Toward Multilingual Identification of Online Registers

Veronika Laippala[1], **Roosa Kyllönen**[1], **Jesse Egbert**[2], **Douglas Biber**[2], **Sampo Pyysalo**[3]

[1] School of Languages and Translation Studies, University of Turku
[2] Applied Linguistics, Northern Arizona University
[3] Department of Future Technologies, University of Turku
[1,3] first.last@utu.fi, [2] first.last@nau.edu

Abstract

We consider cross- and multilingual text classification approaches to the identification of online registers (genres), i.e. text varieties with specific situational characteristics. Register is arguably the most important predictor of linguistic variation, and register information could improve the potential of online data for many applications. We introduce the Finnish Corpus of Online REgisters (FinCORE), the first manually annotated non-English corpus of online registers featuring the full range of linguistic variation found online. The data set consists of 2,237 Finnish documents and follows the register taxonomy developed for the Corpus of Online Registers of English (CORE), the largest manually annotated language collection of online registers. Using CORE and FinCORE data, we demonstrate the feasibility of cross-lingual register identification using a simple approach based on convolutional neural networks and multilingual word embeddings. We further find that register identification results can be improved through multilingual training even when a substantial number of annotations is available in the target language.

1 Introduction

The massive amount of text available online in dozens of languages has created great opportunities for Natural Language Processing (NLP). For instance, methods such as machine translation, automatic syntactic analysis and text generation have benefited from the large-scale data available online (Tiedemann et al., 2016; Zeman et al., 2018; Devlin et al., 2018).

However, the diversity of online data is also a challenge to its use. Documents have little or no information on their communicative purpose or, specifically, on their *register* (*genre*) (Biber, 1988). Register – whether a document is a blog, how-to-page or advertisement – is one of the most important predictors of linguistic variation and affects how we interpret the text (Biber, 2012). Automatic identification of registers could thus improve the potential of online data, in particular for linguistically oriented research (Webber, 2009; Giesbrecht and Evert, 2009).

However, the automatic identification of registers has proven to be difficult. Studies of Web Genre Identification (WGI) have been limited by small and scattered data sets which have resulted in lack of robustness and generalization of the models (Sharoff et al., 2010; Petrenz and Webber, 2011; Pritsos and Stamatatos, 2018; Asheghi et al., 2014). Furthermore, although online data is available in many languages and NLP systems are increasingly focused on multilingual settings (e.g., Zeman et al. (2018)), WGI studies have focused nearly exclusively on English texts. The only large-scale data set representing the full range of online registers is CORE — the Corpus of Online Registers of English — which is based on an unrestricted sample of English documents from the searchable web (Egbert et al., 2015).

In this paper, we extend the scope of modeling online registers to cross- and multilingual settings. We 1) present the first non-English data set of online registers with manual annotations, 2) show that it is possible to identify online registers in a cross-lingual setting, training only on English data while predicting registers also in Finnish, and 3) demonstrate that multilingual training can improve register identification performance even when a substantial number of target language annotations are available. Our approach is based on convolutional neural networks (Kim, 2014) and multilingual word embeddings (Conneau et al., 2018).

Proceedings of the 22nd Nordic Conference on Computational Linguistics (NoDaLiDa), pages 292–297
Turku, Finland, 30 September – 2 October,

2 Previous Work

In WGI, reported performance is often very high due to small and skewed corpora. With six widely used online register corpora composed of 7-70 classes, the best accuracy achieved by Sharoff et al. (2010) was 97% with character n-grams. Similarly, Pritsos and Stamatatos (2018) achieved an F1-score of 79% using two of the same corpora. However, the authors noted that their classifier models identified specific corpus topics rather than generalizable register features. This was further confirmed by Petrenz and Webber (2011) who showed that the applied system performances dropped drastically when the topic distribution of the target data was changed after training.

