Multilingual and Zero-Shot is Closing in on Monolingual Web Register Classification

Samuel Rönnqvist* **Valtteri Skantsi*°** **Miika Oinonen*** **Veronika Laippala***

*TurkuNLP, University of Turku, Finland
°NSE, University of Oulu, Finland
{saanro,valtteri.skantsi,mhtoin,mavela}@utu.fi

Abstract

In this paper, we present experiments in register classification of documents from the unrestricted web, such as news articles or opinion blogs, in a multilingual setting, exploring both the benefit of training on multiple languages and the capabilities for zero-shot cross-lingual transfer. While the wide range of linguistic variation found on the web poses challenges for register classification, recent studies have shown that good levels of cross-lingual transfer from the extensive English CORE corpus to other languages can be achieved. In this study, we show that training on multiple languages 1) benefits languages with limited amounts of register-annotated data, 2) on average achieves performance on par with monolingual models, and 3) greatly improves upon previous zero-shot results in Finnish, French and Swedish. The best results are achieved with the multilingual XLM-R model. As data, we use the CORE corpus series featuring register annotated data from the unrestricted web.

1 Introduction

The focus of this paper is on multilingual training and cross-lingual transfer in register classification of web documents. Text register (or genre) (Biber, 1988), such as discussion forum or encyclopedia article, has been shown to be one of the most important predictors of linguistic variation (Biber, 2012), and register affects also the automatic processing of text (Mahajan et al., 2015; Webber, 2009; Van der Wees et al., 2018). Yet, web data is typically used without register information in many NLP tasks.

Web register classification studies have suffered from the lack of corpora featuring the full range of registers found on the web, as many datasets are based on a priori selection of register categories instead of unrestricted sampling of the web (Asheghi et al., 2016; Pritsos and Stamatatos, 2018). Furthermore, despite the availability of web-scale data in hundreds of languages, until recently, the resources for register identification have focused exclusively on English.

The data for this study consist of four similarly annotated online register collections featuring the CORE corpus series in English (Egbert et al., 2015), Finnish (Laippala et al., 2019), French and Swedish (Repo et al., 2021). All the datasets have been extracted from the unrestricted open web. While the English CORE is extensive, with 34k training examples, the other languages feature merely 2.7–4.6% of that (cf. Table 1).

In this paper, we explore how joint training on the four available CORE corpora can benefit register classification, with a particular interest in improving performance in smaller languages.[1] First, using multilingually pre-trained language models and a custom sampling and training strategy, we compare performance when training on all languages against previous monolingual results on the same corpora, observing gains for the smaller languages. Second, with the aim of creating a universal model fit for all languages, we train a multilingual master model that we evaluate in a zero-shot cross-lingual setting, demonstrating results that land within a relatively short distance from monolingual performances (4–6% F1-score for XLM-R).

2 Related work

Until recently, register identification from the unrestricted web has achieved only modest performance (Sharoff et al., 2010; Asheghi et al., 2014;

[1]For code and model, see: `https://github.com/TurkuNLP/multilingual-register-labeling`

Lang.	Train	Dev.	Test	Total
En	33,915	4,845	9,692	48,452
Fi	1,559	222	445	2,226
Fr	909	363	546	1,818
Sv	1,093	435	654	2,182

Table 1: Data set sizes in number of documents.

Biber and Egbert, 2016). Most importantly, the challenges are caused by the range of linguistic variation found on the web. Texts are written without gatekeepers, and not all registers are equally well-defined with discrete class boundaries (Biber and Egbert, 2018; Sharoff, 2018). To this end, Biber and Egbert (2018) suggest to extend the analysis to *hybrid* documents combining characteristics of several register classes, and Sharoff (2018, 2021) examines web genres by prototypical genre classes and text dimensions featuring communicative functions, such as argumentation or reporting.

