How Universal is Genre in Universal Dependencies?

Max Müller-Eberstein and **Rob van der Goot** and **Barbara Plank**
Department of Computer Science
IT University of Copenhagen, Denmark
mamy@itu.dk, robv@itu.dk, bapl@itu.dk

Abstract

This work provides the first in-depth analysis of genre in Universal Dependencies (UD). In contrast to prior work on genre identification which uses small sets of well-defined labels in mono-/bilingual setups, UD contains 18 genres with varying degrees of specificity spread across 114 languages. As most treebanks are labeled with multiple genres while lacking annotations about which instances belong to which genre, we propose four methods for predicting instance-level genre using weak supervision from treebank metadata. The proposed methods recover instance-level genre better than competitive baselines as measured on a subset of UD with labeled instances and adhere better to the global expected distribution. Our analysis sheds light on prior work using UD genre metadata for treebank selection, finding that metadata alone are a noisy signal and must be disentangled within treebanks before it can be universally applied.

1 Introduction

Identifying document genre automatically has long been of interest to the NLP community due to its immediate applications both in document grouping (Petrenz, 2012) as well as task-specific data selection (Ruder and Plank, 2017; Sato et al., 2017).

Cross-lingual genre identification has however remained a challenge, mainly due to the lack of stable cross-lingual representations (Petrenz, 2012). Recent work has shown that pre-trained masked language models (MLMs) capture monolingual genre (Aharoni and Goldberg, 2020). Do such distinctions manifest in highly multilingual spaces as well? In this work, we investigate whether this property holds for the genre distribution in the 114 language Universal Dependencies corpus (UD version 2.8; Zeman et al., 2021) using the multilingual mBERT MLM (Devlin et al., 2019).

In absence of an exact definition of textual genre (Kessler et al., 1997; Webber, 2009; Plank, 2016), this work will focus on the information specifically denoted by the `genres` metadata tag in UD. We hope that an in-depth, cross-lingual analysis of what this label represents will enable practitioners to better control for the effects of domain shift in their experiments. Previous work using these UD metadata for proxy training data selection have produced mixed results (Stymne, 2020). We investigate possible reasons and identify inconsistencies in genre annotation. The fact that genre labels are only available at the level of treebanks makes it difficult to gather a clear picture of the *sentence-level* genre distribution — especially with some treebanks having up to 10 genre labels. We therefore investigate the degree to which instance-level genre is recoverable using only the treebank-level metadata as weak supervision.

Our contributions entail the, to our knowledge, first detailed definition of all UD metadata genre labels (Section 3), four weakly supervised methods for extracting instance-level genre across 114 languages (Section 4) as well as genre identification experiments which show that our proposed two-step procedure allows for effective genre recovery in multilingual setups where language relatedness typically outweighs genre similarities (Section 5).[1]

[1]Code available at `https://personads.me/x/syntaxfest-2021-code`.

2 Related Work

The largest hurdle for cross-lingual genre classification is the lack of shared representational spaces. Sharoff (2007) use shared POS n-grams in order to jointly classify the genre of English and Russian documents. Petrenz (2012) similarly seek out features which are stable across languages in order to classify English and Chinese documents into four shared genres. A recent data-driven approach finds that monolingual MLM embeddings can be clustered into five groups closely representing the data sources of the original corpus (Aharoni and Goldberg, 2020). In this work, we investigate whether this holds for multilingual settings as well.

Being able to identify textual genre has been crucial for domain-specific fine-tuning (Dai et al., 2020; Gururangan et al., 2020) including dependency parsing. For parser training, in-genre data is typically selected by proxy of the data source (Plank and van Noord, 2011; Rehbein and Bildhauer, 2017; Sato et al., 2017). Data-driven approaches which include automatically inferred topics based on word and embedding distributions (Ruder and Plank, 2017) as well as POS-based approaches (Søgaard, 2011; Rosa, 2015; Vania et al., 2019) have also been found effective.

Universal Dependencies (Nivre et al., 2020) aims to consolidate syntactic annotations for a wide variety of languages and genres under a single scheme. The latest release contains 114 languages — many with fewer than 100 sentences. In order for languages at all resource levels to benefit from domain adaptation, it will continue to be important to identify cross-lingually stable signals for genre. While language labels are generally agreed upon, differences in genre are more subtle. Metadata at the treebank level provides some insights into genres of original data sources, however these are "neither mutually exclusive nor based on homogeneous criteria, but [are] currently the best documentation that can be obtained" (Nivre et al., 2020).

Stymne (2020) performs an initial study on using these treebank metadata labels for the selection of spoken and Twitter data. Results show that training on out-of-language/in-genre data is superior to out-of-language/out-of-genre data. However the best results are obtained using in-language data regardless of genre-adherence. This holds across multiple methods of proxy dataset selection (e.g. treebank embeddings; Smith et al., 2018).

Recently, Müller-Eberstein et al. (2021) have shown that combining UD genre metadata and MLM embeddings can improve proxy training data selection for zero-shot parsing of low-resource languages. The use of genre in their work is more implicit as it is mainly driven by the genre of the target data. In contrast, this work takes a holistic view and explicitly examines the classification of instance-level genre for all sentences in UD.

As genre appears to be a valuable signal, we set out to investigate how it is defined and distributed within UD. Due to the coarse, treebank-level nature of current genre annotations, we hypothesize that a clearer picture can only be obtained by moving to the sentence level. We therefore transition from prior supervised document genre prediction to weakly supervised *instance* genre prediction. Additionally, we expand the linguistic scope from mono- or bilingual corpora to all 114 languages currently in UD.

More generally, this task can be viewed as predicting genre labels for all sentences in all corpora of a collection while only being given the set of labels said to be contained in each corpus.

3 UD-level Genre

We analyze genre as currently used in the `genres` metadata of 200 treebanks from Universal Dependencies version 2.8 (Zeman et al., 2021). Section 3.1 provides an overview of all UD genre types and Section 3.2 analyzes how these global labels relate to the subset of treebanks which do provide treebank-specific, instance genre annotations.

3.1 Available Metadata

UD 2.8 (Zeman et al., 2021) contains 18 genres which are denoted in each treebank's accompanying metadata. Around 36% of treebanks contain a single genre while the remaining majority can contain between 2–10 which are not further labeled at the instance level. There is no official description of each genre label, however they can be roughly categorized as follows:

academic Collections of scientific articles covering multiple disciplines. Note that this label may subsume others such as *medical*.

bible Passages from the bible, frequently from older languages (e.g. Old Church Slavonic-PROIEL by Haug and Jøhndal, 2008). Largely non-overlapping passages are used across treebanks.

blog Internet documents on various topics which may overlap with other genres such as *news*. They are typically more informal in register. Some treebanks group social media content and reviews under this category (e.g. Russian-Taiga by Shavrina and Shapovalova, 2017).

email Formal, written communication. This includes English-EWT's (Silveira et al., 2014) subsection based on the Enronsent Corpus (Styler, 2011) as well as letters attributed to Dante Alighieri as part of Latin-UDante (Cecchini et al., 2020).

fiction Mostly paragraphs from diverse sets of fiction books and magazines.

government The least represented genre, mainly denoting texts from governmental sources. These include political speeches (English-GUM by Zeldes, 2017) as well as inscriptions from Neo-Assyrian kings from around 900 BCE (Akkadian-RIAO by Luukko et al., 2020).

grammar-examples Sentences from teaching or grammatical reference books which are typically short, but cover a wide range of dependency relations (e.g. Tagalog-TRG by Samson and Cöltekin, 2020).