Using the larger Leeds Web Genre (LWG) corpus (Asheghi et al., 2016) of 3,964 documents, Asheghi et al. (2014) showed that online registers can be identified in a representative collection. Their best accuracy was 78.9% on 15 classes based on plain texts and 90.1% based on a semi-supervised graph-based method. However, as the LWG corpus represents only registers exclusive to the web and is compiled by manually selecting the texts, it does not feature the full range of linguistic variation online. By contrast to LWG, CORE (Egbert et al., 2015) is based on an unrestricted sample of the web. Biber and Egbert (2016) evaluated automatic CORE register detection performance with stepwise discriminant analysis, achieving 34% precision and 40% recall.

In previous studies, crosslingual models have been developed using various methods. Andrade et al. (2015), Shi et al. (2010) and Lambert (2015) applied bilingual dictionaries and machine translation to generate target language models in crosslingual topic detection and sentiment analysis. Many recent neural approaches use multilingual embeddings to build the document representations. Approaches such as that of Klementiev et al. (2012) are based on either the combination of multilingual word embeddings or directly learned sentence embeddings. Schwenk and Li (2018) compared their performance in genre classification of a multilingual Reuters corpus, using word embeddings generated by Ammar et al. (2016) and combined to document representations using a one-layer convolutional network and an LSTM-based system as proposed by Schwenk and Douze (2017), finding out that the system based on word embeddings achieved the best performance.

Register	English		Finnish	
Narrative	12,541	(50%)	778	(35%)
Opinion	3,960	(16%)	339	(15%)
D-Informational	3,195	(13%)	379	(17%)
Discussion	2,697	(11%)	140	(6%)
How-to	955	(4%)	144	(7%)
Info-Persuasion	684	(3%)	446	(20%)
Lyrical	576	(2%)	0	(0%)
Spoken	304	(1%)	11	(0%)
Total	24,912		2,237	

Table 1: The sizes of the register classes in the two data sets. The proportions of the classes are given in parentheses.

3 Data

The data for our study come from two sources. The English CORE consists of 48,571 documents coded by four annotators, who used a taxonomy developed in a data-driven manner to cover the full range of linguistic variation found in the Internet. The taxonomy is hierarchical and consists of eight main registers divided into 33 sub-registers. The *Narrative* main register includes sub-registers such as News, Short stories and Personal blogs. The *Opinion* main register consists of texts expressing opinions, such as Opinion blogs and Reviews. *Informational description (D-Informational)* covers informational registers such as Descriptions of a thing and Research articles. The *Discussion* class includes various discussions such as Discussion forums and Question / answer forums. The *How-to / Instructional* main register consists of sub-registers providing different kinds of instructions, such as actual How-to pages, Recipes and Technical support pages. The *Informational persuasion (Info-Persuasion)* main register covers texts that use facts to persuade, such as Editorials and Descriptions with intent to sell. Finally, the *Lyrical* main register includes, e.g., Song lyrics and Poems, and the *Spoken* main register, e.g., Interviews and Video transcripts. For a detailed description of the CORE annotation process and corpus quality, we refer to Egbert et al. (2015).

The Finnish data is based on a sample of the Finnish Internet Parsebank (Luotolahti et al., 2015), a web-crawled corpus that currently consists of nearly 4 billion words. The annotations were done jointly by a supervisor and a dedicated annotator. The Finnish annotations aim to follow the CORE annotation guidelines as closely as pos-