Despite the difficulty, Laippala et al. (2019) show that multi- and cross-lingual modeling of registers between English and Finnish is possible at practical levels of performance, as they propose a convolutional neural network (CNN) model with multilingual word embeddings to model registers. Further, Repo et al. (2021) demonstrate that pretrained neural language models, especially XLM-R, can achieve strong performance monolingually on the four aforementioned languages, as well as achieve strong cross-lingual transfer in a zero-shot learning setting from English to other languages.

The benefits of combining several languages during training has been demonstrated for other NLP tasks. Training the multilingual XLM-RoBERTa (XLM-R), Conneau et al. (2020) showed that adding more languages to training leads to better cross-lingual performance on low-resource languages. Comparing the performance of multiple multilingual models across a number of tasks and languages, Hu et al. (2020) noted as well that adding target language data to training provides higher performance. However, they highlighted that a model's cross-lingual performance varies greatly between languages and tasks – on QA tasks, zero-shot models are very efficient and outperform models trained on 1,000 examples of target-language data. Finally, also the positive effect of sampling under- and overpresented languages has been demonstrated previously; in the context of multilingual semantic parsing, Li et al. (2020) perform up- and downsampling of languages based on frequency as part of their sampling strategy, in order to improve multilingual performance.

3 Data

The four datasets we use in this study—CORE, FinCORE, FreCORE and SweCORE—all feature the unrestricted web, however, they have been compiled in different ways. The English CORE is based on unrestricted search queries of extremely frequent n-grams, while the other datasets are randomly sampled from the 2017 CoNLL Shared Task datasets, originally drawn from Common Crawl (Ginter et al., 2017). Table 1 summarizes the data set sizes.

The four datasets have all manual register annotations following the same register taxonomy that was developed during the compilation of the English CORE. The taxonomy is hierarchical, with eight main registers and approximately 30 subclasses, depending on the language-specific version. In this study, we focus on the main register level, which includes the classes *Narrative* (NA), *Informational Description* (IN), *Opinion* (OP), *Interactive Discussion* (ID), *How-to/Instruction* (HI), *Informational Persuasion* (IP), *Lyrical* (LY) and *Spoken* (SP) (for a detailed description, see (Biber and Egbert, 2018)).

In order to reflect the variation found within the data, *hybrid* documents combining characteristics of several registers are also annotated. On the main register level, these display 11–15% of all other language-specific datasets but Finnish. Perhaps because of the different approaches to gathering the corpora, the register distributions differ also for some other classes between CORE and the others. Specifically, the Informational Persuasion class covers only 2.75% of CORE, and 16.82–24.15% of the other datasets, and also the Opinion class covers 16.23% of CORE and 15,23% of FinCORE, but only 6.63% of FreCORE and 6,60% of SweCORE (for details, see Repo et al. (2021)).

4 Methods

4.1 Multilingual language models

We focus on two multilingual deep learning models, namely Multilingual BERT (mBERT, (Devlin et al., 2019)) and XLM-RoBERTa (XLM-R, (Conneau et al., 2020)), which have been shown

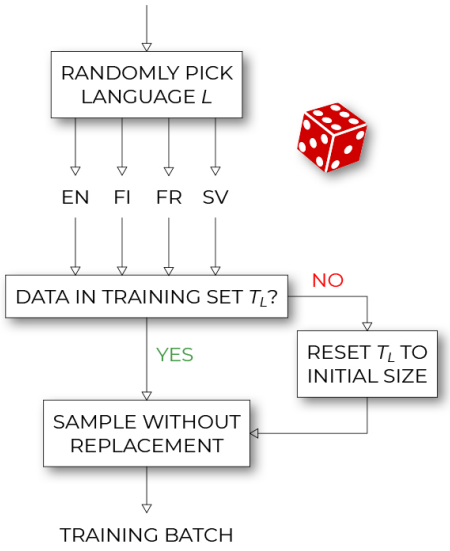

Figure 1: Illustration of the multilingual sampling strategy. Languages are uniformly sampled to generate training batches. A training set is independently reshuffled after a full pass.

to achieve high performance in both monolingual and zero-shot cross-lingual settings of register classification. Repo et al. (2021) show that XLM-R clearly outperforms mBERT by up to 8% points F1-score monolingually and up to 11% points cross-lingually, while both clearly outperform previous state-of-the-art.

