LREC-COLING 2024
Torino

Slot and Intent Detection Resources for Bavarian and Lithuanian: Assessing Translations vs Natural Queries to Digital Assistants

Miriam Winkler, Virginija Juozapaityte, Rob van der Goot, Barbara Plank

mi.winkler@campus.lmu.de
robv@itu.dk
b.plank@lmu.de

NaLiBaSID

Slot and intent detection evaluation dataset for Bavarian and Lithuanian

What will the weather be in New York city this week?

(de-ba) Wia werds Weda in New York City dera Woch?

(lt) Koks oras bus šią savaitę Niujorko mieste?

NaLiBaSID

Effect of Naturalness vs. Translations

→ ‘Translationese’ = traces of source language in the translation

Wos is grod für a weda in äding?

(en) What is the weather like in Altötting right now?

Pateikt rytojaus orų prognozę Vilniuje.

(en) Give tomorrow’s weather forecast for Vilnius.

NaLiBaSID - Translated Datasets

Dataset	Translation Src	Native	Intents	Slots	# sents
de-ba	xSID	–	16	34	800
lt	xSID	–	16	34	800
MAS:de-ba	MASSIVE	–	14	27	2,021
xMAS:de-ba	MASSIVE+xSID	–	16	34	2,821
nat:de-ba	n/a	collected	16	26	315
nat:lt	n/a	collected	16	30	327

- Translations from xSID (van der Goot et al., 2021)
- Translations from MASSIVE (FitzGerald et al., 2022)
  → Cross-dataset performance evaluation

NaLiBaSID - Translated Datasets

Dataset	Translation Src	Native	Intents	Slots	# sents
de-ba	xSID	–	16	34	800
lt	xSID	–	16	34	800
MAS:de-ba	MASSIVE	–	14	27	2,021
xMAS:de-ba	MASSIVE+xSID	–	16	34	2,821
nat:de-ba	n/a	collected	16	26	315
nat:lt	n/a	collected	16	30	327

- Translations from xSID (van der Goot et al., 2021)
- Translations from MASSIVE (FitzGerald et al., 2022)
  → Cross-dataset performance evaluation

Intent	**MASSIVE**	**xSID**
PlayMusic	588	63
weather/find	439	202
set_alarm	225	53
show_alarms	176	29
set_reminder	171	50
show_reminders	169	31
cancel_alarm	104	55
SearchScreeningEvent	59	60
AddToPlaylist	36	53
BookRestaurant	27	69
cancel_reminder	14	26
SearchCreativeWork	6	52
modify_alarm	4	1
snooze_alarm	3	5
RateBook	0	47
time_left_on_alarm	0	4

NaLiBaSID - Translated Datasets

Dataset	Translation Src	Native	Intents	Slots	# sents
de-ba	xSID	–	16	34	800
lt	xSID	–	16	34	800
MAS:de-ba	MASSIVE	–	14	27	2,021
xMAS:de-ba	MASSIVE+xSID	–	16	34	2,821
nat:de-ba	n/a	collected	16	26	315
nat:lt	n/a	collected	16	30	327

- Translations from xSID (van der Goot et al., 2021)
- Translations from MASSIVE (FitzGerald et al., 2022)
  → Cross-dataset performance evaluation

Intent	MASSIVE	xSID
PlayMusic	588	63
weather/find	439	202
set_alarm	225	53
show_alarms	176	29
set_reminder	171	50
show_reminders	169	31
cancel_alarm	104	55
SearchScreeningEvent	59	60
AddToPlaylist	36	53
BookRestaurant	27	69
cancel_reminder	14	26
SearchCreativeWork	6	52
modify_alarm	4	1
snooze_alarm	3	5
RateBook	0	47
time_left_on_alarm	0	4

NaLiBaSID - Natural Datasets

Dataset	Translation Src	Native	Intents	Slots	# sents
de-ba	xSID	–	16	34	800
lt	xSID	–	16	34	800
MAS:de-ba	MASSIVE	–	14	27	2,021
xMAS:de-ba	MASSIVE+xSID	–	16	34	2,821
nat:de-ba	n/a	collected	16	26	315
nat:lt	n/a	collected	16	30	327

- Collected from native speakers of the respective language with questionnaires
- Manually annotated with xSID intents and slots

NaLiBaSID - Natural Bavarian

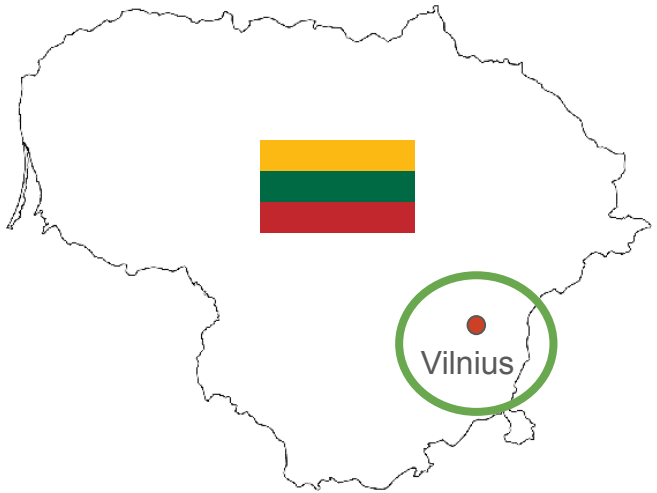
- Spelling variations
- Caused by the lack of standard orthography

