Data point selection for genre-aware parsing

Ines Rehbein
Leibniz ScienceCampus
Heidelberg/Mannheim
rehbein@cl.uni-heidelberg.de

Felix Bildhauer
Institut für Deutsche Sprache
Mannheim
bildhauer@ids-mannheim.de

Abstract

In the NLP literature, adapting a parser to new text with properties different from the training data is commonly referred to as *domain* adaptation. In practice, however, the differences between texts from different sources often reflect a mixture of *domain* and *genre* properties, and it is by no means clear what impact each of those has on statistical parsing. In this paper, we investigate how differences between articles in a newspaper corpus relate to the concepts of *genre* and *domain* and how they influence parsing performance of a transition-based dependency parser. We do this by applying various similarity measures for data point selection and testing their adequacy for creating genre-aware parsing models.

1 Introduction

The work of Biber (1988; 1995) and Biber & Conrad (2009) on language variation has brought valuable insights into the concepts of genre and register and the linguistic features that define them. It has also triggered many studies on genre classification (Kessler et al., 1997; Feldman et al., 2009; Passonneau et al., 2014), trying to automatically predict the genre or register for a particular text. However, despite the amount of work dedicated to genre prediction, the theoretical concept of *genre* remains vague and no agreement has been reached within the NLP (and linguistics) community on how to define it.[1]

This is even more surprising as concepts like *genre* and *domain* seem to be of crucial importance to our field and it is well known that the accuracy of NLP tools trained on one type of text will decrease noticeably when applying the same tools to another type of text with underlying properties that are different from the training data (Sekine, 1997; Gildea, 2001; McClosky et al., 2006). This might be due to either domain or genre differences, however, in NLP we usually refer to both as *out-of-domain* effects. While many studies have successfully shown how we can adapt tools to new domains (or genres) (Blitzer et al., 2006; Titov, 2011; Mitchell and Steedman, 2015), less is known about the underlying properties that are responsible for the decrease in performance. Out-of-domain (including out-of-genre) effects might be due to a large amount of unknown words introduced by topic shifts but might also be caused by a higher structural complexity in the data, by longer dependencies or a higher amount of non-projectivity.

Intuitively, we assume that domain differences can be captured by content-related features (e.g. from topic modelling) while we expect that functional differences between genres are reflected in structural features such as part-of-speech n-grams and other morpho-syntactic features. In the paper, we address these issues and investigate how differences between articles in a newspaper corpus relate to the concepts of *genre* and *domain* and how they impact parsing performance of a transition-based dependency parser.

[1]This is especially true for distinguishing *genre* from closely related concepts such as *register* and *text type*, which is why we will use *genre* in a broad sense here, i. e., as a cover term for *genre*, *register*, *text type* and similar.

Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT16), pages 95–105,
Prague, Czech Republic, January 23–24, 2018.

Keywords: genre and register variation, parser adaptation, dependency parsing

2 Related work

2.1 Register, Genre and Topic

It is hard to find a clear definition for concepts such as *register*, *genre* or *domain* in the literature.[2] We follow Biber and Conrad (2009) and consider *register* and *genre* not as different concepts but rather as different perspectives on the same thing. On this view, *genre* focusses on the "linguistic characteristics that are used to structure complete texts" (Biber and Conrad, 2009, p. 15). Passonneau et al. (2014) follow the functional view of Biber and Conrad and describe *genre* as a "set of shared regularities among written or spoken documents that enables readers, writers, listeners and speakers to signal discourse function, and that conditions their expectations of linguistic form". The *domain* concept, on the other hand, is orthogonal to the concept of *register* and *genre*. It reflects the main topic of the text (e. g., the sports domain) and can include texts from various genres with different communicative functions, such as soccer news, a report of a tennis match or an interview with a golf player.

While genre/register classification of documents can be a daunting task for humans, automatic genre/register classification of unrestricted text does not even reach 50% classification accuracy in recent state-of-the-art experiments (Biber and Egbert, 2015). One reason for this is, of course, the fact that there is no general consent about the number and boundaries of relevant categories to be included in a taxonomy of genres/registers. Moreover, as Petrenz and Webber (2011) point out, a text can not only have more than one topic but can also belong to multiple genres, which makes the *genre* concept even more complex and also casts some doubt on the validity of the task of genre classification on the document level. The authors discuss the correlation between genre and topic and show for a large newspaper corpus that there is a substantial correlation between the two, and that this correlation is not stable over time but undergoes significant changes. Petrenz and Webber (2011) also show that linguistic features that correlate with topic can decrease results in a genre prediction task. The authors argue that a meaningful evaluation of genre classification should thus control for topic, to avoid overly optimistic results that do not generalise to new texts with a topic distribution different from the one in the training data.