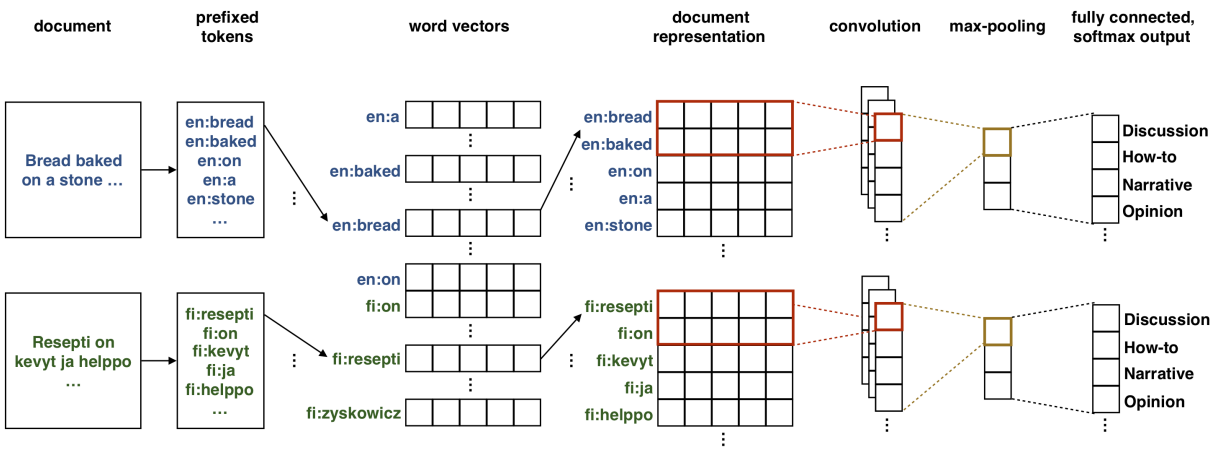

Figure 1: Illustration of text classification approach. Tokens are prefixed with language tags to differentiate e.g. the English word *on* from the Finnish word *on* 'is'. Multilingual word vectors are used and the same network applied regardless of language to allow cross-lingual and multilingual training and classification. (Following in part Kim (2014))

sible. The process advances through a decision tree, where the annotator 1) evaluates the mode of the text (spoken or written), 2) determines whether the text is interactive (multiple authors) or non-interactive (one author) and 3) identifies the general register of the text. Finally, the most accurate sub-register is selected if applicable. If the text appears to have more than one appropriate register, the annotator may choose up to three registers. Texts with several registers are called *hybrid texts*.

In this paper, we focus on the main register level because of the small size of the Finnish data set. Furthermore, to simplify the task setting, we use only the CORE documents for which at least three out of four annotators agreed on the register, thus excluding English hybrid texts, and similarly exclude the Finnish documents that were identified as hybrids. Finally, as the numbers of annotated Finnish texts in the main registers *Spoken* and *Lyrical* were too low for meaningful evaluation (11 and 0, respectively), these registers were excluded from the experiments.

The distribution of documents in the data used in our experiments is shown in Table 1. We note that the classes are very unevenly distributed, and the distributions are quite different in the two languages. In English, Narrative represents half of the data, with Opinion being the second most frequent with 16%. For Finnish, Narrative covers only 35%, and the second most frequent register is Informational persuasion, at 20%. For English, this is one of the least frequent classes, with only 3% of the data.

Both data sets were split into training, development and test sets using a stratified 70%/10%/20% split. The test data was held out during method development and parameter selection and only used for the final experiments.

4 Methods

Our approach is based on a simple convolutional neural network (CNN) architecture following Kim (2014) and illustrated in Figure 1. Documents are first tokenized using the Turku Neural Parser (Kanerva et al., 2018) trained on language-specific Universal Dependencies (Nivre et al., 2016) resources. The input is represented as a word vector sequence to a convolution layer with ReLU activation, followed by max-pooling and a fully-connected output layer. Similarly to Schwenk and Li (2018), we use pretrained multilingual word embeddings for multi- and cross-lingual classification; to differentiate between the same word forms in different languages, we simply prefix a language tag to each token and modify word vector indexing analogously. We use English and Finnish word vectors from the Multilingual Unsupervised and Supervised Embeddings (MUSE) library[1] (Conneau et al., 2018) in all experiments. As MUSE word vectors are uncased, we lowercase text following tokenization.

Based on initial experiments on the development set, we set the maximum number of word vectors to 100000, the number of CNN filters to

[1] https://github.com/facebookresearch/MUSE

Setting	Monolingual				Cross-/Multilingual	
Method	fastText		CNN		CNN	
Training data	Finnish	English	Finnish	English	English	En + Fi
Test data	Finnish	English	Finnish	English	Finnish	Finnish
D-Informational	67.1%	93.9%	75.4%	94.1%	69.0%	75.4%
Discussion	86.5%	93.3%	83.1%	96.5%	80.1%	86.5%
How-to	84.6%	94.9%	88.3%	94.8%	82.9%	89.7%
Info-Persuasion	84.5%	93.2%	84.7%	95.2%	74.0%	85.5%
Narrative	76.3%	91.9%	85.2%	92.7%	79.8%	86.3%
Opinion	78.2%	86.6%	86.2%	88.2%	85.8%	88.3%
Average	79.5%	92.3%	83.8%	93.6%	78.6%	85.3%

Table 2: Evaluation results (AUC scores) in mono-, cross-, and multilingual training settings.