‘stellen’ (to set)

- ‘stei’
- ‘stö’
- ‘steu’
- ‘stoi’

NaLiBaSID - Natural Lithuanian

- Digital assistants not very common in Lithuania
  → Led to production of unsuitable sentences for NaLiBaSID

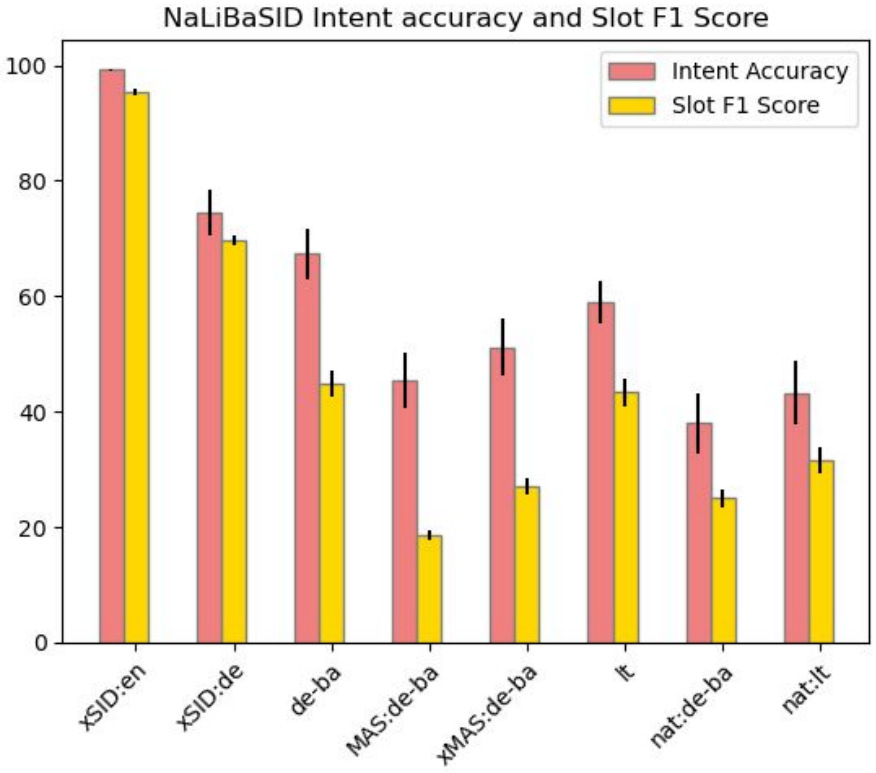


Experimental Setup

MaChAmp toolkit + **mBERT**

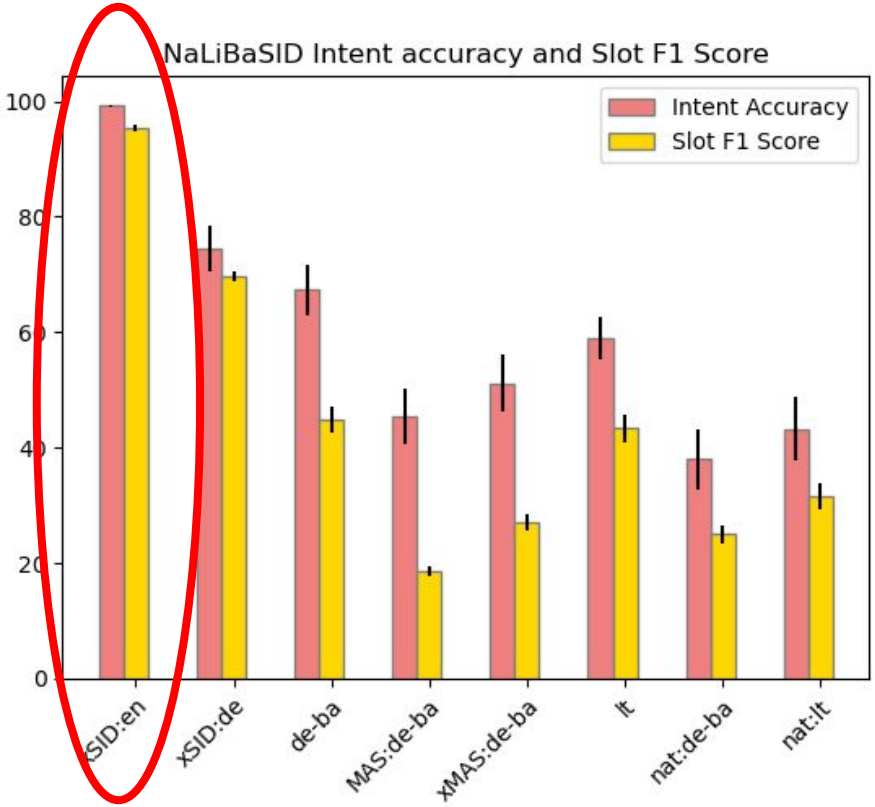
(van der Goot et al., 2021) (Devlin et al., 2019)