Both mBERT and XLM-R are based on the BERT architecture, the first being trained on Wikipedia in 104 languages and the latter on cleaned Common Crawl data in 100 languages. While both models lack an explicit cross-lingual signal, XLM-R has more than double the vocabulary size and was trained on significantly more data for a longer time. We use the large version of XLM-R, whereas mBERT is only available in base size. In various multilingual tasks, XLM-R has been shown to outperform mBERT, which tends to struggle especially with smaller languages such as Finnish and Swedish (Rönnqvist et al., 2019). Nevertheless, we include both models in order to study their relative performances as we introduce a multilingual sampling strategy.

The experiments are performed as multi-label classification in order to support hybrid registers. We use TensorFlow checkpoints of the models through the Huggingface Transformers library and repository (Wolf et al., 2020). We train a decision layer on top of the top-layer CLS embedding, while also fine-tuning the language model parameters, with a binary cross-entropy loss. The models are evaluated using micro-averaged F1-score and a fixed prediction threshold of 0.5.

4.2 Sampling and training strategy

Since the training sets in the different language corpora we use differ, they risk skewing the class distributions when training on multiple languages at once. In particular, the English set is much larger than the others, and exhibits a somewhat different class distribution (see Section 3, Repo et al. (2021)).

In order to mitigate this problem, we propose a sampling strategy which samples all languages in equal parts during training. The strategy is illustrated in Figure 1. First, for each mini-batch, the language is selected with uniform probability, and then training samples are randomly sampled without replacement. The examples in a language set are reshuffled when they have all been sampled, such that the smaller sets are repeated more often. One training epoch consists of $N \cdot B_1$ mini-batches, where N is the number of languages and B_1 the number of mini-batches in the smallest training set.

In combination with this mode of sampling, we train the models for longer than reported by Repo et al. (2021), typically on the order of 100 epochs, in order to avoid explicitly disregarding any data in the larger training sets. We apply an early stopping criterion on the validation set F1-score, in order to avoid excessive training and to empirically determine when the data sets have been sufficiently repeated. We also use a learning rate about an order of magnitude lower than in the previously reported work to match the longer training.

5 Experiments

We first train models jointly on all four languages following the sampling strategy introduced above, and optimize hyperparameters[2] for each target language separately, based on development set performance. The optimal model for each language is tested on the respective test set. We compare the multilingual results to the previous state-of-the-art results in monolingual settings, i.e., where one and

[2]We test learning rates in the range $4e^{-6}$ to $7e^{-5}$ and maximum number of epochs 25 to 175 (affecting rate of warm-up and learning rate decay). Batch size is 7 (capped by available GPU memory) and patience 5 epochs.

	Monolingual (baseline)				Multilingual (ours)				
mBERT	Dev.		Test		Dev.		Test		Test diff.
Target	F1 (%)	Std.	F1 (%)	Std.	F1 (%)	Std.	F1 (%)	Std.	F1 (%)
En	72.80	(0.21)	73.06	(0.09)	68.20	(1.36)	68.63	(1.39)	-4.43
Fi	65.91	(0.85)	64.83	(1.16)	69.25	(1.75)	65.95	(1.06)	1.12
Fr	70.74	(1.67)	68.66	(0.63)	72.49	(0.54)	69.55	(0.36)	0.89
Sv	76.91	(0.45)	76.43	(0.46)	78.49	(0.85)	78.22	(1.17)	1.79
Average			70.75				70.59		-0.16
excl. En			69.97				71.24		0.91
XLM-R									
Target	F1 (%)	Std.	F1 (%)	Std.	F1 (%)	Std.	F1 (%)	Std.	
En	75.80	(0.12)	75.68	(0.05)	72.03	(0.89)	72.43	(0.48)	-3.25
Fi	76.25	(0.45)	73.18	(1.35)	77.53	(0.94)	75.00	(0.53)	1.82
Fr	77.38	(0.51)	76.92	(0.24)	78.72	(0.49)	77.54	(0.99)	0.62
Sv	82.61	(0.37)	83.04	(0.62)	83.92	(0.34)	83.92	(0.34)	0.90
Average			**77.21**				**77.22**		0.01
excl. En			**77.71**				**78.82**		0.83

Table 2: Performance of models trained in monolingual and multilingual settings, optimized for each language separately. F1-scores are means, N=3.