learner-essays Small genre occurring in three single-genre treebanks. Sentences were written by second-language learners and either contain original errors (English-ESL by Berzak et al., 2016), manual corrections (IT-Valico by Di Nuovo et al., 2019) or both (Chinese-CFL by Lee et al., 2017).

legal Relatively frequent genre based mostly on laws and legal corpora within the public domain.

medical Scientific articles/books in the field of medicine (e.g. cardiology, diabetes, endocrinology for Romanian-SiMoNERo by Mitrofan et al., 2019). It is subsumed by *academic* for some treebanks (e.g. Czech-CAC by Hladká et al., 2008).

news The highest-resource genre by a large margin corresponding to news-wire texts as well as online newspapers on specific topics (e.g. IT-news in German-HDT by Borges Völker et al., 2019).

nonfiction Second most frequent genre with a high degree of variance, subsuming e.g. *academic* and *legal*. German-LIT (Salomoni, 2019) contains three philosophical books from the 18th century. Other *non-fiction* treebanks can originate from multiple sources (e.g. books and internet) and time spans.

poetry Smaller, yet distinct genre covering mostly older texts and language variations (e.g. Old French-SRCMF by Stein and Prévost, 2013).

reviews Medium-resource genre covering informal online reviews with unnormalized orthography (e.g. English-EWT) as well as formal reviews (e.g. newspaper film reviews in Czech-CAC).

social Encompasses social media data such as tweets (e.g. Italian-TWITTIRÒ by Cignarella et al., 2019) as well as newsgroups (e.g. English-EWT). Some *spoken* data is co-labeled with this genre when it refers to colloquial speech (e.g. South Levantine Arabic-MADAR by Zahra, 2020).

spoken Distinct genre which typically consists of spoken language transcriptions. Sentences contain filler words and may have abrupt boundaries. Sources range from elicited speech of native speakers (Komi Zyrian-IKDP by Partanen et al., 2018) to radio program transcriptions (Frisian Dutch-Fame by Braggaar and van der Goot, 2021).

web Similarly ambiguous genre as *non-fiction*. It occurs in conjunction with specific genres such as *blog* and *social* and never appears alone (e.g. Persian-PerDT by Sadegh Rasooli et al., 2020).

wiki Denotes data from Wikipedia for which cross-lingual authoring guidelines exist.

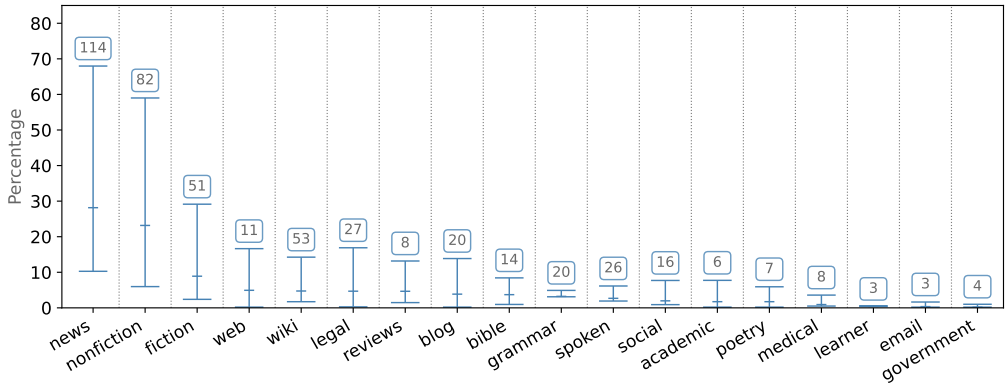


Figure 1: **Genre Distribution in UD Version 2.8.** Ranges indicate upper/lower bounds for sentences per genre inferred from UD metadata. Center marker reflects the distribution under the assumption that genres within treebanks are uniformly distributed. Labels above the bars indicate the number of treebanks which contain each genre.

Figure 1 shows the approximated distribution of these genres in UD. Maximum/minimum sentence counts are inferred from the size of single-genre treebanks plus the size of all treebanks in which a genre is said to occur. The center line denotes the distribution under the assumption that genres are uniformly distributed within each treebank.

It is clear that *news* and *non-fiction* constitute more than half of the entire dataset. Specialized genres such as *medical* are less represented. For broader genres such as *web*, which frequently co-occurs with others, the exact number of sentences is hard to estimate, but must lie between 0–20%. Considering these large variances, access to instance-level genre will likely be crucial for effective proxy data selection and downstream domain adaptation.

3.2 Instance-level Annotations

In addition to the aforementioned 18 treebank-level genre labels, some treebanks provide instance-level genre annotations in the comment-metadata before each sentence. We find such annotations in 26 out of 200 treebanks in UD 2.8 amounting to 124k or 8.25% of all sentences.

Out of this set, 20 treebanks belong to the Parallel Universal Dependencies (PUD; Nivre et al., 2017). They are split 500/500 between *news* and *wiki*, as denoted by sentence IDs beginning with `n` and `w` respectively. The parallel nature of PUD makes it interesting for analyzing cross-lingual genre identification performance. However these two genres only represent a small fraction of non-fiction texts and furthermore, each PUD-treebank is test-split-only. Note also that Polish-PUD as an exception has the metadata labels *news* and *non-fiction*.

The remaining six treebanks for which we were able to identify instance-level genre annotations are Belarusian-HSE (Lyashevskaya et al., 2017), Czech-CAC (Hladká et al., 2008), English-EWT (Silveira et al., 2014), German-LIT (Salomoni, 2019), Polish-LFG (Patejuk and Przepiórkowski, 2018) and Russian-Taiga (Shavrina and Shapovalova, 2017). They cover a wider set of 12 genres. Annotation schema vary across treebanks and are neither fully compatible amongst each other nor with the 18 UD labels. Approximate mappings can however be drawn thanks to source data documentation by the respective authors (Section 4.2).

Further comment-metadata which may guide genre separation within treebanks includes document, paragraph and source identifiers. Again, these are unfortunately not available for all sentences (although coverage of these metadata reaches up to 45%) and their values do not provide further indications about genre adherence.

4 Instance Genre from Treebank Labels

From the previous analysis, it is evident that finer-grained genre labels are needed before domain adaptation can be successful across all languages.

Formally, the task of predicting instance-level UD genre can be defined as assigning a set of labels $\mathcal{L} = \{l_0, l_1, \ldots, l_K\}$ (i.e. genres) to all instances x_n of a corpus $\mathcal{X}$ (i.e. UD). The corpus consists of S distinct subsets $\mathcal{X} = \{\mathcal{X}_0 \cup \mathcal{X}_1 \cup \ldots \cup \mathcal{X}_S\}$ (i.e. treebanks) each with a subset of labels $\mathcal{L}_s \subseteq \mathcal{L}$. As no instance-level labels $x_n \rightarrow l$ are available, models must learn this mapping based solely on the subset of labels said to be contained in each data subset $\mathcal{X}_s \rightarrow \mathcal{L}_s$.

4.1 Genre Prediction Methods

As instance-level labels are noisy and sparse, we investigate two classification-based and two clustering-based approaches for inferring instance genre labels from the treebank metadata $\mathcal{L}_s$ alone. Building on Müller-Eberstein et al. (2021), our proposed methods leverage latent genre information in the pre-trained mBERT language model (Devlin et al., 2019).