These observations are relevant also for adapting a parser to text from a new genre or domain, as most studies do not distinguish between content-based and structural features when measuring domain and genre similarity but use both evenhandedly. To our best knowledge, there are no studies on parser adaptation that try to separate domain from genre effects.

2.2 Adapting parsers to new genres and domains

Many parsing studies have addressed the problem of parser adaptation to new genres or domains, often focussed on adapting a Penn treebank-trained parser to biomedical text or to web data.[3] Different techniques have been tested for parser adaptation, such as transformations applied to the target data (Foster, 2010), ensemble parsing (Dredze et al., 2007) or co-training (Baucom et al., 2013). Other studies have tried to distinguish between features specific to the source data and general features that also occur in the target data (Dredze et al., 2007), or to create domain- or genre-specific parsing models and select the model combination that most probably will maximise parsing scores on the target data (McClosky et al., 2010; Plank and Sima'an, 2008). Plank and van Noord (2011) and Mukherjee et al. (2017) create new training sets that reflect the distribution in the target data by identifying the source data most similar to the target, based on measures that assess structural or topic similarity between both.

Features used in these experiments (McClosky et al., 2010; Plank and van Noord, 2011; Mukherjee et al., 2017) include known and unknown words, character n-grams and LDA topics but do not (or only implicitly) capture *structural* similarity. The authors show that content and surface features are successful in selecting appropriate training data for the new domain and also work better than using genre labels assigned by humans (Plank and van Noord, 2011). Søgaard (2011), however, has shown that data point selection based on structural similarity can improve parsing accuracy significantly in a cross-lingual parser adaptation setting and Rehbein (2011) shows a similar effect for in-domain self-training. Based

[2] A full survey of work on register, genre or domain variation is beyond the scope of this paper. We refer to Biber (1988) and especially Lee (2001) for a review of how these terms have been used in various theoretical frameworks.

[3] See, e. g., the CoNLL 2007 Shared Task on Domain Adaptation (Nivre et al., 2007) and the SANCL 2012 Shared Task on Parsing the Web (Petrov and McDonald, 2012).

	# articles	# sent	# token	avg. sent length	# sent train pool	# sent testset
portrait	42	1,195	24,035	20.1	695	500
letter	102	1,789	34,923	19.5	1,289	500
documentary	72	3,162	61,534	19.5	2,662	500
agency	617	5,278	84,944	16.1	4,750	528
interview	102	7,585	120,215	15.8	6,826	759
commentary	333	9,613	178,347	18.5	8,652	961
taz report	2,376	66,973	1,283,803	19.2	66,973	–

Table 1: Distribution of different genres in the TüBa-D/Z and training/test sizes.

on these results, we are interested in comparing the adequacy of *surface* and *content* features for data point selection with features that capture *structural* similarity in the data.

We evaluate the features in a setting where we try to improve the performance of a dependency parser on different genres in a newspaper corpus by training genre-aware parsing models. We would like to know whether the different feature types capture similar properties in the data. We consider content-related features to be characteristic for certain domains while we expect that functional differences between genres are reflected in structural differences between texts and can be captured by features such as part-of-speech n-grams. In addition, we compare the potential of content and structural features to measure domain and genre similarity with linguistically defined features, inspired by the work of Biber (1988; 1995) on register variation.

3 Experiments

In our experiments, we use the TüBa-D/Z treebank (Telljohann et al., 2004), a corpus of German newspaper text from the *taz*, a German daily newspaper, that includes more than 95,000 sentences annotated with constituency trees and grammatical function labels. The data has been automatically converted to dependencies. Webber (2009) has shown for the Penn treebank (Marcus et al., 1994) that even newspaper corpora should not be considered as homogeneous objects but typically also consist of multiple genres. Similar to the Penn treebank, the TüBa-D/Z (v10) includes articles from a variety of genres. The genre labels in the TüBa-D/Z have been assigned by the editors of the *taz* and are: *reports, commentaries, documentaries, letters to the editor, interviews, portraits* and messages from *news agencies*. It is, however, not clear to what extend these labels correspond to linguistically well-defined categories, i. e., whether documents within a specific genre category share "linguistic characteristics that are used to structure complete texts" as suggested by Biber and Conrad (2009). The vast majority of the articles in the treebank is labelled as *taz reports* (Table 1).

