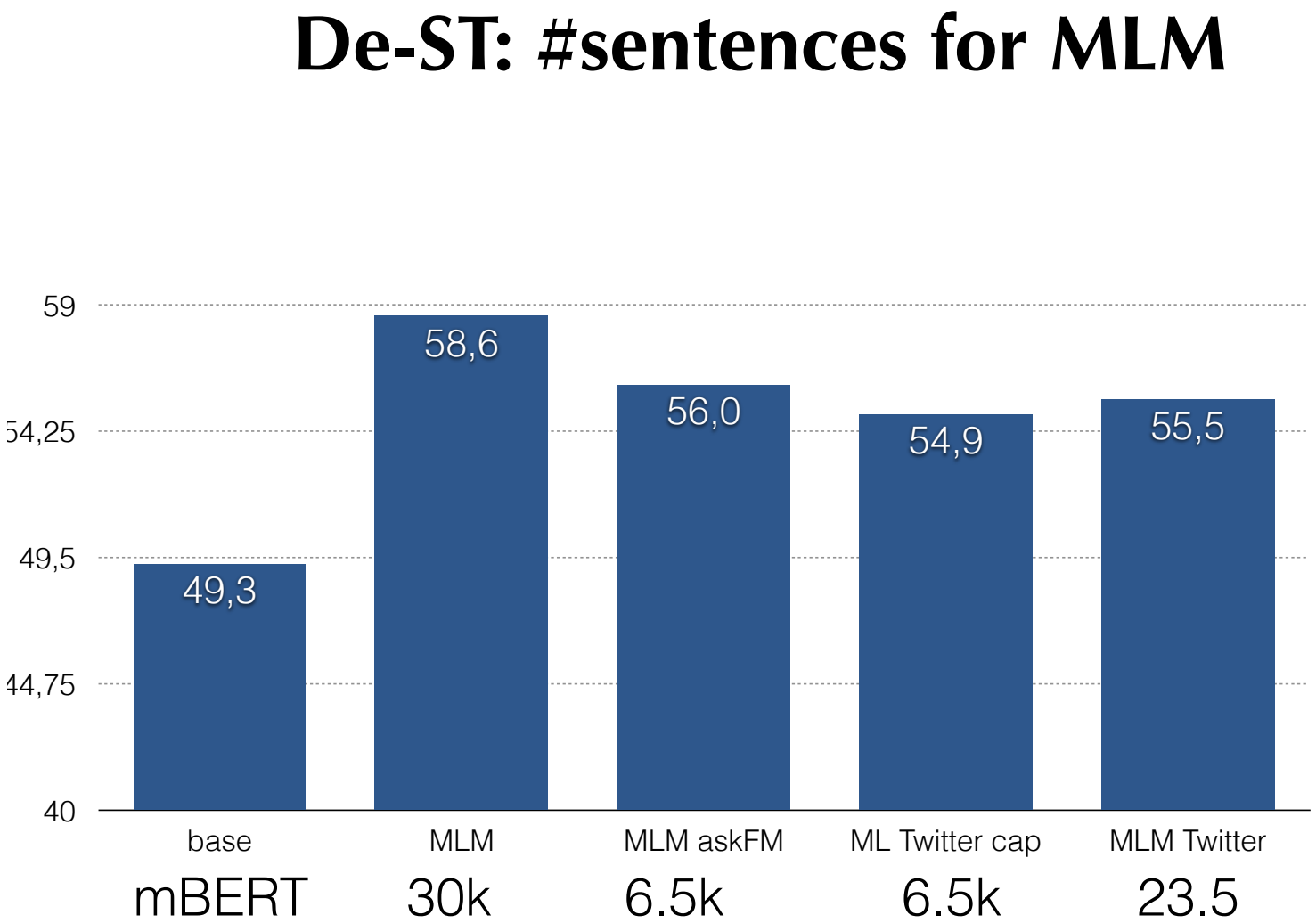
Results (intents)



Results (slots)



A closer look at South-Tyrolean



77

Conclusions

- xSID: More language (varieties), label (domain/topic) variety and high quality annotation
- For intents, nmt-transfer is hard to beat, and auxiliary tasks are not beneficial
- MLM performs robust as auxiliary task for slots
- For unseen languages, UD performs even better (XLM15)

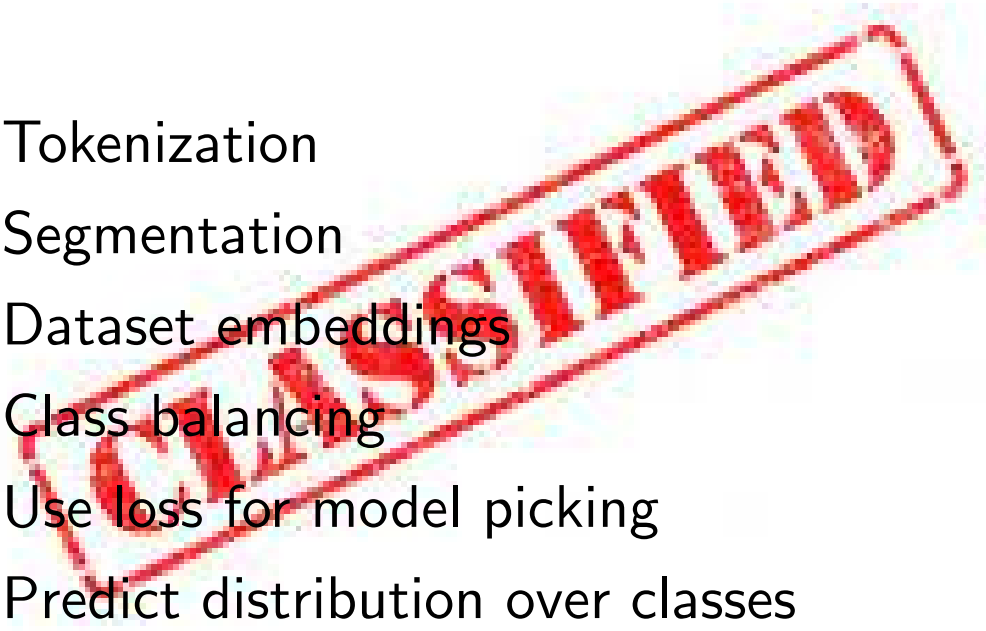
Other interesting MaChAmp features

- Dataset smoothing
- Loss-weighing
- Sequential and joint training
- Massively multi-lingual parsing models available
- Slack channel for support

Future MaChAmp features

Future MaChAmp features

- Tokenization
- Segmentation
- Dataset embeddings
- Class balancing
- Use loss for model picking
- Predict distribution over classes
- Embed output of previous task



Recommended things to check if you'd like to know more:

- Poster MaChAmp
- Presentation MaChAmp
- Presentation xSID

If you would like to try/use MaChAmp (or need a new feature), let me know!

Recommended things to check if you'd like to know more:

- Poster MaChAmp
- Presentation MaChAmp
- Presentation xSID

If you would like to try/use MaChAmp (or need a new feature), let me know!
We are also always interested in adding languages to our datasets, annotation is easy!