BOOT In order to select proxy training data which matches the genre of an unseen target, Müller-Eberstein et al. (2021) propose a bootstrapping-based approach to genre classification (BOOT). An mBERT-based classifier (Devlin et al., 2019) is initially trained on sentences from single-genre treebanks, corresponding to standard supervised classification. Above a confidence threshold (i.e. softmax probability of 0.99), sentences from treebanks containing a known genre in mixture are bootstrapped as single-genre training data for the next round. After bootstrapping sentences from all known genres, the remaining unclassified instances of any treebank containing a single unknown genre are inferred to be of that last genre. While this method was previously used for targeted data selection, we investigate the degree to which it actually recovers instance-level genre.

CLASS With approximate classification (CLASS), we simplify BOOT to naively learn instance genre labels from weak supervision. It fine-tunes the same mBERT MLM with a 18-genre classification layer on the `[CLS]`-token. For single-genre treebanks it is possible to measure the exact cross entropy between the predicted probability and the target (i.e. $x_n \rightarrow l$ with $l \in \mathcal{L}_s$ and $|\mathcal{L}_s| = 1$). For multi-genre treebanks with $|\mathcal{L}_s| > 1$, this is not possible as the gold label is unknown. For the CLASS approach, each sentence from a k-genre treebank is therefore classified k times — once for each class in $\mathcal{L}_s$.

GMM In addition to classification, we also evaluate two common clustering algorithms. First we investigate whether clusters formed by untuned MLM sentence embeddings (mean over sentence subwords) represent genre to such a degree that Gaussian Mixture Models can recover the 18 UD genre groups. For monolingual data from five genres, such clusters were shown to be recoverable (Aharoni and Goldberg, 2020). We extend this approach to the 114 language setting of UD.

LDA As all methods so far are to some degree dependent on the pre-trained MLM representations, we also evaluate the recoverability of genre using Latent Dirichlet Allocation (Blei et al., 2003) with lexical features. Feature vectors are constructed using the frequency of character 3–6-grams.

Cluster Labeling Both clustering methods produce 18 groups of sentences from UD, however these will not carry meaningful labels as with classification. While labels could be assigned manually post-hoc by matching representative sentences in each cluster to one of the 18 global UD genres, this process is bound to be subjective and also depends on the annotator to be fluent in most of the 114 languages.

In order to automate this procedure, we propose **GMM+L** and **LDA+L** which combine clustering and classification. Both methods start by clustering each treebank $\mathcal{X}_s$ into the number of genres specified by its metadata (note that standard GMM and LDA cluster all of UD at once, i.e. $\mathcal{X}$).

Next, the mean embedding of each cluster is computed such that they can be compared in a single representational space. Note that this would not be possible using monolingual models as their latent spaces are not as cross-lingually aligned. Analogous to BOOT, single-genre treebanks can then be used as a single-label signal such that the closest cluster from each treebank containing the respective genre

can be extracted. Newly identified clusters are added to the pool of single-genre clusters. This process need only be repeated for three rounds before all sentences in UD can be assigned a single label.

Using these four methods, we aim to assign a single genre label to each sentence in UD. By comparing model ablations, we further depart from prior work and explicitly quantify the genre information in MLM embeddings as well as how it manifests within and across treebanks in UD.

4.2 Supervised Evaluation

For the 26 treebanks with instance genre labels, we are able to measure standard F1 after applying a mapping from the treebank-specific labels to the 18 global UD genre labels. The mapping was created according to the following criteria.

First, we only allowed treebank-specific genre labels to be mapped to the set of UD genre labels specified in each treebank's metadata.

Second, if possible treebank labels are mapped to UD labels of the same name (e.g. `fiction` $\rightarrow$ *fiction*) or to the closest subsuming category (e.g. `spoken (prepared)` $\rightarrow$ *spoken*).

Third, decisions involving subjective uncertainty were based on the label which covers the majority of data sources. E.g., Czech-CAC has the metadata label set {*legal, medical, news, non-fiction, reviews*} and only three types of instance labels (`aw`, `nw`, `sw`). The `sw` (scientific-written) label is attached to many medical articles, but also to articles on philosophy or music. While *academic* may be the most fitting label, it is not in the metadata. As such we chose the broader *non-fiction* as the target label.

The full mapping is in Appendix A and we hope future work will be able to expand upon it.

4.3 Unsupervised Evaluation

For the remaining 174 treebanks without sentence-level gold labels it is difficult to measure the exact quality of the predicted genre distributions. Nonetheless, treebank annotations provide enough information for approximate, global comparisons.

Based on label/cluster assignments, it is possible to compute the standard cluster purity measure (PUR; Schütze et al., 2008). Across treebanks of the same genre, the majority of sentences should belong to the same label/cluster. We measure this using the ratio of cross-treebank label agreement (AGR). As in prior work (Aharoni and Goldberg, 2020) it is important to note that the aforementioned metrics can be misleading when taken on their own: A perfect score can for example be achieved by simply assigning all instances to the same genre.

To mitigate this issue we turn to the expected overlap of inter-treebank genre distributions. For multi-genre treebanks, it is known which genres are present, but not how they are distributed. Since treebanks are expected to have a certain amount of overlap, we can however estimate a global error. A {*fiction, spoken, wiki*} treebank should for example have no clusters in common with a {*news*} treebank, but should have many sentences in the same clusters as a {*fiction, medical, spoken*} one. Assuming that genres are uniformly distributed within each treebank, the first pair would share 0 mass between distributions while the second pair would share $\frac{2}{3}$. Intuitively, a good prediction would produce a global genre distribution that falls precisely between the metadata range bars of Figure 1, close to the center markers.

To quantify the overlap between two treebank genre distributions p and q over the genres in $\mathcal{L}_s$, we use the discrete Bhattacharyya coefficient:

$$BC(p, q) = \sum_{l \in \mathcal{L}_s} \sqrt{p(l)q(l)} \tag{1}$$

which has often been applied to distributional comparisons (Choi and Lee, 2003; Ruder and Plank, 2017). It is computed for all pairs of treebanks such that the overlap error $\Delta\text{BC} \in [0, 100]$ is the mean absolute difference between the expected distributional overlap of each treebank pair and the predicted one (i.e. lower is better).

While none of these metrics can individually provide an exact measure of a prediction method's fit to the UD-specified distribution, they complement each other as to allow for global comparisons in absence of any sentence-level annotations.

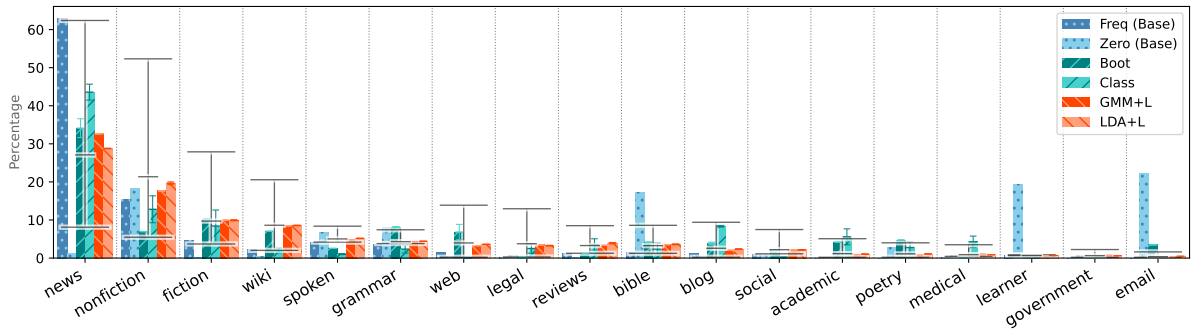


Figure 2: **Genre Predictions on UD (Test).** Ranges indicate upper/lower bounds inferred from UD metadata and the distribution under treebank-level uniformity at the center marker. Bars show averaged distribution predictions with standard deviations by FREQ, ZERO, BOOT, CLASS, GMM+L and LDA+L.