Results - Intent Accuracy



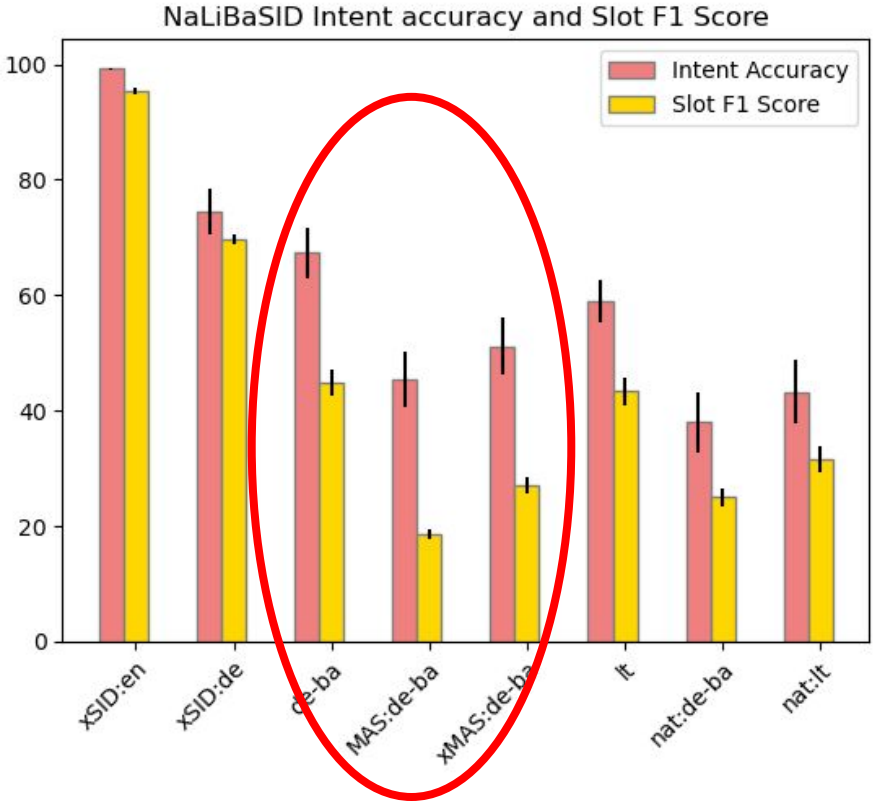
Intent classification is easy for standard languages

Results - Intent Accuracy



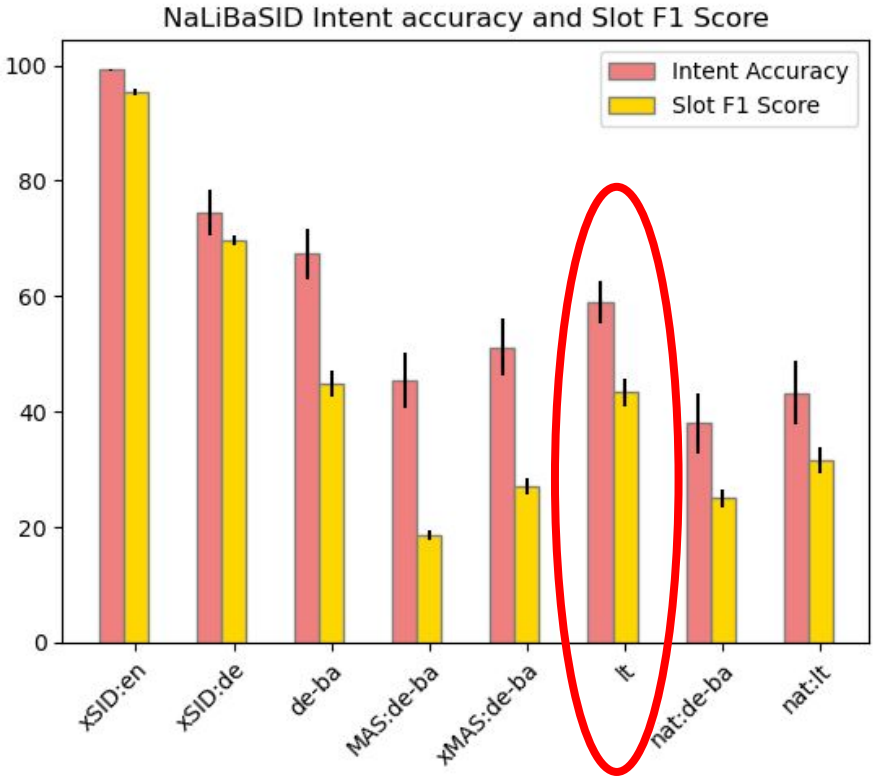
Intent classification is easy for standard languages