	Multilingual master model			
mBERT	Common dev.		Test	
Target	F1 (%)	Std.	F1 (%)	Std.
En	71.32	(1.51)	66.27	(2.33)
Fi			65.27	(1.56)
Fr			69.76	(2.24)
Sv			77.92	(1.21)
Average			69.81	
excl. En			70.98	
XLM-R	Common dev.		Test	
Target	F1 (%)	Std.	F1 (%)	Std.
En	78.20	(0.04)	72.37	(1.17)
Fi			75.05	(0.81)
Fr			78.81	(0.89)
Sv			82.36	(0.54)
Average			77.15	
excl. En			78.74	

Table 3: Performance of models validated against a common development set that is balanced between the languages, and tested on the language-specific test sets. F1-scores are means, N=3.

	Zero-shot, from English (baseline)		**Zero-shot, multilingual (ours)**	
mBERT	Test		Test	
Target	F1 (%)	Std.	F1 (%)	Std.
En	–	–	55.15	(2.58)
Fi	50.21	(0.74)	58.46	(0.76)
Fr	55.04	(0.66)	62.82	(1.86)
Sv	62.53	(0.78)	69.48	(0.72)
Average	–		61.48	
excl. En	55.93		63.59	
XLM-R				
Target	F1 (%)	Std.	F1 (%)	Std.
En	–	–	**63.32**	(0.25)
Fi	61.35	(1.26)	**69.60**	(0.55)
Fr	64.27	(1.58)	**72.85**	(1.74)
Sv	69.22	(1.66)	**79.49**	(0.95)
Average	–		**71.31**	
excl. En	64.95		**73.98**	

Table 4: Performance of models trained in zero-shot cross-lingual settings, from English to target language (left), and from all other languages to target (right). F1-scores are means, N=3.