3.1 Genre differences

The first question we are interested in is whether we can cluster the data according to the labels assigned by the *taz*, to see whether these labels reflect systematic linguistic differences in the data. For this, we divide the data into genre-specific samples of 10,000 tokens each. First, we concatenate all articles from the same genre and split them into smaller samples of 10,000 tokens, so that sentences from the same article end up in the same sample most of the time. Then we run a Principle Components Analysis (PCA), a) based on the frequency of POS tags in the data (Figure 1 left), and b) based on topic distributions from an LDA (Figure 1 right).[4]

Figure 1 (left) shows crucial differences between samples taken from articles that have been assigned different labels. Interviews and agency messages are separated from the other samples along the second principal component (Dim2) while both can be separated from each other along the first component (Dim1). Reports and documentaries are positioned more central, with the reports a bit more to the left and the documentaries a bit more to the right. The commentaries cluster together in the lower part of the space and the letters are at the boundary between documentaries and commentaries. While we can

[4]For topic modelling we use the Mallet implementation from `http://mallet.cs.umass.edu/` and learn 100 topics on the lemmatised version of the TüBa-D/Z. We compute the PCA using the R PCA function from the `FactoMineR` package. To increase readability, we only include the first 20 samples from the *report* genre.

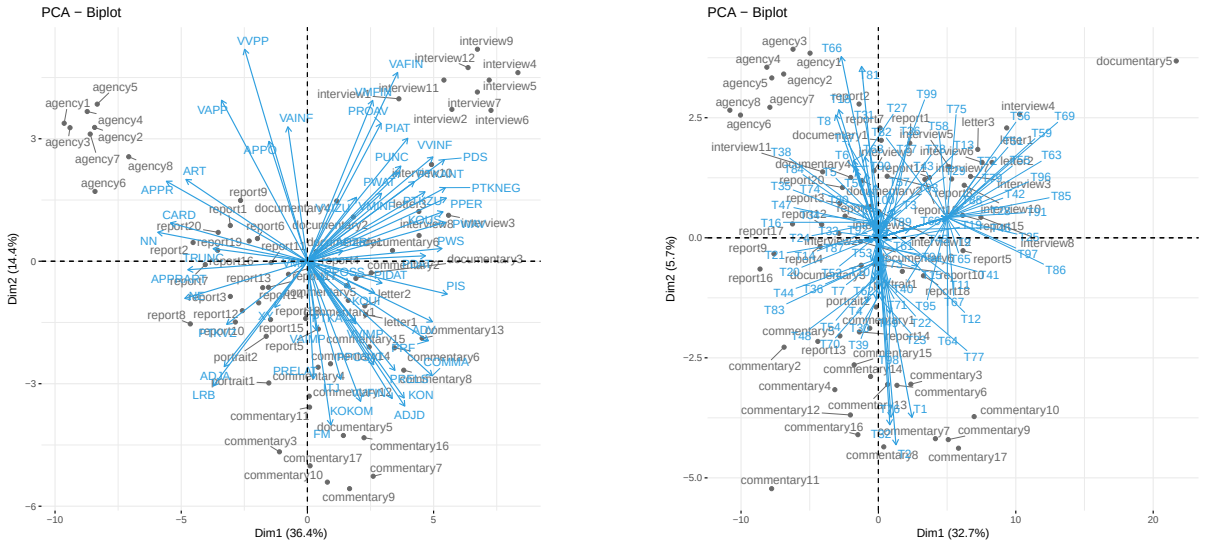

Figure 1: PCA based on frequency of POS (left) and on topic distributions (right) from LDA topic modelling in samples of 10,000 tokens from the TüBa-D/Z.

observe strong tendencies, the distinction between these samples is not as pronounced as the one between the interviews and the agency messages.

Most interestingly, we can see similar trends for the PCA based on the LDA topics (Figure 1, right). The most important difference, however, is that topic-wise we observe a similarity between the interviews and the letters while in the PCA based on POS tags the letters are positioned between the commentaries. The PCA shows the strong correlation between topics and genres that has already been pointed out by Petrenz and Webber (2011). It also shows that the labels assigned by the *taz* correspond to systematic differences between the texts and can be used at least as an approximation to linguistically defined genres.