5 Experiments

5.1 Setup

Data From the 1.5 million sentences in UD, we construct global training, development and testing splits. All original test splits are left unchanged and gathered into one global test split containing 204k sentences. Note that test-only treebanks and languages are thereby never seen during training or tuning. For instance-level, supervised evaluation, this means that all PUD treebanks and German-LIT are excluded, leaving five treebanks for tuning.

Next, all original training and development splits are concatenated and split 10/90 into a global training and development split with 102k and 915k sentences respectively. The reason for this small "training" split is that it is only required for training CLASS and BOOT. Within it, we again split the data 70/30 (71k and 31k sentences) for classifier training and held-out data for early stopping. All exact splits are provided in Appendix A.

Baselines For our comparisons, we use a maximum frequency baseline (FREQ) which labels all sentences within a treebank with the metadata genre label that is most frequent overall. For example, in any treebank containing *news*, all instances are labeled as such.

In order to measure the untuned classification performance of mBERT, we propose an additional zero-shot classification baseline (ZERO). Prior research has found that classifying sentences based solely on their cosine similarity to genre label strings in MLM embedding space can be remarkably effective (Veeranna et al., 2016; Yin et al., 2019; Davison, 2020). For example, a sentence is labeled as *academic* if this is the closest embedded label out of all 18 genre strings.

Training Every method from Section 4.1 is run with three initializations. CLASS and BOOT are trained for a maximum of 30 epochs with an early stopping patience of 3. ZERO, GMM+L and LDA+L (by extension GMM, LDA) do not require training and can be directly applied to the target data. Implementation details and development results are reported in Appendices B and C.

5.2 Results

Using the 8% subset of annotated instances (Section 4.2) in addition to the unsupervised metrics from Section 4.3, we can gather an estimate of each method's performance in Table 1. UD-level genre predictions in addition to instance-level confusions are further visualized in Figures 2 and 3.

Baselines The FREQ baseline highlights the issue of using individual unsupervised metrics for estimating performance. As it assigns all sentences per treebank to the same genre, it automatically achieves 100% single-genre treebank purity and agreement. Considering that the instance-level F1 covers 12 genres, a baseline score of 47 is also competitive. Note that this is mostly due to the data imbalance towards *news*. This unlikely distribution predicted by FREQ is also reflected in Figure 2.

METHOD	PUR	AGR	ΔBC	F1
FREQ	100±0.0	100±0.0	21±0.0	47±0.0
ZERO	46±0.0	56±0.0	47±0.0	12±0.0
CLASS	83±1.4	63±3.9	34±1.1	32±0.9
BOOT	86±0.4	70±0.7	29±0.3	38±1.2
GMM	90±0.5	45±2.6	31±0.3	—
+LABELS	100±0.0	100±0.0	4±0.2	54±2.1
LDA	77±0.8	34±2.6	31±0.2	—
+LABELS	100±0.0	100±0.0	2±0.1	51±1.5

Table 1: **Results of Genre Prediction on UD (Test).** Purity (PUR ↑), agreement (AGR ↑), overlap error (ΔBC ↓) and micro-F1 over instance-labeled TBs (F1 ↑) for FREQ, ZERO, CLASS, BOOT and GMM, LDA with/without cluster label predictions (+LABELS). Standard deviation denoted ±.

ZERO-shot classification is not fine-tuned on UD-specific signals and as such predicts a genre distribution that does not adhere to the metadata at all (see Figure 2). It severely underpredicts high-frequency genres such as *news* and overpredicts less frequent genres such as *email*. This reflects in our metrics, with ZERO obtaining the lowest PUR, AGR and F1 while having the highest ΔBC of 47.

Classification With regard to explicit genre fine-tuning, CLASS increases purity by 38 points compared to ZERO. Agreement across treebanks also improves, while overlap error decreases. These differences are also reflected in Figure 2 in that the predicted distribution is more within the range that would be expected given the metadata.

BOOT fits the UD genre distribution more closely, resulting in a purity that is 4 points higher and agreement that is 11 points higher than CLASS. F1 also increases by 6 points while overlap error decreases by 4 points, indicating that these improvements are not merely due to e.g. assigning all sentences to the same genre. While instance-level F1 is below the FREQ baseline, both methods improve upon the untuned ZERO by a factor of 3.

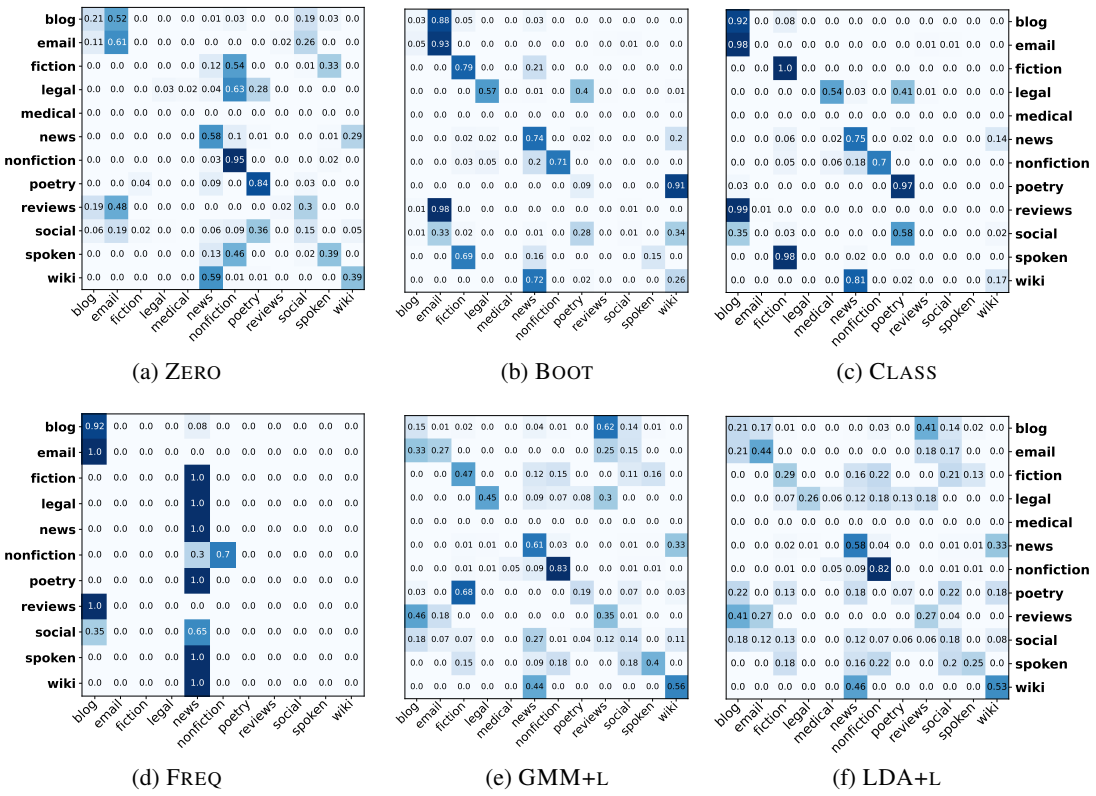
The benefits of the less noisy training signal are visible in Figure 2: Compared to CLASS, BOOT predicts labels in a way that more closely resembles the expected distribution even when the label only occurs in multi-genre treebanks and is ambiguous (e.g. *web*). While BOOT agrees upon the same genre-label across languages (e.g. all *social* treebanks are labeled as such), CLASS tends to overassign the globally most frequent labels (e.g. half of *social* treebanks are labeled *wiki*) and has a larger variance in its assignments across initializations.