Results - Intent Accuracy



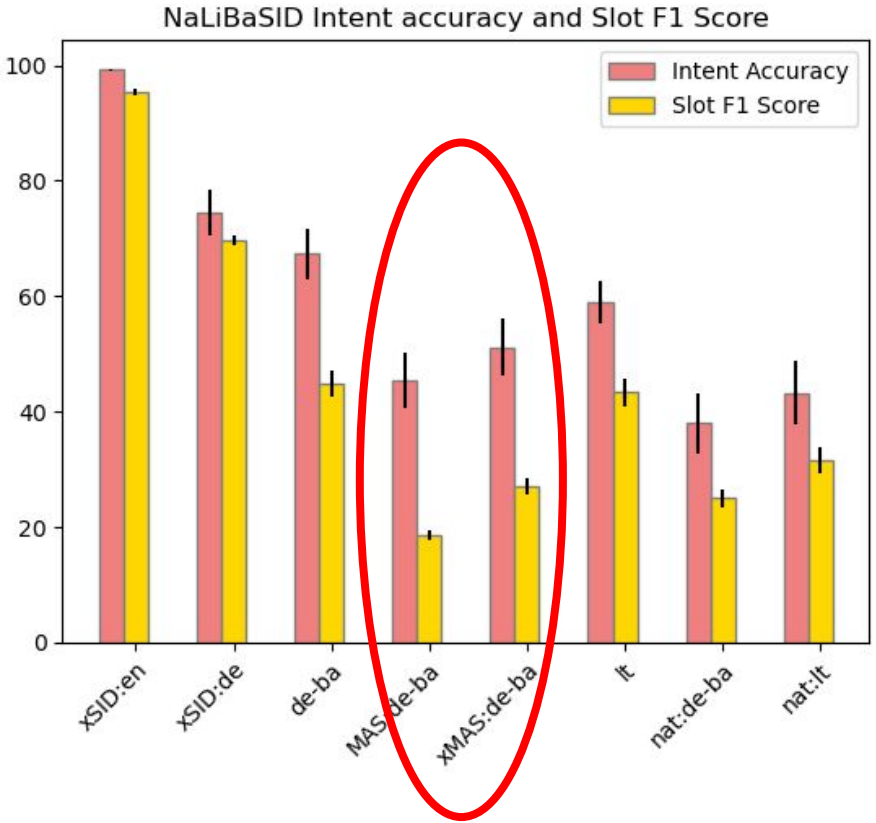
Accuracy drops on the translated Bavarian datasets but overall good performance

Results - Intent Accuracy



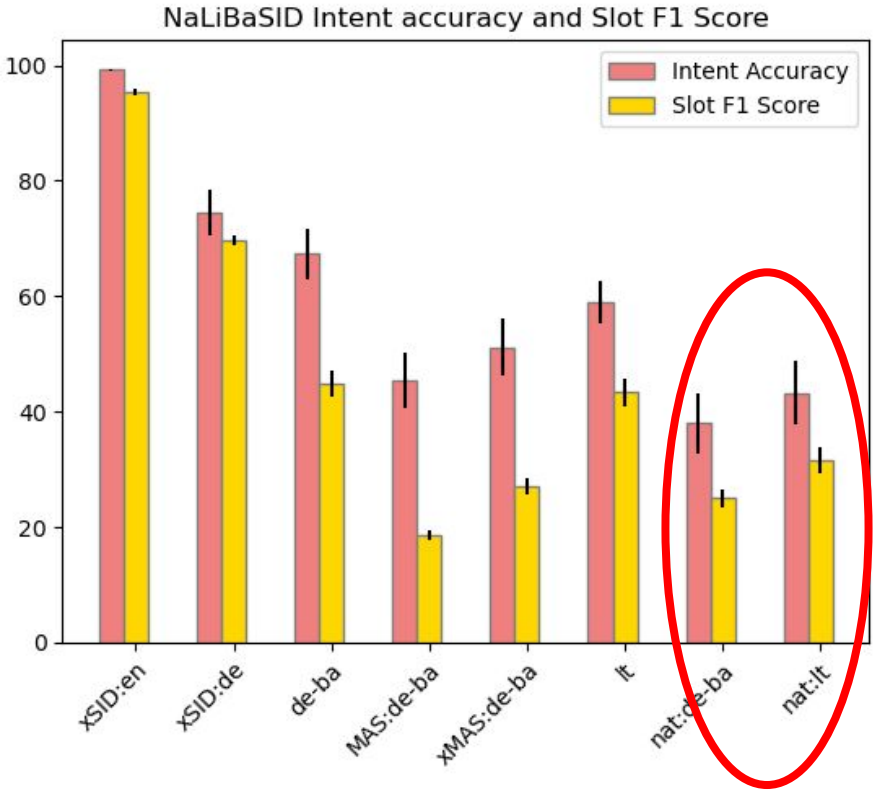
Similar performance for the Lithuanian translated data

