Genre as Weak Supervision for Cross-lingual Dependency Parsing

Max **Müller-Eberstein**, Rob **van der Goot** and Barbara **Plank**

PERSONADS.ME/X/EMNLP-2021

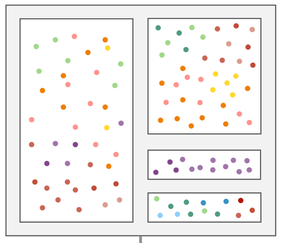
@mxmeij, @robvanderg and @barbara_plank

{mamy, robv, bapl}@itu.dk

MOTIVATION + SETUP

Universal Dependencies v2.7
Zeman et al., 2020

177 TREEBANKS 104 LANGUAGES
1.38M SENTENCES 18 GENRES



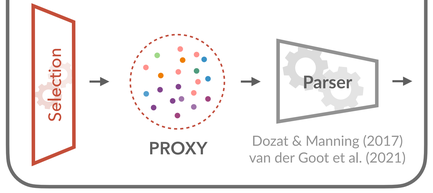
- **Cross-lingual proxy data selection** may benefit from incorporating textual **genre**
- Most commonly, **entire treebanks** containing relevant genres are used as proxy data
- Proxy data will be more effective if it only contains the most **genre-relevant instances**
- Instance genre labels are mostly unavailable

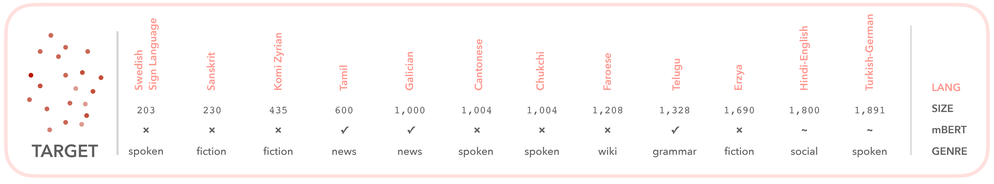
How do we select the best possible instances given only treebank-level genre metadata?

RESEARCH QUESTION

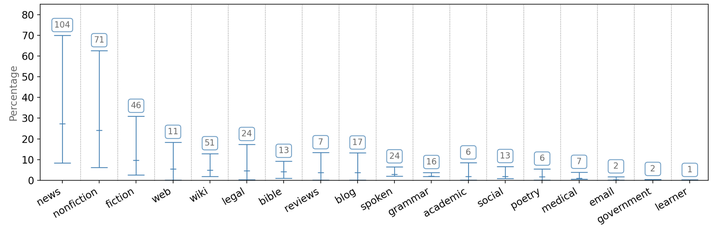
When selecting proxy training data for parsing...

Can genre guide the selection of cross-lingual proxy training data for a low-resource, zero-shot target?





ANALYSIS



Genre Distribution in UD. Upper/lower bounds for sentences per genre. Center marker reflects the distribution should genres within treebanks be uniformly distributed. Labels indicate the number of treebanks which contain each genre.

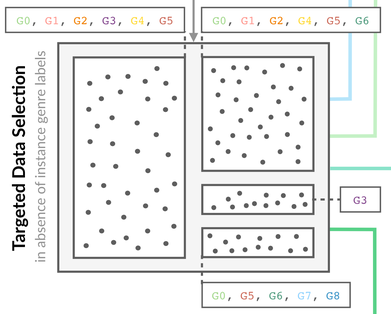
Genre in Universal Dependencies
as estimated from treebank metadata

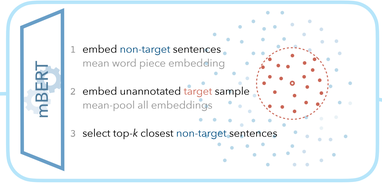
- *News* and *non-fiction* likely make up over half of the entire UD dataset
- Specialized genres (e.g. *spoken*, *social*, *medical*) are much less represented
- Some genres (e.g. *web*) only occur in mixture

Low-resource language/genre combinations benefit most from targeted instance selection.