Clustering GMM clusters from untuned mBERT embeddings follow the distribution specified by UD metadata more than the LDA clusters produced from lexical information. Although sentence representations are gathered using a naive mean-pooling approach, the resulting clusters reach 90% PUR compared to 77% for LDA. AGR follows a similar pattern and ΔBC is equivalent.

Turning to our cluster labelling approaches, both GMM+L and LDA+L obtain the highest overall F1 scores, outperforming both baselines. They achieve 100% PUR and AGR by the same process as the FREQ-baseline while their overlap error is significantly lower at 4 and 2 points respectively. Figure 2 reflects this, as GMM+L and LDA+L are always closest to the expected genre distribution, regardless of overall genre frequency. This shows how focusing on treebank-internal differences before applying a global labelling procedure combines the benefits of local clustering with the benefits of bootstrapped classification, resulting in an effective overall method.

5.3 Analysis

From the F1 scores in Table 1 it is clear that predicting instance genre based on treebank metadata alone — while accounting for its skewed distribution and inter-treebank shifts of genre definitions — is a difficult task. In the following we analyze the performance characteristics of each method.



(a) ZERO (b) BOOT (c) CLASS

(d) FREQ (e) GMM+L (f) LDA+L

Figure 3: **Confusions of Instance-level Genre.** Ratios of predicted labels (columns) per target (row) for ZERO, BOOT, CLASS, FREQ, GMM+L, LDA+L on test splits of 26 instance-annotated treebanks.

Overall, trends of the unsupervised metrics follow the supervised F1, leading us to believe that the methods would behave comparatively should labels for all instances in UD be available. The confusion matrices with prediction ratios per gold label in Figure 3 reflect our previous observations.

Baselines The FREQ baseline's predictions are clearly dominated by the most frequent *news* genre, followed by the similarly high frequency *non-fiction* and *blog* (see Figure 3d).

ZERO appears to follow a pattern similar to BOOT (e.g. *blog* and *email*), however it also makes more predictions away from the diagonal (see Figure 3a).

Classification Both CLASS (Figure 3c) and BOOT (Figure 3b) assign most instances of a genre to a single prediction label, often strongly aligning with the target diagonal. CLASS more often assigns a single label per target instead of spreading out predictions across multiple labels as in BOOT. Nonetheless, both methods make some unintuitive errors such as BOOT classifying parts of *poetry* as *wiki*. For these 68 samples from Russian-Taiga, BOOT likely overfits the language signal from Russian-GSD (McDonald et al., 2013; *wiki*).

Compared to ZERO which approximates the predictions of an untuned mBERT model, BOOT and CLASS fine-tuning appears to amplify existing patterns and shifts some predictions to better align with genres as defined in UD (e.g. *fiction* and *legal* in BOOT).

Clustering Grouping all 1.5 million sentences of UD into 18 unlabeled clusters using GMM and LDA results in purity and ΔBC comparable to CLASS and BOOT. However, looking into the cluster contents of the former reveals that they are oversaturated with large treebanks such as German-HDT. Cosine similarities of cluster centroids from the mBERT-based GMM further indicate that proximity corresponds foremost to language similarity.

Some clusters predominantly contain *news*, *wiki* or *social*. This corresponds to cases such as the Italian Twitter treebank TWITTIRÒ in which specific tokens (e.g. "@user") are distinct enough to override the

language signal. Overall, most UD-level clusters do not have clear genre distinctions and are influenced more strongly by language than genre, resulting in high treebank purity while having low intra-treebank agreement. Attempting to cross-lingually cluster all sentences in UD directly is therefore not as effective for recovering instance-level genre as it was in the monolingual setting (Aharoni and Goldberg, 2020).

Initially constructing clusters within each treebank as in the GMM+L and LDA+L methods appears to restore the benefits observed in the monolingual setting. A qualitative analysis of the treebank-level LDA clusters reveals that *wiki* clusters often contain lexical indicators for the genre, such as brackets, while *news* features often contain n-grams which may be related to spoken quotes such as "said", "Ik␣" (first person pronoun).

Attaching labels to these clusters using the globally shared mBERT space yields confusion plots for GMM+L and LDA+L which most closely follow the diagonal (see Figures 3e and 3f). Overall, their predictions follow a similar pattern indicating that clustering at the treebank-level using either mBERT embeddings or lexical features results in similar sentence groups.

Within the instance-labeled subset, all models share confusions between *news* and *wiki* (mainly from PUD). While *wiki* is often predicted as *news*, both GMM+L and LDA+L substantially improve upon this "*news*-bias" with a confusion ratio that is 13%–56% lower compared to all other methods. The sentence-bounded context from which all models must make their genre predictions nonetheless limits the amount of improvement possible. For example, using the aforementioned LDA features the algorithm would very likely be unable to distinguish between *news* and *wiki* (both non-fiction, edited texts describing facts) for cases such as, *"Weiss was honored with the literature prizes from the cities of Cologne and Bremen."*

6 Discussion and Conclusion

This work provided an in-depth analysis of the 18 genres in Universal Dependencies (UD) and identified challenges for projecting this treebank metadata to the instance level. As these genre labels were not part of the first UD releases, but were added in later versions, we identified large variations in the way they are interpreted and applied — resulting in far less universal definitions of genre than for syntactic dependencies. Most treebanks furthermore contain multiple genres while not providing finer-grained instance-level annotations thereof. This also sheds light on prior work which used UD metadata for training data selection, where treebank-level genre improved in-language parsing performance (Stymne, 2020) and where moving to instance-level genre signals lead to additional increases even across languages (Müller-Eberstein et al., 2021).

Building on the latent genre information stored in MLM embeddings, we investigated four methods for projecting treebank-level labels to the instance level. In contrast to prior monolingual work, immediately clustering multilingual embeddings yielded clusters dominated by language similarity instead of genre (Section 5.3). Similarly, zero-shot labelling using the untuned mBERT latent space proved to be insufficient for producing a genre distribution which adheres to the UD metadata. The classification-based CLASS and BOOT methods are able to extract a stronger genre signal from mBERT than ZERO.

Our proposed GMM+L and LDA+L methods which combine local treebank clusters with the global, cross-lingual representation space reach the best overall performance, outperforming both baselines as well as both classification methods at a much lower computational cost (Section 5.2; Appendix B). This highlights how the current genre annotations are far from universal, yet can still guide our local-to-global instance-level genre predictors in identifying cross-lingually consistent, data-driven notions of genre.

Future work may be able to improve instance genre prediction by using a more consistent label set or human annotations. The definition of genre macro-classes or a broader taxonomy covering existing annotations could also guide further investigations into cross-lingual language variation. Nonetheless, we expect the task of predicting sentence genre to remain difficult due to the short context within which both annotators and models must make their predictions.

Within the complex scenario of highly cross-lingual, instance-level genre classification, our methods have nonetheless demonstrated that genre is recoverable across the 114 languages in UD — shedding light on prior genre-driven work as well as enabling future research to more deliberately control for additional dimensions of language variation in their data.

Acknowledgements

We would like to thank the NLPnorth group for insightful discussions on this work — in particular Elisa Bassignana and Mike Zhang. Thanks to Héctor Martínez Alonso for feedback on an early draft as well as ITU's High-performance Computing Cluster team. Finally, we thank the anonymous reviewers for their helpful feedback. This research is supported by the Independent Research Fund Denmark (DFF) grant 9063-00077B and an Amazon Faculty Research Award (ARA).