3.2 Impact of genre differences on parsing

Given that we are able to discriminate documents from different genres based on the distribution of POS in the data, we expect that the genre differences also impact parsing accuracy. To investigate this, we split the texts into a pool of training data and test data as follows. From each genre, we create test sets with 10% of the tokens for this genre or, for genres with less than 50,000 sentences, we select 500 sentences from the pool for the test set. For the test sets we also control for topic by selecting articles so that the similarity with regard to topic distribution is maximised.[5] The rest of the data is used as a pool from which we create different-sized training sets (see Table 1).

For the first experiment, we create training sets of size $N = \{10000, 20000, 380000\}$ tokens by *randomly* selecting articles from the *report* data that constitute the largest part of the training pool.[6] We would like to know whether we can observe systematic differences between the genres with regard to their "parsability", i. e., how hard it is to predict the correct parse. We train the IMSTrans parser (Björkelund and Nivre, 2015), a transition-based dependency parser, on the randomly extracted training sets and report LAS for the different genres. All results are based on automatic POS and morphological tags predicted by Marmot (Mueller et al., 2013)[7] and include punctuation in the evaluation (Table 2). We report average LAS and standard deviation (σ) over 5 runs.

As expected, we observe substantial differences in parsing scores between the genres. Over all sample sizes, agency messages achieve the highest parsing accuracy, followed by portraits and documentaries while letters and commentaries seem to be more difficult to parse.

[5]We compute the topic distribution for articles in the TüBa-D/Z, based on LDA and then compute the Manhattan distance between the topic distribution for each pair of articles from the same genre. Then we select articles for each genre in the test set so that the accumulated distance between the articles for each genre is minimised.

[6]We count tokens instead of sentences as the differences in sentence length between the genres would impact results.

[7]We also use predicted POS/morphological information in the training data. We use the pre-trained SPMRL models kindly provided by the developers: http://cistern.cis.lmu.de/marmot/models/CURRENT.

size (token)		agency	commentary	documentary	interview	letter	portrait
10,000	*avg.*	85.65	78.47	79.20	78.64	77.61	80.06
	σ	0.31	0.47	0.31	0.23	0.36	0.33
50,000	*avg.*	89.87	83.31	84.61	83.95	83.08	85.66
	σ	0.24	0.26	0.24	0.17	0.21	0.28
380,000	*avg.*	93.15	87.71	89.40	88.41	87.96	89.58
	σ	0.22	0.13	0.17	0.14	0.28	0.12

Table 2: LAS for different genres: ***random*** training sets from reports (avg. LAS over 5 runs and std dev.)

3.3 Genre effects as an out-of-domain problem

In the next set of experiments, we investigate whether the performance changes when we train the parser on the same amount of data, but this time using sentences from the same genre (according to the *taz* labels) as in the test set. In other words, we would like to know whether the differences in parsing accuracy reflect an out-of-domain problem and will vanish when we train on "in-domain" (or rather, *in-genre*) data. As before, we randomly select articles from the pool of training data, but now we control for genre. This is possible only for the smaller training set sizes ($N = 2500, 4500$) and we also have to exclude the genres for which we do not have enough data in the pool (*letter*, *portrait*).

Table 3 shows the improvements we obtain when training the parser on data from the same genre, as opposed to training it on a randomly selected dataset from the *taz* reports. An exception are the *documentaries* which seem to be closer to the *reports* (see figure 1) so that the effect of training on in-genre data is levelled out. As before, we note substantial differences between the parsing scores for the different genres. This shows that the gap in results is not due to missing in-domain (or *in-genre*) training data but that certain genres are in fact harder to parse than others. To find out what it is that makes agency texts so much easier to parse than the commentaries and letters, we compare linguistic properties of the texts in the different genres that have been associated with syntactic complexity and parsing difficulty in the literature (Roark et al., 2007; McDonald and Nivre, 2007; Gulordava and Merlo, 2016).

size (token)		agency	commentary	documentary	interview	letter	portrait
10,000	*random*	85.65	78.47	**79.20**	78.64	77.61	80.06
10,000	*in-domain*	**86.30**	**79.01**	79.13	**78.83**	**79.53**	**80.52**
50,000	*random*	89.87	83.31	84.61	83.95	83.08	85.66
50,000	*in-domain*	**90.68**	**83.80**	**84.82**	**84.47**	n.a.	n.a.

Table 3: LAS for *random* training sets from reports and *in-domain* training sets (avg. LAS over 5 runs).