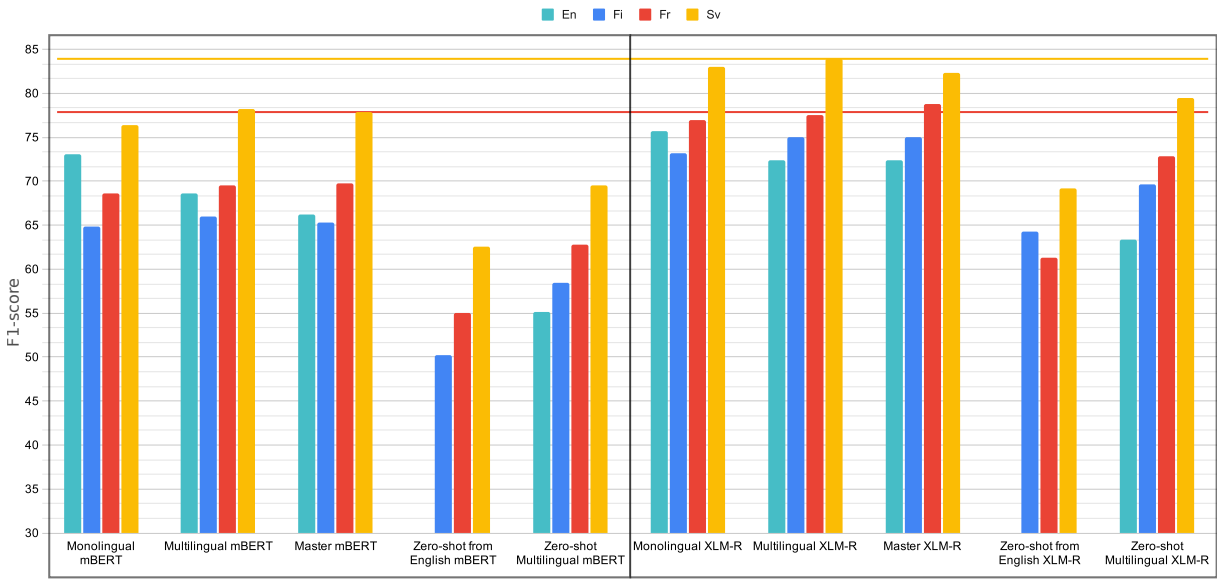

Figure 2: Comparison of all F1-scores. The left box presents the performance of mBERT in the different settings (bar groups) for all languages (color coded) and the right box presents those of XLM-R. Inter-annotator agreement levels (horizontal colored lines) for French and Swedish provide points of reference indicating potential upper bounds for modeling.

the same language is used to train, validate and test the models.

Table 2 presents the results of these experiments (right hand side), as well as the monolingual baseline performances reported by Repo et al. (2021) (left hand side). We observe that both mBERT (above) and XLM-R (below) perform better in multilingual training for all languages except for English. The gains are on average (excluding English) 0.8–0.9% F1-score for the two models, indicating some degree of cross-lingual knowledge transfer from the extra data. Meanwhile, performance for English drops by 3.3–4.4% points, which is likely due to the class distribution being pushed to its disadvantage by the uniform sampling of the otherwise more homogeneous corpora. In terms of average F1-score, the multilingual performance is on par with the previous monolingual models.

Second, after optimizing on each language individually, we perform another hyperparameter search for training a single multilingual model that should favor each language equally, which we call a *master model*. In order to train the master model, we create a common development set based on the individual sets of the languages. The development sets differ in size due to different sizes of the corpora and different data split ratios (see Table 1). We create the common set by upsampling the Finnish and French and downsampling the English set to the size of the Swedish set; the sets are then concatenated to a total size of 1740. The master model is validated against this set during training. In particular, when to stop training is determined based on the performance on this set, i.e., on the average performance across languages.

Table 3 lists the best performance on the common development set for both mBERT and XLM-R, as well as the performance of both models in each language-specific test set. The level of performance remains stable for the master model, with an average decrease of 0.78% for mBERT and only 0.08% for XLM-R compared to the multilingual results in Table 2.

Third, in order to estimate the performance that can be expected of the master model on an unseen language, we still perform an experiment where each of our four languages is in turn taken as target, and a model is trained with the previously optimized hyperparameters, using the remaining three languages for training and validation (controlling early stopping). The models are tested in each language separately.

The results of this zero-shot cross-lingual experiment are listed in Table 4 (right hand side), along baseline results from previous work studying cross-lingual transfer from English to the other languages (Repo et al., 2021) (left hand side).