References

Roee Aharoni and Yoav Goldberg. 2020. Unsupervised Domain Clusters in Pretrained Language Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7747–7763, Online. Association for Computational Linguistics.

Yevgeni Berzak, Jessica Kenney, Carolyn Spadine, Jing Xian Wang, Lucia Lam, Keiko Sophie Mori, Sebastian Garza, and Boris Katz. 2016. Universal Dependencies for learner English. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 737–746, Berlin, Germany, August. Association for Computational Linguistics.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Emanuel Borges Völker, Maximilian Wendt, Felix Hennig, and Arne Köhn. 2019. HDT-UD: A very large Universal Dependencies treebank for German. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 46–57, Paris, France, August. Association for Computational Linguistics.

Anouck Braggaar and Rob van der Goot. 2021. Challenges in annotating and parsing spoken, code-switched, Frisian-Dutch data. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 50–58, Kyiv, Ukraine, April. Association for Computational Linguistics.

Flavio M Cecchini, Rachele Sprugnoli, Giovanni Moretti, and Marco Passarotti. 2020. Udante: First steps towards the universal dependencies treebank of dante's latin works. In *Seventh Italian Conference on Computational Linguistics*, pages 1–7. CEUR-WS. org.

Euisun Choi and Chulhee Lee. 2003. Feature extraction based on the Bhattacharyya distance. *Pattern Recognition*, 36:1703–1709, 08.

Alessandra Teresa Cignarella, Cristina Bosco, and Paolo Rosso. 2019. Presenting TWITTIRÒ-UD: An Italian Twitter treebank in Universal Dependencies. In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pages 190–197, Paris, France, August. Association for Computational Linguistics.

Xiang Dai, Sarvnaz Karimi, Ben Hachey, and Cecile Paris. 2020. Cost-effective selection of pretraining data: A case study of pretraining BERT on social media. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1675–1681, Online, November. Association for Computational Linguistics.

Joe Davison. 2020. Zero-Shot Learning in Modern NLP, May. Accessed December 4th, 2020.

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, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Elisa Di Nuovo, Cristina Bosco, Alessandro Mazzei, and Manuela Sanguinetti. 2019. Towards an italian learner treebank in universal dependencies. In *6th Italian Conference on Computational Linguistics, CLiC-it 2019*, volume 2481, pages 1–6. CEUR-WS.

Fonticons. 2021. Font Awesome Icons. CC-BY 4.0 License.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online, July. Association for Computational Linguistics.

Dag TT Haug and Marius Jøhndal. 2008. Creating a parallel treebank of the old Indo-European Bible translations. In *Proceedings of the second workshop on language technology for cultural heritage data (LaTeCH 2008)*, pages 27–34.

Barbora Hladká, Jan Hajič, Jirka Hana, Jaroslava Hlaváčová, Jiří Mírovskỳ, and Jan Raab. 2008. The Czech academic corpus 2.0 guide. *The Prague Bulletin of Mathematical Linguistics*, 89(1):41–96.

Brett Kessler, Geoffrey Nunberg, and Hinrich Schutze. 1997. Automatic detection of text genre. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 32–38, Madrid, Spain, July. Association for Computational Linguistics.

John Lee, Herman Leung, and Keying Li. 2017. Towards Universal Dependencies for learner Chinese. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 67–71, Gothenburg, Sweden, May. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *Computing Research Repository*, arXiv: 1711.05101. version 3.

Mikko Luukko, Aleksi Sahala, Sam Hardwick, and Krister Lindén. 2020. Akkadian treebank for early neo-assyrian royal inscriptions. In *Proceedings of the 19th International Workshop on Treebanks and Linguistic Theories*, pages 124–134, Düsseldorf, Germany, October. Association for Computational Linguistics.

Olga Lyashevskaya, Angelika Peljak-Łapińska, and Daria Petrova. 2017. UD_Belarusian-HSE. `https://github.com/UniversalDependencies/UD_Belarusian-HSE`.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal Dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria, August. Association for Computational Linguistics.

Maria Mitrofan, Verginica Barbu Mititelu, and Grigorina Mitrofan. 2019. MoNERo: a biomedical gold standard corpus for the Romanian language. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 71–79, Florence, Italy, August. Association for Computational Linguistics.

Max Müller-Eberstein, Rob van der Goot, and Barbara Plank. 2021. Genre as weak supervision for cross-lingual dependency parsing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4786–4802, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Joakim Nivre, Željko Agić, Lars Ahrenberg, Lene Antonsen, Maria Jesus Aranzabe, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, John Bauer, Kepa Bengoetxea, Riyaz Ahmad Bhat, Eckhard Bick, Cristina Bosco, Gosse Bouma, Sam Bowman, Aljoscha Burchardt, Marie Candito, Gauthier Caron, Gülşen Cebiroğlu Eryiğit, Giuseppe G. A. Celano, Savas Cetin, Fabricio Chalub, Jinho Choi, Yongseok Cho, Silvie Cinková, Çağrı Çöltekin, Miriam Connor, Marie-Catherine de Marneffe, Valeria de Paiva, Arantza Diaz de Ilarraza, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Marhaba Eli, Ali Elkahky, Tomaž Erjavec, Richárd Farkas, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta Gonzáles Saavedra, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Nizar Habash, Jan Hajič, Jan Hajič jr., Linh Hà Mỹ, Kim Harris, Dag Haug, Barbora Hladká, Jaroslava Hlaváčová, Petter Hohle, Radu Ion, Elena Irimia, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşıkara, Hiroshi Kanayama, Jenna Kanerva, Tolga Kayadelen, Václava Kettnerová, Jesse Kirchner, Natalia Kotsyba, Simon Krek, Sookyoung Kwak, Veronika Laippala, Lorenzo Lambertino, Tatiana Lando, Phuong Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Cheuk Ying Li, Josie Li, Nikola Ljubešić, Olga Loginova, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Gustavo Mendonça, Anna Missilä, Verginica Mititelu, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Shunsuke Mori, Bohdan Moskalevskyi, Kadri Muischnek, Nina Mustafina, Kaili Müürisep, Pinkey Nainwani, Anna Nedoluzhko, Luong Nguyễn Thị, Huyền Nguyễn Thị Minh, Vitaly Nikolaev, Rattima Nitisaroj, Hanna Nurmi, Stina Ojala, Petya Osenova, Lilja Øvrelid, Elena Pascual, Marco Passarotti, Cenel-Augusto Perez, Guy Perrier, Slav Petrov, Jussi Piitulainen, Emily Pitler, Barbara Plank, Martin Popel, Lauma Pretkalniņa, Prokopis Prokopidis, Tiina Puolakainen, Sampo Pyysalo, Alexandre Rademaker, Livy Real, Siva Reddy, Georg Rehm, Larissa Rinaldi, Laura Rituma, Rudolf Rosa, Davide Rovati, Shadi Saleh, Manuela Sanguinetti, Baiba Saulīte, Yanin Sawanakunanon, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan

Seraji, Lena Shakurova, Mo Shen, Atsuko Shimada, Muh Shohibussirri, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Antonio Stella, Jana Strnadová, Alane Suhr, Umut Sulubacak, Zsolt Szántó, Dima Taji, Takaaki Tanaka, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Gertjan van Noord, Viktor Varga, Veronika Vincze, Jonathan North Washington, Zhuoran Yu, Zdeněk Žabokrtský, Daniel Zeman, and Hanzhi Zhu. 2017. Universal dependencies 2.0 – CoNLL 2017 shared task development and test data. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France, May. European Language Resources Association.