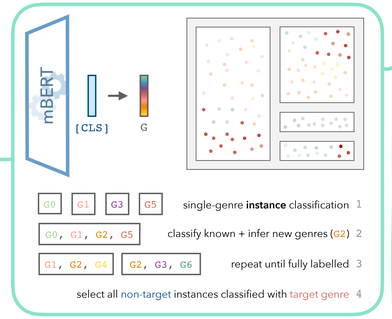
Targeted Data Selection

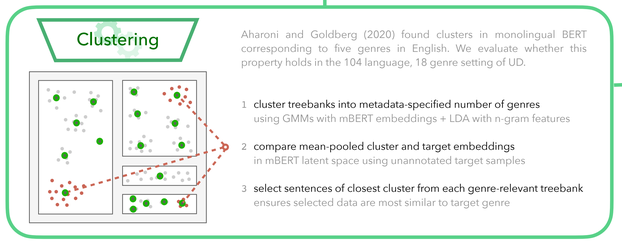
in absence of instance genre labels





select all treebanks containing the target genre

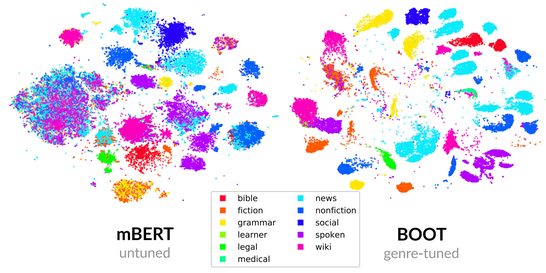




	SWL	SA	KPV	TA	GL	YUE	CKT	FO	TE	MYV	QHE	QTD	Ø	
TARGET	28.0	15.7	13.4	64.1	80.9	–	–	49.6	83.6	–	62.7	55.0	50.3	even with target data, these low-resource targets are difficult to parse
RAND	3.7	24.8	10.9	50.7	77.7	33.3	15.5	61.9	67.7	20.0	27.0	44.6	36.5	selects data according to treebank sizes
SENT	3.6	23.7	13.7	47.9	77.6	35.8	16.4	62.5	68.1	22.9	26.5	42.8	36.8	less targeted data selection in this multilingual setting
META	6.5	24.3	10.2	50.4	76.6	31.2	11.6	61.2	64.9	20.4	9.42	42.6	34.1	lowest overall performance despite selecting up to 8x more data
BOOT	5.2	21.8	*21.1	49.4	76.7	*49.9	18.4	*66.3	65.6	19.5	14.8	43.8	37.7	outperforms baselines despite not having access to any target data
GMM	4.9	22.9	*20.9	*51.5	77.8	*49.9	*19.8	*68.3	67.9	20.2	15.1	45.4	38.7	highest overall LAS
LDA	6.6	23.7	*22.3	49.2	77.0	*49.4	*19.1	*68.3	*68.6	20.5	15.1	44.7	38.7	better on non-mBERT

Zero-shot Parsing Results. LAS for test splits of target treebanks using training data from target/alternative in-language treebanks (**TARGET**; where available), random sentence selection (**RAND**), closest sentence selection (**SENT**), treebanks containing target genre (**META**), instances classified as target genre (**BOOT**) and closest cluster selection (**GMM** and **LDA**). Scores marked with * significantly outperform TARGET, RAND, SENT and META.

tSNE embedding plots show much clearer genre segmentation



Analysis of Selected Instances
BOOT, GMM and LDA identify data-driven genre

- **BOOT**, **GMM** and **LDA** perform best, while **META** performs worst
  - Using treebanks in bulk – as is currently common – is not fine-grained enough
  - **Genre-targeted instance selection** appears to be key
- Combining **higher-level genre information** with **latent information** from embeddings leads to performance increases not achievable by using either in isolation (i.e. META and SENT)
- Similar performance patterns and overlapping selections indicate similar, **data-driven notions of genre**
- Our proposed genre-driven methods significantly outperform prior work (van der Goot et al., 2021) using an identical parser architecture on 5/12 treebanks (i.e. SA, KPV, YUE, CKT, FO) **without annotated in-language data**

CONCLUSIONS

IT-UNIVERSITETET I KØBENHAVN / north