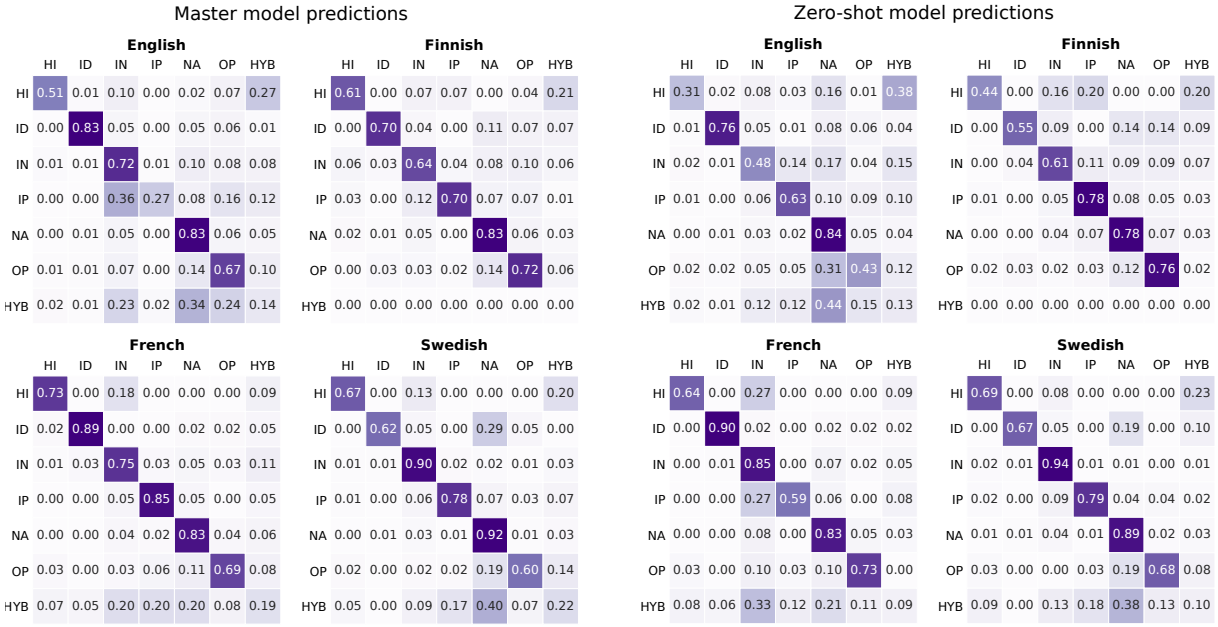

Figure 3: Confusion matrices for predictions in each language using the master multilingual model (left) and the zero-shot cross-lingual models (right). Columns represent predictions and rows true labels for the most common classes, with all hybrid instances represented by HYB only.

The numbers show a significant gain for multilingual modeling over cross-lingual modeling from English only; for mBERT the increase is 5.55% (7.66% excluding English) and for XLM-R 6.36% (9.03%).

Finally, Figure 2 summarizes the F1-scores from the aforementioned tables in a side-by-side comparison. We especially observe how the zero-shot multilingual results take the lead over the baseline of zero-shot from English, in order to approach the levels of the monolingual and multilingual models for which target language is also used for training. The levels of inter-annotator agreement, as reported by Repo et al. (2021), were counted prior to any discussions between the annotators. Although this level should be considered as a lower bound of human agreement, it sets a theoretical boundary for automatic register identification.

6 Error analysis

In order to gain a more detailed understanding of the types of errors the models are making, we study the confusion matrices in Figure 3. These present the correct classifications (diagonal) and misclassifications (rest), both in a single language setting using the master multilingual model and in a cross-lingual setting using the zero-shot model. The matrices include the six most frequent classes and a separate hybrid class, as the confusions matrix is not defined for the multi-label setting.

We observe that hybrid documents overall are difficult to recognize as such, in particular hybrids composed of Narrative (NA) and another class are often predicted as NA only. Comparing the master model (left) and the zero-shot models (right), we see that the overall patterns are quite similar, while the cross-lingual performance, for instance, in English and Finnish is worse for How-to/Instruction (HI) and Interactive discussion (ID). In Swedish, however, ID performs better cross-lingually, and Swedish generally exhibits the smallest differences between the settings.

Informational description (IN) and Informational persuasion (IP) are difficult to distinguish in English for the master model, whereas the cross-lingual model handles these classes much better, although there is still room for improvement. Distinguishing purely informational texts and those with an intent to persuade is difficult for other zero-shot models as well.

Comparing the cross-lingual matrices with those reported by Repo et al. (2021) for transfer from English to the other languages, we note that our diagonals are significantly crisper, i.e., the classes more frequently correctly predicted. In

their results, especially the classes HI and IP are generally more dispersed, as well as Opinion (OP) for French, NA for Swedish and IN for Finnish (vertically, i.e., other classes are mistaken for IN).

Finally, comparing class-wise F1-score between the master and zero-shot models we observe a 3.1% mean decrease for NA (sd. 1.5%), 5.9% for OP (sd. 2.6%) and 7.6% for IP (sd. 5.4%). Most of the classes are too infrequent in our data for meaningful interpretation of class-wise differences, or the patterns are inconsistent across languages.