Table 4 shows the average sentence length, the number of finite verbs per sentence (as an approximation of the complexity of the sentence structure), the number of unknown words, the average dependency length, the average entropy in arc direction (Liu, 2010) (whether the head of a dependent is found to its left, to its right, or can be positioned either way), and the percentage of non-projective sentences in the test sets. When fitting a linear regression model to the data, arc direction entropy was the only significant predictor for parsing accuracy (β = -2.44, p <.01). However, given the small size of the test sets we used, we prefer to consider our results as preliminary pending confirmation on larger data sets.

genre	LAS	sent.len	Vfin/sent	# unk	dep.len	arc.ent	non-proj
agency	93.15	16.1	1.2	16.3	2.7	13.1	7.0
portrait	89.58	19.5	1.8	16.2	2.6	14.7	9.2
documentary	89.40	21.9	1.6	14.7	3.1	14.9	13.0
interview	88.41	19.2	1.5	19.0	2.7	15.1	10.0
letter	87.96	17.7	1.5	13.4	2.6	15.1	11.1
commentary	87.71	18.3	1.6	12.9	2.8	14.7	10.2

Table 4: Differences between test sets (avg. sentence length, no. of finite verbs per sentence, % of unknown tokens, avg. dependency length, entropy of arc direction, % of non-projective trees) and LAS for each test set when training the parser on a randomly selected training set from reports (N =380,000).

	setting	feature types: *surface*	*structural*	*content*	*linguistic*	raw data	description	no. of features
Exp01	KNS	✓				✓	Text statistics and word frequencies	33
Exp02	POS n-gram		✓			✓	LM perplexity based on pos n-grams	n.a.
Exp03	LDA topics			✓			Distance from topic distribution	100
Exp04	COReX	✓			✓		Text statistics, morpho-syntactic features	40

Table 5: Overview of the settings and features used for data point selection.

3.4 Data point selection for genre-aware parsing

We now explore whether we can train genre-aware parsing models on larger data by selecting out-of-genre data points that are similar to the *target* genre. To that end, we test the adequacy of different feature types for measuring similarity.

Kessler et al. (1997) (KNS) have obtained consistently good results for genre prediction across topics (Petrenz and Webber, 2011), based only on *surface* features.[8] Plank and van Noord (2011) and Mukherjee et al. (2017) have trained domain-specific parsing models based on *content* (LDA topics) and *surface* features (word frequencies and character n-gram frequencies). In our experiments, we would like to compare the adequacy of *content* and *surface* features with data selection based on *structural* similarity (where similarity is operationalised as the perplexity of a language model (LM) based on POS n-grams) and features that take into account the *linguistic* properties of a text, relying on Biber-style features. Table 5 gives an overview over the different settings and features used for data selection.

3.5 Selecting the training sets

For our different settings, we select training data from the pool as follows. For the structural setting, we make use of additional unannotated data from the *taz* with articles from 1989-1999, with information on article boundaries and genre labels. We select all articles from 1989, 1992, 1995, 1997, 1999 and remove those that are included in the TüBa-D/Z treebank from the raw text corpus. We automatically predict POS tags and lemma forms, using the Treetagger (Schmid, 1994) with the standard parameter file provided by the developer.

Exp01 We create a version of the raw newswire data where we replace all words with their POS and divide the data so that we have one sample per genre. We use the CMU SLM toolkit (Clarkson and Rosenfeld, 1997) to train a LM for each genre, based on POS n-grams in the samples. Then we compute the perplexity for each article in the training pool of the TüBa-D/Z and assign articles to the genre for which they show the lowest perplexity, i. e., to which they are most similar. Based on this, we extract one training set per genre with N=380,000 tokens from the most similar articles. Please note that the articles do not need to come from the same genre but only need to be similar to the raw text files in this genre.

Exp02 For the next experiment, we extract the features described in Kessler et al. (1997) from the large, POS tagged newswire corpus. We aggregate the scores for each feature over all files from the same genre and normalise by the number of articles. Then we extract the same features from the articles in the TüBa-D/Z training pool and compute the similarity of each article in the training pool to the genres in the raw text corpus, based on the aggregated feature scores, using the Manhattan distance as similarity measure. We select the most similar articles for each genre and extract the first N=380,000 tokens for the genre-specific training sets.

Exp03 For topic modelling, we use a lemmatised version of all articles in the TüBa-D/Z data. The test data for each genre has been merged into one document per genre while the articles from the training pool are included as separate documents (each article is one document).[9] We use the topic distributions for each document to compute the similarity of each article in the training pool to the different genre testsets, and create genre-specific training sets by selecting the most similar articles for each test set. As a similarity measure we use the Manhattan distance. Please note that –in contrast to the other settings–

[8] We reimplemented the KNS features (text statistics and frequencies for particular word forms) described at `http://homepages.inf.ed.ac.uk/s0895822/SCTG/features.html` for German.