Niko Partanen, Rogier Blokland, KyungTae Lim, Thierry Poibeau, and Michael Rießler. 2018. The first Komi-Zyrian Universal Dependencies treebanks. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 126–132, Brussels, Belgium, November. Association for Computational Linguistics.

Agnieszka Patejuk and Adam Przepiórkowski. 2018. *From Lexical Functional Grammar to Enhanced Universal Dependencies: Linguistically informed treebanks of Polish*. Institute of Computer Science, Polish Academy of Sciences, Warsaw.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Philipp Petrenz. 2012. Cross-lingual genre classification. In *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 11–21, Avignon, France, April. Association for Computational Linguistics.

Barbara Plank and Gertjan van Noord. 2011. Effective measures of domain similarity for parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1566–1576, Portland, Oregon, USA, June. Association for Computational Linguistics.

Barbara Plank. 2016. What to do about non-standard (or non-canonical) language in nlp. In *KONVENS*, Bochum, Germany, September.

Ines Rehbein and Felix Bildhauer. 2017. Data point selection for genre-aware parsing. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 95–105, Prague, Czech Republic.

Rudolf Rosa. 2015. Parsing natural language sentences by semi-supervised methods. *Computing Research Repository*, arXiv: 1506.04897. version 1.

Sebastian Ruder and Barbara Plank. 2017. Learning to select data for transfer learning with Bayesian optimization. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 372–382, Copenhagen, Denmark, September. Association for Computational Linguistics.

Mohammad Sadegh Rasooli, Pegah Safari, Amirsaeid Moloodi, and Alireza Nourian. 2020. The Persian dependency treebank made universal. *arXiv e-prints*, pages arXiv–2009.

Alessio Salomoni. 2019. UD_German-LIT. `https://github.com/UniversalDependencies/UD_German-LIT`.

Stephanie Samson and Cagrı Cöltekin. 2020. UD_Tagalog-TRG. `https://github.com/UniversalDependencies/UD_Tagalog-TRG`.

Motoki Sato, Hitoshi Manabe, Hiroshi Noji, and Yuji Matsumoto. 2017. Adversarial training for cross-domain universal dependency parsing. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Vancouver, Canada. Association for Computational Linguistics.

Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*, volume 39. Cambridge University Press.

Serge Sharoff. 2007. Classifying web corpora into domain and genre using automatic feature identification. In *Proceedings of the 3rd Web as Corpus Workshop*, pages 83–94.

Tatiana Shavrina and Olga Shapovalova. 2017. To the methodology of corpus construction for machine learning:"Taiga" syntax tree corpus and parser. In *Proceedings of "CORPORA-2017" International Conference*, pages 78–84.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Chris Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Aaron Smith, Bernd Bohnet, Miryam de Lhoneux, Joakim Nivre, Yan Shao, and Sara Stymne. 2018. 82 treebanks, 34 models: Universal Dependency parsing with multi-treebank models. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 113–123, Brussels, Belgium, October. Association for Computational Linguistics.

Anders Søgaard. 2011. Data point selection for cross-language adaptation of dependency parsers. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 682–686, Portland, Oregon, USA, June. Association for Computational Linguistics.

Achim Stein and Sophie Prévost. 2013. Syntactic annotation of medieval texts. *New methods in historical corpora*, 3:275.

Will Styler. 2011. The Enronsent Corpus.

Sara Stymne. 2020. Cross-lingual domain adaptation for dependency parsing. In *Proceedings of the 19th International Workshop on Treebanks and Linguistic Theories*, pages 62–69, Düsseldorf, Germany, October. Association for Computational Linguistics.

Clara Vania, Yova Kementchedjhieva, Anders Søgaard, and Adam Lopez. 2019. A systematic comparison of methods for low-resource dependency parsing on genuinely low-resource languages. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1105–1116, Hong Kong, China, November. Association for Computational Linguistics.

Sappadla Prateek Veeranna, Jinseok Nam, Eneldo Loza Mencıa, and Johannes Fürnkranz. 2016. Using semantic similarity for multi-label zero-shot classification of text documents. In *Proceeding of European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges, Belgium: Elsevier*, pages 423–428.

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, Suntec, Singapore, August. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 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, Online, October. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *Computing Research Repository*, arXiv: 1609.08144. version 2.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3914–3923, Hong Kong, China, November. Association for Computational Linguistics.

Shorouq Zahra. 2020. Parsing low-resource Levantine Arabic: Annotation projection versus small-sized annotated data.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Gabrielė Aleksandravičiūtė, Ika Alfina, Lene Antonsen, Katya Aplonova, Angelina Aquino, Carolina Aragon, Maria Jesus Aranzabe, Bilge Nas Arıcan, Hórunn Arnardóttir, Gashaw Arutie, Jessica Naraiswari Arwidarasti, Masayuki Asahara, Deniz Baran Aslan, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Keerthana Balasubramani, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Starkaður Barkarson, Victoria Basmov, Colin Batchelor, John Bauer, Seyyit Talha Bedir, Kepa Bengoetxea, Gözde Berk, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Kristín Bjarnadóttir, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Anouck Braggaar, Kristina Brokaitė, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Lauren Cassidy, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Neslihan Cesur, Savas Cetin, Özlem Çetinoğlu, Fabricio Chalub, Shweta Chauhan, Ethan Chi, Taishi Chika, Yongseok Cho, Jinho Choi, Jayeol Chun, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Mihaela Cristescu, Philemon. Daniel, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Mehmet Oguz Derin, Elvis de Souza, Arantza Diaz de Ilarraza, Carly Dickerson, Arawinda Dinakaramani, Elisa Di Nuovo, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Hanne Eckhoff, Sandra Eiche, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Aline Etienne, Wograine Evelyn, Sidney Facundes, Richárd Farkas, Marília Fernanda, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Fabrício Ferraz Gerardi, Kim Gerdes, Filip Ginter, Gustavo Godoy, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Griciūtė, Matias Grioni, Loïc Grobol, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Tunga Güngör, Nizar Habash, Hinrik Hafsteinsson, Jan Hajič, Jan Hajič jr., Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Muhammad Yudistira Hanifmuti, Sam Hardwick, Kim Harris, Dag Haug, Johannes Heinecke, Oliver Hellwig, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Eva Huber, Jena Hwang, Takumi Ikeda, Anton Karl Ingason, Radu Ion, Elena Irimia, Ọlájídé Ishola, Kaoru Ito, Tomáš Jelínek, Apoorva Jha, Anders Johannsen, Hildur Jónsdóttir, Fredrik Jørgensen, Markus Juutinen, Sarveswaran K, Hüner Kaşıkara, Andre Kaasen, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Neslihan Kara, Boris Katz, Tolga Kayadelen, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Elena Klementieva, Arne Köhn, Abdullatif Köksal, Kamil Kopacewicz, Timo Korkiakangas, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Parameswari Krishnamurthy, Oğuzhan Kuyrukçu, Aslı Kuzgun, Sookyoung Kwak, Veronika Laippala, Lucia Lam, Lorenzo Lambertino, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phuong Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Maria Levina, Cheuk Ying Li, Josie Li, Keying Li, Yuan Li, KyungTae Lim, Bruna Lima Padovani, Krister Lindén, Nikola Ljubešić, Olga Loginova, Andry Luthfi, Mikko Luukko, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Büşra Marşan, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Hiroshi Matsuda, Yuji Matsumoto, Alessandro Mazzei, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Niko Miekka, Karina Mischenkova, Margarita Misirpashayeva, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, AmirHossein Mojiri Foroushani, Judit Molnár, Amirsaeid Moloodi, Simonetta Montemagni, Amir More, Laura Moreno Romero, Giovanni Moretti, Keiko Sophie Mori, Shinsuke Mori, Tomohiko Morioka, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Mariam Nakhlé, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Manuela Nevaci, Luong Nguyễn Thị, Huyền Nguyễn Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Alireza Nourian, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Adédayọ̀ Olúòkun, Mai Omura, Emeka Onwuegbuzia, Petya Osenova, Robert Östling, Lilja Øvrelid, Şaziye Betül Özateş, Merve Özçelik, Arzucan Özgür, Balkız Öztürk Başaran, Hyunji Hayley Park, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Angelika Peljak-Łapińska, Siyao Peng, Cenel-Augusto Perez, Natalia Perkova, Guy Perrier, Slav Petrov, Daria Petrova, Jason Phelan, Jussi Piitulainen, Tommi A Pirinen, Emily Pitler, Barbara Plank, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Peng Qi, Andriela Rääbis, Alexandre Rademaker, Taraka Rama, Loganathan Ramasamy, Carlos Ramisch, Fam Rashel, Mohammad Sadegh Rasooli, Vinit Ravishankar, Livy Real, Petru Rebeja, Siva Reddy, Georg Rehm, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Eiríkur Rögnvaldsson, Mykhailo Romanenko, Rudolf Rosa, Valentin Roșca, Davide Rovati, Olga Rudina, Jack Rueter, Kristján Rúnarsson, Shoval Sadde, Pegah Safari, Benoît Sagot, Aleksi Sahala, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Ezgi Sanıyar, Dage Särg, Baiba Saulīte, Yanin Sawanakunanon, Shefali Saxena, Kevin Scannell, Salvatore Scarlata, Nathan Schneider, Sebastian Schuster, Lane Schwartz, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Yana Shishkina, Muh Shohibussirri, Dmitry Sichinava, Janine Siewert, Einar Freyr Sigurðsson, Aline Silveira, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Maria Skachedubova, Aaron Smith, Isabela Soares-Bastos, Carolyn Spadine, Rachele Sprugnoli, Steinþór Steingrímsson, Antonio Stella, Milan Straka, Emmett Strickland,