7 Discussion

Our results show that multilingual training brings clear advantages to web register identification, in particular for the languages with small amounts of training examples. When allowing training on target data, performance is somewhat improved for these languages, while it remains on par in average. In the zero-shot setting, however, the performance is greatly improved compared to the recent and already strong state-of-the-art results. As illustrated in Figure 2, the multilingual zero-shot XLM-R is closing in on its top-performing counterparts trained monolingually or on all languages.

The fact that the multilingual performance on English is lagging behind is expected, as its class distribution differs notably from that of the other languages, and the uniform sampling is designed to allow the model to learn a mean distribution across the languages. In the zero-shot experiments, the English-targeted model will see relatively little data compared to the other models, which likely works to its disadvantage. In the context of pre-trained language models, English monolingual models are also known to be high-performing; similar results on a multilingual model outperforming other monolingual models but not English have been reported by Hu et al. (2020).

To test how the multilingual model performs in a zero-shot setting, we experimented with a leave-one-out version of the multilingual setting, where a model was trained on all except for the target language data on which the model was tested. Although the results were, as expected, lower than the monolingual and multilingual results where target language was included in training, the gap is closing quickly. With the baseline methods, the average gap between the cross-lingual models and monolingual models has been 12.76% points F1-score—in our study, it is 3.73% excluding English, 5.9% including English (with XLM-R).

With an average F1-score of 73.98% for Finnish, French and Swedish, we demonstrate that applying this multilingual register classification model in zero-shot settings can be done at very practical levels of performance. This indicates that our multilingual model can be applied without significant loss of accuracy on languages without existing register-annotated corpora, which is an important step toward being able to perform register identification on the truly unrestricted web, also in terms of language.

In particular, these performances are competitive considering the difficulty of the task. As discussed above, the inter-annotator agreements of 78% for French and 84% for Swedish serve as a potential upper bound in modeling. The monolingual models are already very close to this level, and the multilingual zero-shot models are not far.

The competitiveness of multilingual training is particularly interesting in the case of registers. Although the advantages of this multilingual training have been noted before (see Section 2), it is not evident that register identification can benefit from it. Registers are specific to the situation and to the culture where they have been produced. For instance, Opinion blogs can express their points of view differently depending on cultural context, and the level of formality of Speeches and News reports (subregisters of Spoken and Narrative) may vary according to the culture. Also the linguistic means to express functional characteristics associated with registers, such as narration or interaction, differ across languages. These differences can have a drastic effect on the success of the modeling even if the transfer itself works. In the current study, the included languages are all European, which makes also the transfer easier, whereas including more languages and more distant cultures remains a research desideratum.

8 Conclusion

To sum up, our study corroborates the power of multilingual training when modeling registers in languages with a limited amount of training data. We train and make available a multilingual master model for register classification, whose performance is competitive with existing monolingual models. Its zero-shot performance is approaching that of monolingual models, as it improves upon

already strong state-of-the-art results. Considering the estimated level of human agreement on the task, the margin for further improvement is relatively slim. Nevertheless, it is our goal to continue this work in order to achieve robust zero-shot performance in a wide range of languages up to the level of monolingual models. Furthermore, it would be interesting to test the robustness and generalizability of our models by evaluating them against the prototypical web genre categories and Function Text Dimensions presented in (Sharoff, 2018, 2021).

Finally, in the future, we will also investigate register-specific differences in their transfer. Registers differ in terms of how well they are linguistically defined, which naturally also affects their identification (Laippala et al., 2021). For instance, while the linguistic characteristics of many blogs can vary extensively, those of encyclopedia articles remain very similar across texts. This tendency concerns also the cross-lingual similarities of registers, and similarities have already been discovered in particular in the spoken register.

Acknowledgements

We thank the Emil Aaltonen Foundation and Academy of Finland for financial support. We also wish to acknowledge CSC – IT Center for Science, Finland, and the NVIDIA Corporation GPU Grant Program, for computational resources.

References

Noushin Rezapour Asheghi, Serge Sharoff, and Katja Markert. 2016. Crowdsourcing for web genre annotation. *Language Resources and Evaluation*, 50(3):603–641.