[9] We set the number of topics to N=100. We use standard settings und remove stopwords but do not lowercase the lemmas.

	Setting	agency	commentary	documentary	interview	letter	portrait
Baseline	*random*	93.15	87.71	89.40	88.41	87.96	89.58
Exp01	POS n-gram LM	93.38	87.68	89.40	88.61	88.57*	89.45
Exp02	KNS	**93.83***	87.52	89.26	88.31	88.50*	89.32
Exp03	LDA topics	93.67*	**88.19***	89.61	89.15**	**88.60***	**90.05**
Exp04	COReX	93.70*	87.93	**89.80**	**89.25****	88.14	89.50

Table 6: Results for different data selection methods for genre-aware parsing (LAS; * indicates a significant improvement over the baseline with $p < 0.05$; ** with $p < 0.01$).

here we directly maximise the similarity between training and test data, while in the other settings we use the unannotated newspaper data as a proxy for determining genre similarity. This means that a direct comparison of the results might not be fair. On the other hand, this approach allows us to investigate the interaction between topic and genre. We will get back to this issue in Section 3.6.

Exp04 For the COReX setting, training sets are created based on the linguistic properties of each genre. We use a fine-grained set of 40 linguistic features obtained from COReX (Bildhauer and Schäfer, 2017), a framework for lexico-grammatical document annotation for large German corpora. The COReX features we use inlcude frequencies for POS, morphosyntactic features, named entity-based features and stylistic markers in the text, inspired by Biber (1988). We extract these features for each article in the training pool[10] and aggregate the scores over all articles in the same genre. We then compute the similarity (or, rather, the distance) of the feature vector for one article to the aggregated vectors for each genre, again using Manhattan distance as similarity measure. For each genre, we select the documents that showed the highest genre similarity, based on the distance between the feature vectors, and create new training sets for each genre with N=380,000 tokens.

We create genre-aware parsing models by training the parser on the different datasets and evaluate the different models on the test sets from each genre (based on the human-assigned genre labels, see Table 1). Another possible approach would be to select the best model for each text that we want to parse based on its similarity to the different training sets. We refrain from doing so as extracting the COReX features is costly and includes several preprocessing steps such as POS and morphological tagging, topological parsing and NER. In our setup, we only have to run the pipeline once for creating the genre-aware parsing models. Doing the same again for each text that we want to parse seems exorbitant. Our setup, however, assumes that we have genre information for the texts we want to parse. In a different scenario where genre labels are not given we could either refer to the COReX pipeline or test how far we get when using similarity measures based on simple POS and surface frequencies. We leave this to future work.

3.6 Results for genre-aware parsing

Table 6 shows parsing results for the genre-aware models based on different similarity measures. The *structural* model (Exp01) fails to outperform the random baseline for all genres but the letters.[11] The KNS model (Exp02) that is based on *surface* features gives a significant improvement for agency messages and letters but also fails to improve LAS for the other genres. Most interestingly, the topic setting (Exp03) is the only model where we see an improvement over the baseline for *all* genres. The COReX features improve results for nearly all genres, however not always significantly.

As already pointed out above, the success of the topic model might be due to the fact that we directly optimised the selection of training instances based on their similarity to the articles in the test set (and not, as done for the other settings, by approximating genres via unlabelled data (Exp01-02) or by computing similarity against an averaged score obtained from all articles in the treebank (Exp04).

We thus run another experiment where we create additional test data for each genre by selecting articles from this particular genre but with a topic distribution different from the one in the previous test data. We do this by selecting new articles so that we maximise the Manhattan distance to the articles in the old test sets. Then we use the same parsing models (Exp03) to parse the data and compare the results to the

[10]Feature counts are normalised per 1,000 words.
[11]For significance testing we use Bikel's Randomized Parsing Evaluation Comparator.

size (token)		agency	commentary	documentary	interview	letter	portrait
380,000	*random*	93.80	**90.62**	**90.44**	92.54	87.13	90.26
	σ	0.09	0.17	0.27	0.09	0.13	0.29
380,000	*topic*	**94.54***	90.37	90.15	92.55	**87.71**	**90.46**

Table 7: LAS for baseline (*random* training sets from reports; avg. LAS over 5 runs and standard deviation) and for *topic* setting where data selection is *not* optimised on the test set.

ones we get when parsing the new test sets with the baseline parsing models from Exp01.