128, the filter size to one word, and froze the word vector weights. Wider filters and word vector fine-tuning appeared to give modest benefit in monolingual settings, and reduced performance in cross-lingual settings. The latter results were expected given that wider filters capture aspects of word order that are not consistent cross-lingually, and fine-tuned word vectors may no longer align across languages. Input texts are padded or truncated to 1000 tokens. We train the CNN for 10 epochs using Adam with the default settings suggested by Kingma and Ba (2014). We refer to Kim (2014) and Conneau et al. (2018) for further information on the model and word vectors, and our open-source release[2] for implementation details.

For reference, we also report results using fastText (Joulin et al., 2016), a popular text classification method based on word vector averages that emphasizes computational efficiency. We initialize fastText with the same word vectors and train for the same number of epochs as the CNN, and retain its parameters otherwise at their defaults. As fastText does not support cross-lingual classification, we only use it in the monolingual setting.

The class disbalance and the different class distributions in the two languages represent challenges for cross-lingual generalization and evaluation. We opted to focus on ranking and evaluate performance for each register in a one-versus-rest setting using the distribution-independent area under the receiver operating characteristic curve (AUC) measure. Additionally, to account for random variation from classifier initialization, we repeat each experiment ten times and report averages over these runs.

[2] https://github.com/TurkuNLP/multiling-cnn

5 Results

The primary results are summarized in Table 2. First, we briefly note that in a monolingual setting, the CNN and fastText results are broadly comparable, with the CNN achieving slightly higher performance for both English and Finnish overall as well as for most individual classes. This confirms that the somewhat restricted nature of the CNN (e.g. frozen word vector weights) does not critically limit its performance at the task. As expected, performance is notably higher for English, which has more than 10 times the number of annotated examples for Finnish.

In the cross-lingual setting, we find that when trained on English data and tested on Finnish, the CNN clearly outperforms the random baseline (50%) for all classes, confirming the basic feasibility of the approach to cross-lingual register identification. As expected, performance is below the comparable monolingual results (Finnish-Finnish), but the differences are encouragingly small; in particular, the cross-lingual CNN performance is very close to the monolingual fastText baseline.

The best results for Finnish are achieved when training on the combination of English and Finnish data, both overall as well as for most individual classes. Given the different languages and independent development histories of these two corpora, it is far from given that this corpus combination would be successful, and this result is very positive in indicating both the basic compatibility of these specific resources as well as the broader ability to generalize the CORE register classification and annotation strategy to new languages.

To gain further insight into the effectiveness of multilingual training, we evaluated Finnish regis-

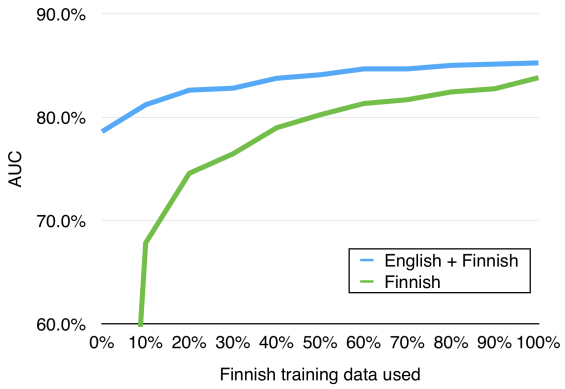

Figure 2: Average AUC for Finnish register prediction when training with varying proportions of Finnish training data, contrasting performance with and without additional English training data.

ter classification performance using subsets (10%, 20%, ...) of the Finnish training data both in the monolingual (Finnish only) and multilingual (English and Finnish) training settings. All of the English training data was used in the latter setting. The results, summarized in Figure 2, show that with these corpora, multilingual training is beneficial regardless of the size of available target language data, and that zero-shot cross-lingual classification (no target language data) outperforms monolingual classification with up to 900 examples of target language data.