Rezapour Noushin Asheghi, Katja Markert, and Serge Sharoff. 2014. Semi-supervised graph-based genre classification for web pages. In *Proceedings of TextGraphs-9*, pages 39–47. Association for Computational Linguistics.

Douglas Biber. 1988. *Variation across speech and writing*. Cambridge University Press.

Douglas Biber. 2012. Register as a predictor of linguistic variation. *Corpus linguistics and linguistic theory*, 8(1):9–37.

Douglas Biber and Jesse Egbert. 2016. Using grammatical features for automatic register identification in an unrestricted corpus of documents from the open web. *Journal of Research Design and Statistics in Linguistics and Communication Science*, 2:3–36.

Douglas Biber and Jesse Egbert. 2018. *Register variation online*. Cambridge University Press.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. https://doi.org/10.18653/v1/2020.acl-main.747 Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Jesse Egbert, Douglas Biber, and Mark Davies. 2015. Developing a bottom-up, user-based method of web register classification. *Journal of the Association for Information Science and Technology*, 66:1817–1831.

Filip Ginter, Jan Hajič, Juhani Luotolahti, Milan Straka, and Daniel Zeman. 2017. http://hdl.handle.net/11234/1-1989 CoNLL 2017 shared task - Automatically annotated raw texts and word embeddings. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. http://arxiv.org/abs/2003.11080 Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization.

Veronika Laippala, Jesse Egbert, Douglas Biber, and Aki-Juhani Kyröläinen. 2021. Exploring the role of lexis and grammar for the stable identification of register in an unrestricted corpus of web documents. *Lang Resources Evaluation*.

Veronika Laippala, Roosa Kyllönen, Jesse Egbert, Douglas Biber, and Sampo Pyysalo. 2019. https://www.aclweb.org/anthology/W19-6130 Toward multilingual identification of online registers. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 292–297. Linköping University Electronic Press.

Haoran Li, Abhinav Arora, Shuohui Chen, Anchit Gupta, Sonal Gupta, and Yashar Mehdad. 2020. http://arxiv.org/abs/2008.09335 Mtop: A comprehensive multilingual task-oriented semantic parsing benchmark.

Anuj Mahajan, Sharmistha Jat, and Shourya Roy. 2015. https://doi.org/10.18653/v1/K15-1034 Feature selection for short text classification using wavelet packet transform. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 321–326. Association for Computational Linguistics.

Dimitrios Pritsos and Efstathios Stamatatos. 2018. https://doi.org/10.1007/s10579-018-9418-y Open set evaluation of web genre identification. *Language Resources and Evaluation*, 52(4):949–968.

Liina Repo, Valtteri Skantsi, Samuel Rönnqvist, Saara Hellström, Miika Oinonen, Anna Salmela, Douglas Biber, Jesse Egbert, Sampo Pyysalo, and Veronika Laippala. 2021. Beyond the english web: Zero-shot cross-lingual and lightweight monolingual classification of registers. In *Proceedings of the EACL 2021 Student Research Workshop*.

Samuel Rönnqvist, Jenna Kanerva, Tapio Salakoski, and Filip Ginter. 2019. Is multilingual BERT fluent in language generation? In *Proceedings of the 1st NLPL Workshop on Deep Learning for Natural Language Processing*.

Serge Sharoff. 2018. Functional text dimensions for the annotation of web corpora. *Corpora*, 1(13):65–95.

Serge Sharoff. 2021. Genre annotation for the web: text-external and text-internal perspectives. *Register Studies*.

Serge Sharoff, Zhili Wu, and Katja Markert. 2010. The web library of babel: evaluating genre collections. In *Proceedings of LREC)*.

Bonnie Webber. 2009. Genre distinctions for discourse in the Penn TreeBank. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 674–682. Association for Computational Linguistics.

Marlies Van der Wees, Arianna Bisazza, and Christof Monz. 2018. Evaluation of machine translation performance across multiple genres and languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.