Jana Strnadová, Alane Suhr, Yogi Lesmana Sulestio, Umut Sulubacak, Shingo Suzuki, Zsolt Szántó, Dima Taji, Yuta Takahashi, Fabio Tamburini, Mary Ann C. Tan, Takaaki Tanaka, Samson Tella, Isabelle Tellier, Marinella Testori, Guillaume Thomas, Liisi Torga, Marsida Toska, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Utku Türk, Francis Tyers, Sumire Uematsu, Roman Untilov, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Sowmya Vajjala, Rob van der Goot, Martine Vanhove, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Natalia Vlasova, Aya Wakasa, Joel C. Wallenberg, Lars Wallin, Abigail Walsh, Jing Xian Wang, Jonathan North Washington, Maximilan Wendt, Paul Widmer, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Mary Yako, Kayo Yamashita, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Arife Betül Yenice, Olcay Taner Yıldız, Zhuoran Yu, Zdeněk Žabokrtský, Shorouq Zahra, Amir Zeldes, Hanzhi Zhu, Anna Zhuravleva, and Rayan Ziane. 2021. Universal dependencies 2.8.1. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Appendix

A Universal Dependencies Setup

All experiments make use of Universal Dependencies v2.8 (Zeman et al., 2021). From the total set of 202 treebanks, we use all except for the following two (due to licensing restrictions): *Arabic-NYUAD* and *Japanese-BCCWJ*. In total 1.51 million sentences are used in our experiments.

Data Splits The experiments in Section 5 use the 204k global test split. Initial comparisons were performed on the 915k dev set. The 102k training split was used to fine-tune CLASS and BOOT. For early stopping, 31k sentences from the latter split were used as a held-out set. The exact instances are available in the associated code repository for future reproducibility.

Genre Mapping For 26 treebanks with instance-level genre labels in the metadata comments before each sentence, we created mappings from the treebank genre labels to the UD genre label set according to the guidelines described in Section 4.2. The genre metadata typically either follow the format `genre = X` or are implied by the document source specified in the sentence ID (e.g. `sent_id = genre-...`). There are a total of 91 mappings which will be made available with the codebase upon publication.

B Model and Training Details

The following describes architecture and training details for all methods. When not further defined, default hyperparameters are used. Implementations and predictions are available in the code repository at `https://personads.me/x/syntaxfest-2021-code`.

Infrastructure Neural models are trained on an NVIDIA A100 GPU with 40 GB of VRAM.

Language Model This work uses mBERT (Devlin et al., 2019) as implemented in the Transformers library (Wolf et al., 2020) as `bert-base-multilingual-cased`. Embeddings are of size $d_{\text{emb}} = 768$ and the model has 178 million parameters. To create sentence embeddings, we use the mean-pooled WordPiece embeddings (Wu et al., 2016) of the final layer.

Classification CLASS and BOOT build on the standard mBERT architecture as follows: mBERT $\rightarrow$ `CLS`-token $\rightarrow$ linear layer ($d_{\text{emb}} \times 18$) $\rightarrow$ softmax. The training has an epoch limit of 30 with early stopping after 3 iterations without improvements on the development set. Backpropagation is performed using AdamW (Loshchilov and Hutter, 2017) with a learning rate of 10^{-7} on batches of size 16. The fine-tuning procedure requires GPU hardware which can host mBERT, corresponding to 10 GB of VRAM. Training on the 71k relevant instances takes approximately 10 hours.

Clustering Both *Gaussian Mixture Models* (GMM) and *Latent Dirichlet Allocation* (Blei et al., 2003; LDA) use scikit-learn v0.23 (Pedregosa et al., 2011). LDA uses bags of character 3–6-grams which occur in at least 2 and in at most 30% of sentences. GMMs use the mBERT sentence embeddings as input. Both methods are CPU-bound and cluster all treebanks in UD in under 45 minutes.

Random Initializations Each experiment is run thrice using the seeds 41, 42 and 43.

C Additional Results

Table 2 shows results on the 915k development split of UD. Performance patterns are similar to those on the test split: the labeled clustering methods GMM+L and LDA+L perform best out of our proposed methods and outperform the baselines on the majority of metrics. With respect to classification, BOOT outperforms both the noisier CLASS and ZERO. Note that the frequency baseline FREQ performs especially well on the dev set, since only 5 of 26 instance labeled treebanks are included and 4 of these have the majority genre *news*.

METHOD	PUR	AGR	ΔBC	F1
FREQ	100±0.0	100±0.0	23±0.0	27±0.0
ZERO	43±0.0	66±0.0	50±0.0	5±0.0
CLASS	87±1.2	77±3.9	29±1.9	9±4.5
BOOT	95±0.2	100±0.0	24±0.3	16±1.0
GMM	92±0.1	55±5.5	30±0.7	—
+LABELS	100±0.0	100±0.0	5±0.1	17±1.6
LDA	88±1.0	42±2.2	30±0.2	—
+LABELS	100±0.0	100±0.0	5±0.0	15±0.9

Table 2: **Results of Genre Prediction on UD (Dev).** Purity (PUR ↑), agreement (AGR ↑), overlap error (ΔBC ↓) and micro-F1 over instance-labeled TBs (F1 ↑) for FREQ, ZERO, CLASS, BOOT and GMM, LDA with/without labels. Standard deviation denoted ±.