Results - Intent Accuracy



Bavarian MASSIVE translations perform worse than Bavarian xSID

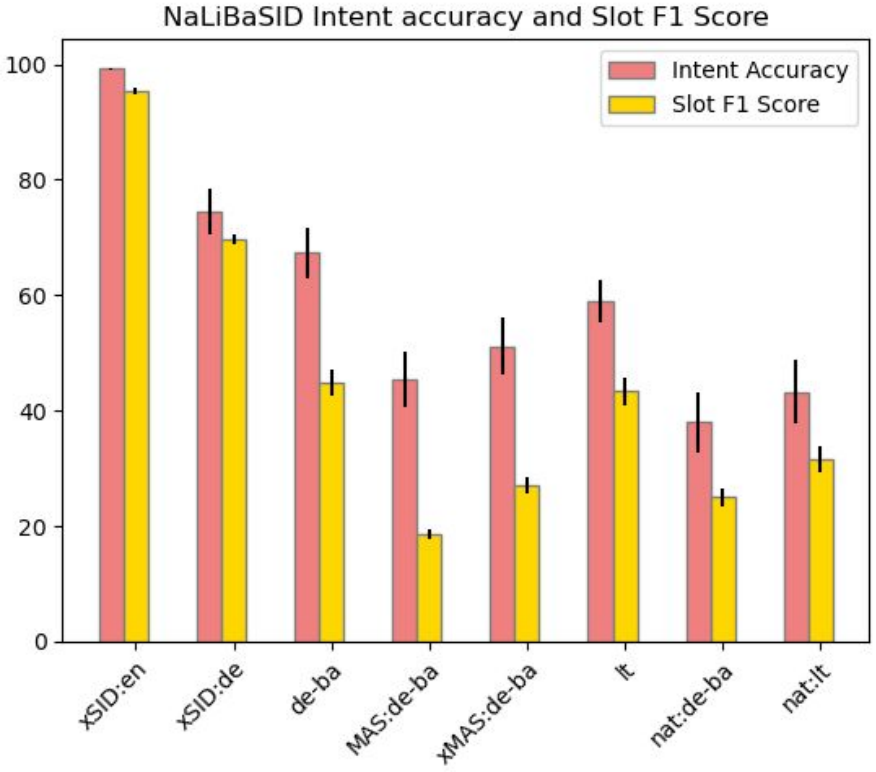
Results - Intent Accuracy



Lower scores on natural data than on translated data

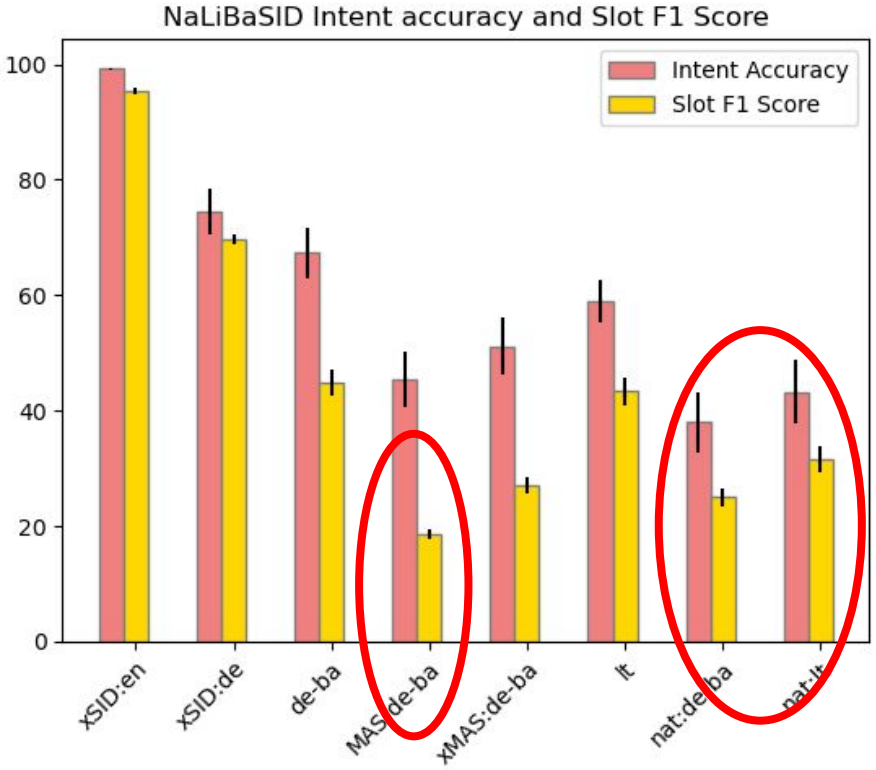
→ Impact of ‘translationese’ and cross-dataset setup