6 Discussion and future work

In this paper, we explored the identification of registers in Internet texts in cross- and multilingual settings. We introduced FinCORE, the first non-English corpus annotated following the guidelines of the CORE corpus, the largest online register corpus representing the full range of linguistic variation found online. Evaluation using a simple CNN with multilingual word vectors indicated that cross-lingual register classification is feasible, and that combination of the large CORE corpus data with smaller target language data further benefits classification performance. This positive result also confirmed the compatibility of the English and Finnish corpus annotations.

While our study has only considered a single language pair, we note that the general approach is immediately applicable to any language for which a tokenizer and multilingual word vectors are available, including 30 languages in MUSE at the time of this writing. As the approach avoids many language-specific features (e.g. word order) and is demonstrated on a pair of languages that are not closely related, we are optimistic regarding its ability to generalize to other languages.

This is an early study in a relatively new area and leaves open several avenues to explore. For example, our approach is based on a straightforward application of convolutional neural networks for text classification, and it is likely possible to improve performance through further model development and parameter optimization. Future work should also consider the effectiveness of more advanced deep learning methods, such as multilingual transformer architectures. In current and planned future work, we are building on these initial results to address additional languages as well as the full CORE register hierarchy.

All of the data and methods newly introduced in this work are available under open licenses from `https://github.com/TurkuNLP/FinCORE`.

Acknowledgements

We thank Fulbright Finland, Kone foundation and Emil Aaltonen Foundation for financial support.

References

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A Smith. 2016. Massively multilingual word embeddings. *arXiv preprint arXiv:1602.01925*.

Daniel Andrade, Kunihiko Sadamasa, Akihiro Tamura, and Masaaki Tsuchida. 2015. Cross-lingual text classification using topic-dependent word probabilities. In *Proceedings of HLT-NAACL*.

Noushin 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.

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*.

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.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. *CoRR*, abs/1710.04087.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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

Eugenie Giesbrecht and Stefan Evert. 2009. Is part-of-speech tagging a solved task? an evaluation of pos taggers for the german web as corpus. In *Web as Corpus Workshop (WAC5)*, pages 27–36.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Jenna Kanerva, Filip Ginter, Niko Miekka, Akseli Leino, and Tapio Salakoski. 2018. Turku neural parser pipeline: An end-to-end system for the CoNLL 2018 shared task. In *Proceedings of CoNLL 2018 Shared Task*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *Proceedings of EMNLP*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of COLING*.

Patrik Lambert. 2015. Aspect-level cross-lingual sentiment classification with constrained SMT. In *Proceedings of ACL*.

Juhani Luotolahti, Jenna Kanerva, Veronika Laippala, Sampo Pyysalo, and Filip Ginter. 2015. Towards universal web parsebanks. In *Proceedings of Depling'15*, pages 211–220.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of LREC 2016*, pages 1659–1666.

Philipp Petrenz and Bonnie Webber. 2011. Stable classification of text genres. *Computational Linguistics*, 37(2):385–393.

Dimitrios Pritsos and Efstathios Stamatatos. 2018. Open set evaluation of web genre identification. *Language Resources and Evaluation*.

Holger Schwenk and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. *Proceedings of the 2nd Workshop on Representation Learning for NLP*.

Holger Schwenk and Xian Li. 2018. A corpus for multilingual document classification in eight languages. In *Proceedings of the 11th Language Resources and Evaluation Conference*.

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

Lei Shi, Rada Mihalcea, and Mingjun Tian. 2010. Cross language text classification by model translation and semi-supervised learning. In *EMNLP*.

Jörg Tiedemann, Fabienne Cap, Jenna Kanerva, Filip Ginter, Sara Stymne, Robert Östling, and Marion Weller-Di Marco. 2016. Phrase-based SMT for finnish with more data, better models and alternative alignment and translation tools. In *Proceedings of the First Conference on Machine Translation*, volume 2, pages 391–398.

Bonnie Webber. 2009. Genre distinctions for discourse in the Penn treebank. In *Proceedings of ACL-IJCNLP*, pages 674–682.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. Conll 2018 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of CoNLL 2018 Shared Task*, pages 1–21.