Table 7 shows that the parsing models we trained based on topic similarity do not necessarily generalise well to other data from the same genre. Only for agency messages results improve, while for all other genres results are in the same range or even decrease.

3.7 Discussion

Our results showed that our initial hypothesis about the *structural* similarity being more suitable for capturing genre similarity than *surface* and *content* features does not seem to bear out. We take this as evidence that the concept of genre can not easily be defined (or reduced to) structural properties in the texts, at least not in the way as operationalised in our experiments.

We also showed that data selection based on LDA topics in the data can improve parsing scores, as has been shown before by Plank and van Noord (2011) and Mukherjee et al. (2017). This approach, however, requires to compute topics over the joint training and test data which might not always be possible in practice. In addition, our experiments showed that while there is a correlation between topic and genre, the topics we learn are by no means representative for a particular genre. Our results are in line with the results of Petrenz and Webber (2011) for genre prediction. We thus argue that LDA topic modelling might be appropriate for domain adaptation for highly diverse sources such as biomedical data and data from the newswire. For more homogeneous source texts as we have in our setup, however, relying on content similarity might not be the right approach.

This brings us back to our original research question: How can we model genre distinctions for parsing? So far, our experiments showed that distinguishing genre from domain is by no means an easy task. We argue that the human-labelled categories in the TüBa-D/Z reflect both, genre and domain properties, and both seem to have an impact on parsing. We also showed that content similarity based on LDA topics might be useful for parser adaptation to new domains but not for adapting the parser to a new genre.

4 Conclusions

We presented an approach to genre-aware parsing where we only have access to small amounts of annotated training data for each genre. Our approach tests several ways to operationalise similarity and makes use of large unannotated data to learn genre-specific distributions of features. Based on this, we extract training sets for each genre by selecting sentences from the pool of annotated training data that are similar to the target genre. We computed similarity based on surface features, structural features, text topic and fine-grained linguistic features, and showed that different feature types work best for different datasets. We take that as evidence that for parser adaptation we have to deal with a mixture of genre and domain effects, and to obtain optimal results we need to model both. However, using content features such as topics for modelling genre similarity might be dangerous as those features do not generalise well.

In future work we would like to test our approach in a setting where no human-assigned genre labels are available, and also apply self-training to extend the training data size for genre-aware parsing.

Acknowledgments

This research has been partially supported by the Leibniz Science Campus "Empirical Linguistics and Computational Modeling", funded by the Leibniz Association under grant no. SAS-2015-IDS-LWC and by the Ministry of Science, Research, and Art (MWK) of the state of Baden-Württemberg.

References

Eric Baucom, Levi King, and Sandra Kübler. 2013. Domain adaptation for parsing. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*. Hissar, Bulgaria, pages 56–64.

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

Douglas Biber. 1995. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge: Cambridge University Press.

Douglas Biber and Susan Conrad. 2009. *Register, genre, and style*. Cambridge: Cambridge University Press.

Douglas Biber and Jesse Egbert. 2015. 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(1):3–36.

Felix Bildhauer and Roland Schäfer. 2017. COReX und COReCO: A lexico-grammatical document annotation framework for large German corpora. Poster at the Computational linguistics poster session at the DGfS 2017.

Anders Björkelund and Joakim Nivre. 2015. Non-deterministic oracles for unrestricted non-projective transition-based dependency parsing. In *Proceedings of the 14th International Conference on Parsing Technologies*. Bilbao, Spain, pages 76–86.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. Sydney, Australia, EMNLP '06, pages 120–128.

Philip Clarkson and Ronald Rosenfeld. 1997. Statistical language modeling using the CMU-Cambridge toolkit. In *ESCA Eurospeech*. pages 2707–2710.

Mark Dredze, John Blitzer, Partha Pratim Talukdar, Kuzman Ganchev, João V. Graça, and O Pereira. 2007. Frustratingly hard domain adaptation for dependency parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*. EMNLP-CoNLL.

S. Feldman, M. A. Marin, M. Ostendorf, and M. R. Gupta. 2009. Part-of-speech histograms for genre classification of text. In *Proceedings of the 2009 International Conference on Acoustics, Speech and Signal Processing*. Washington, DC, IEEE'09, pages 4781–4784.

Jennifer Foster. 2010. "cba to check the spelling" investigating parser performance on discussion forum posts. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Los Angeles, California, HLT '10, pages 381–384.