Results - Slot-F1



Similar trends as intent accuracy for Slot F1 scores

Results - Slot-F1



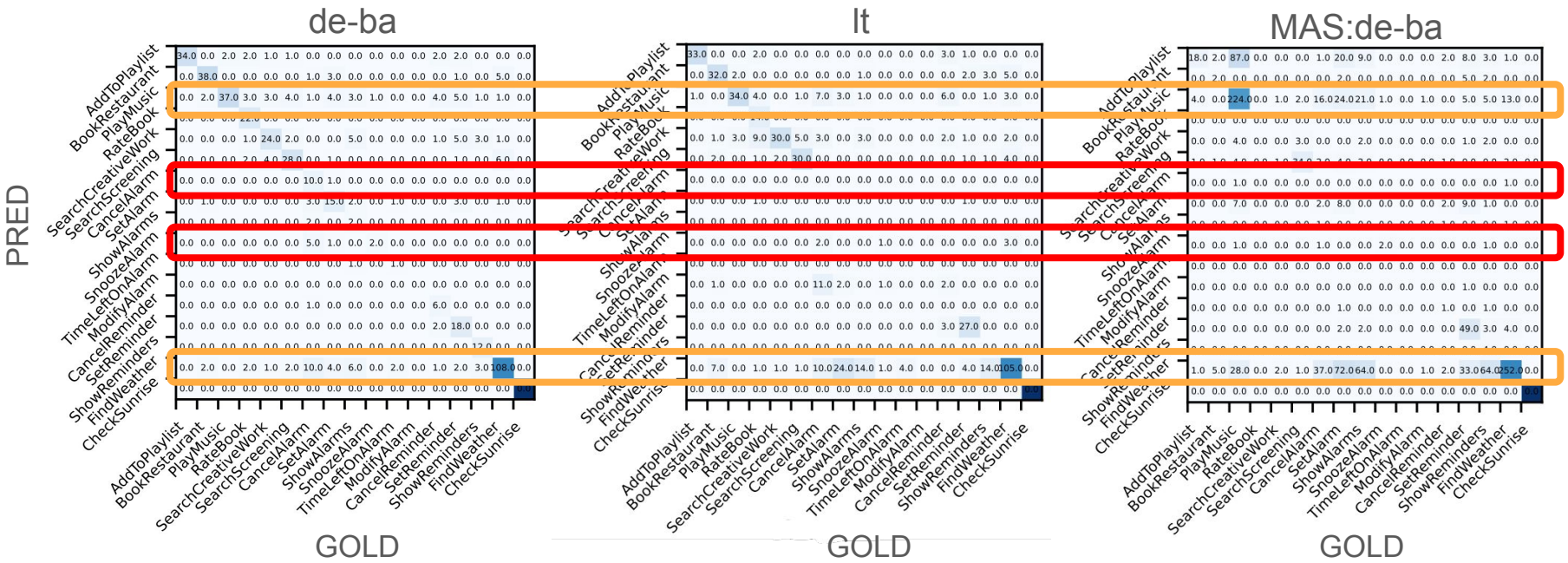
Natural datasets produce better results than MASSIVE translations

→ Cross-dataset experiments are challenging

Analyses - Intent Confusion Matrices

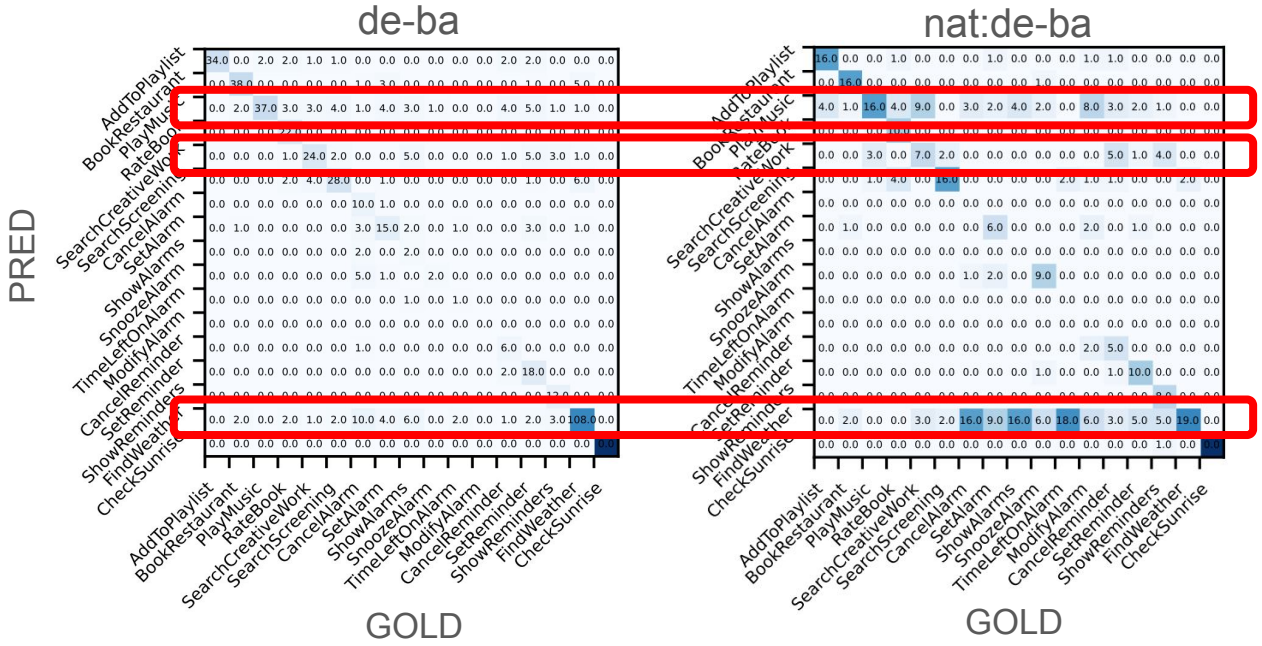
difficult intents: ‘snooze_alarm’, ‘cancel_alarm’

overpredicted: ‘PlayMusic’, ‘weather/find’



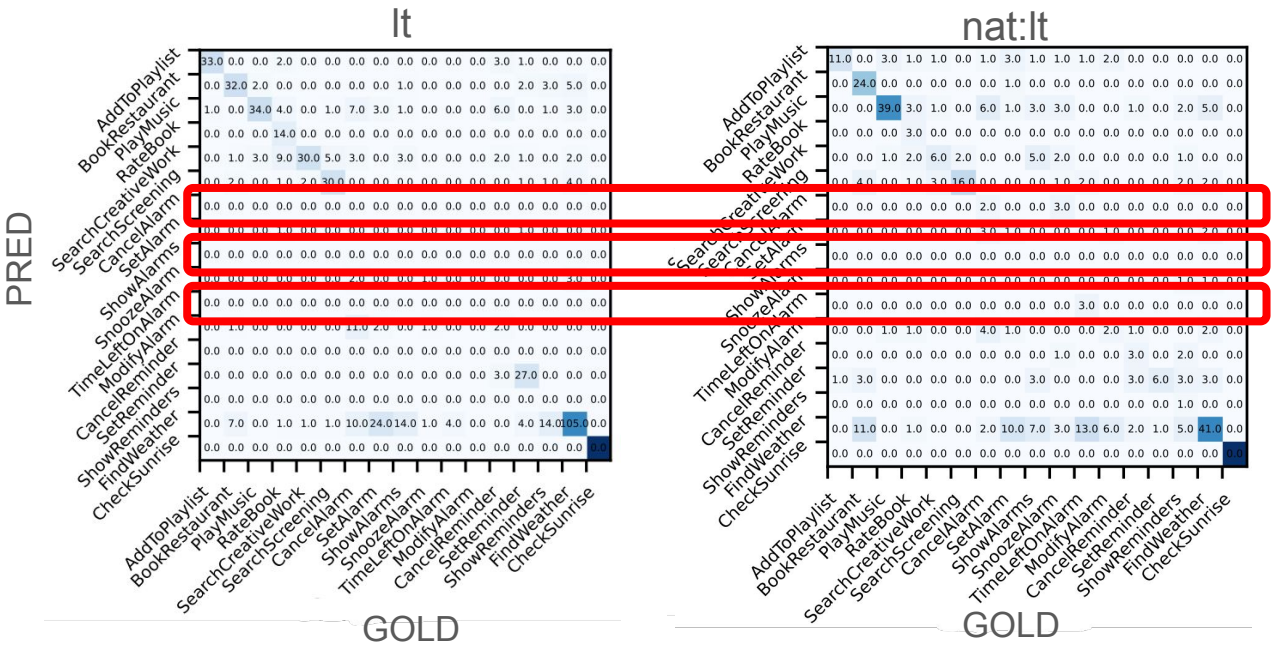
Analyses - Intent Confusion Matrices

similar error types for natural data

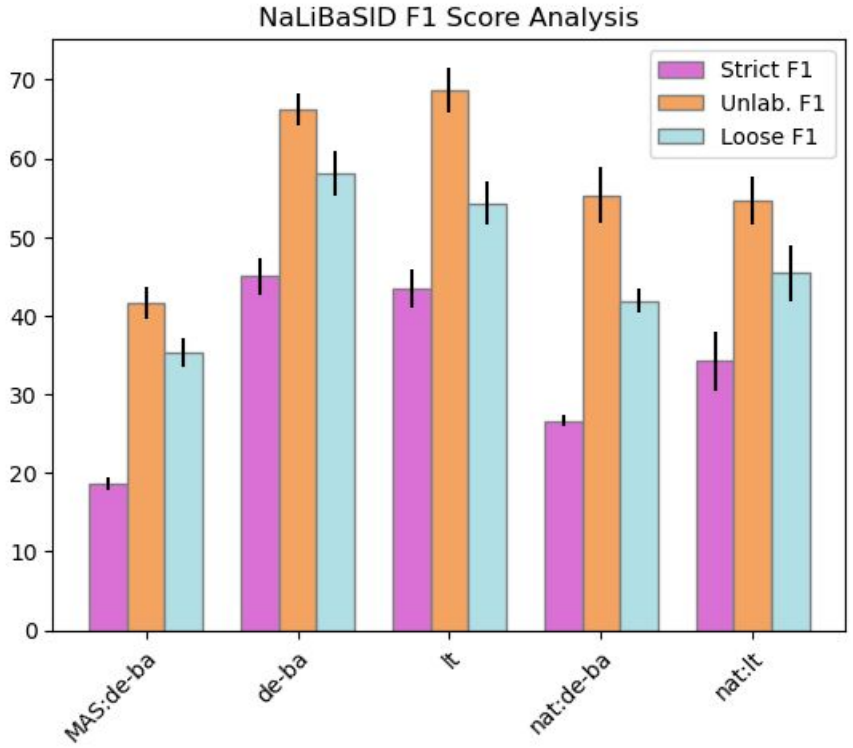


Analyses - Intent Confusion Matrices

similar error types for natural data



Analyses - Slot F1



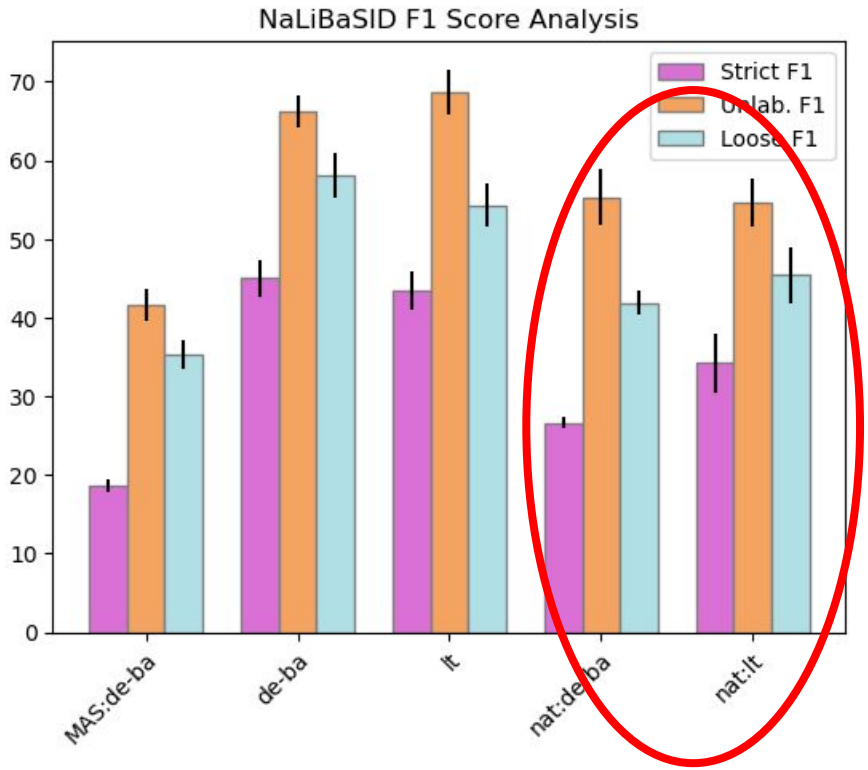
Unlab. F1:

ability to identify slots regardless of correct labeling

Loose F1:

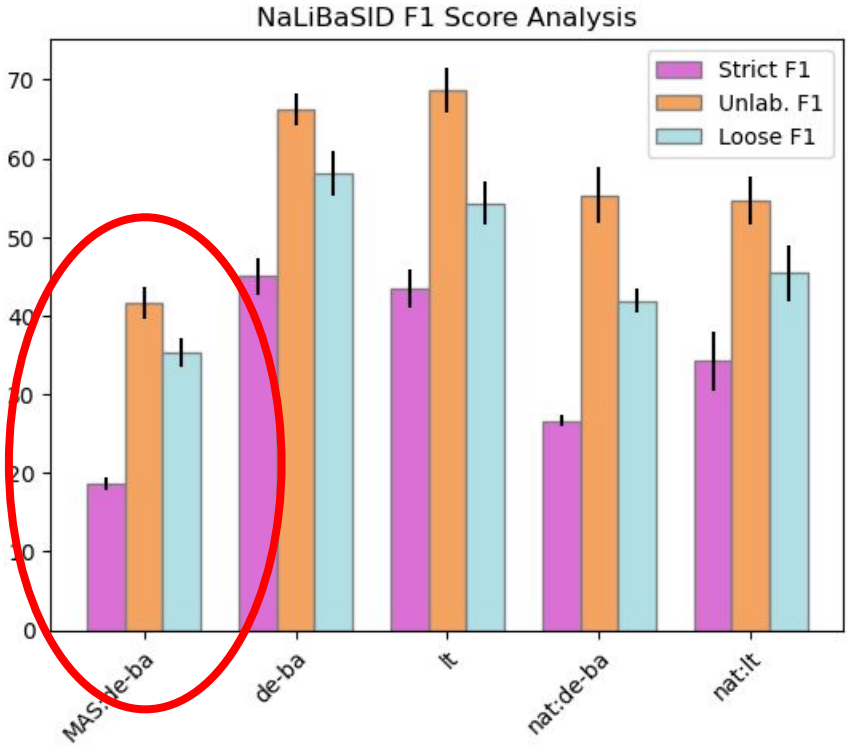
checks whether correctly labeled slots have correct boundaries

Analyses - Slot F1



mostly only partially correct predictions on natural sentences

Analyses - Slot F1



More difficult to find correct label in cross-dataset experiments than the exact slot span

Conclusions

- Translated datasets can lead to overoptimistic performance estimates
- The gathering method of data has an impact on model performance
  → Focus on cross-dataset experiments

Our contribution for these challenges:

- Data in two low-resource languages translated from a cross-lingual benchmark (xSID)
- Data for cross-dataset evaluation (MASSIVE)
- Natural data generated by native speakers
- Analysis of SID models on the data