Daniel Gildea. 2001. Corpus variation and parser performance. In *Empirical Methods in Natural Language Processing*. EMNLP '01, pages 167–202.

Kristina Gulordava and Paola Merlo. 2016. Multi-lingual dependency parsing evaluation: a large-scale analysis of word order properties using artificial data. *TACL* 4:343–356.

Brett Kessler, Geoffrey Nunberg, and Hinrich Schütze. 1997. Automatic detection of text genre. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*. Morristown, NJ, ACL'97, pages 32–38.

David Yw Lee. 2001. Genres, registers, text types, domains and styles: clarifying the concepts and navigating a path through the bnc jungle. *Technology* 5:37–72.

Haitao Liu. 2010. Dependency direction as a means of word-order typology: A method based on dependency treebanks. *Lingua* 120(6):1567–1578.

Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The penn treebank: Annotating predicate argument structure. In *Proceedings of the Workshop on Human Language Technology*. Plainsboro, NJ, HLT '94, pages 114–119.

David McClosky, Eugene Charniak, and Mark Johnson. 2006. Reranking and self-training for parser adaptation. In *Proceedings of the International Conference on Computational Linguistics and the Association for Computational Linguistics*. Sydney, Australia, COLING-ACL, pages 337–344.

David McClosky, Eugene Charniak, and Mark Johnson. 2010. Automatic domain adaptation for parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Los Angeles, California, HLT '10, pages 28–36.

Ryan McDonald and Joakim Nivre. 2007. Characterizing the errors of data-driven dependency parsing models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. EMNLP-CoNLL '07, pages 122–131.

Jeff Mitchell and Mark Steedman. 2015. Parser adaptation to the biomedical domain without re-training. In *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis*. Lisbon, Portugal, pages 79–89.

Thomas Mueller, Helmut Schmid, and Hinrich Schütze. 2013. Efficient higher-order CRFs for morphological tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Seattle, Washington, USA, pages 322–332. http://www.aclweb.org/anthology/D13-1032.

Atreyee Mukherjee, Sandra Kübler, and Matthias Scheutz. 2017. Creating POS tagging and dependency parsing experts via topic modeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Valencia, Spain, EACL'17, pages 347–355.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The CoNLL 2007 Shared Task on Dependency Parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*. Prague, Czech Republic, pages 915–932.

Rebecca J. Passonneau, Nancy Ide, Songqiao Su, and Jesse Stuart. 2014. Biber Redux: Reconsidering Dimensions of Variation in American English. In *Proceedings of the 25th International Conference on Computational Linguistics*. COLING'14, pages 565–576.

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

Slav Petrov and Ryan McDonald. 2012. Overview of the 2012 shared task on parsing the web. Notes of the First Workshop on Syntactic Analysis of Non-Canonical Language (SANCL).

Barbara Plank and Khalil Sima'an. 2008. Subdomain sensitive statistical parsing using raw corpora. In European Language Resources Association (ELRA), editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco.

Barbara Plank and Gertjan van Noord. 2011. Effective measures of domain similarity for parsing. In *Proceedings of the Second International Conference on Human Language Technology Research*. San Francisco, CA, pages 82–86.

Ines Rehbein. 2011. Data point selection for self-training. In *Proceedings of the Second Workshop on Statistical Parsing of Morphologically Rich Languages*. Dublin, Ireland, SPMRL '11, pages 62–67.

Brian Roark, Margaret Mitchell, and Kristy Hollingshead. 2007. Syntactic complexity measures for detecting mild cognitive impairment. In *Proceedings of the ACL 2007 Workshop on Biomedical Natural Language Processing*. BioNLP, pages 1–8.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*. Manchester, UK, pages 44–49.

Satoshi Sekine. 1997. The domain dependence of parsing. In *Applied Natural Language Processing*. ANLP '01, pages 96–102.

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: Short Papers - Volume 2*. Portland, Oregon, HLT '11, pages 682–686.

Heike Telljohann, Erhard Hinrichs, and Sandra Kübler. 2004. The Tüba-D/Z Treebank: Annotating German with a Context-Free Backbone. In *In Proceedings of the Fourth International Conference on Language Resources and Evaluation*. Lisbon, Portugal, LREC'04, pages 2229–2235.

Ivan Titov. 2011. Domain adaptation by constraining inter-domain variability of latent feature representation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*. Portland, Oregon, HLT '11, pages 62–71.

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: Volume 2 - Volume 2*. Association for Computational Linguistics, Suntec, Singapore, ACL '09, pages 674–682.