Fine-grained Genre Classification using Structural Learning Algorithms

Zhili Wu
Centre for Translation Studies
University of Leeds, UK
z.wu@leeds.ac.uk

Katja Markert
School of Computing
University of Leeds, UK
scskm@leeds.ac.uk

Serge Sharoff
Centre for Translation Studies
University of Leeds, UK
s.sharoff@leeds.ac.uk

Abstract

Prior use of machine learning in genre classification used a list of labels as classification categories. However, genre classes are often organised into hierarchies, e.g., covering the subgenres of fiction. In this paper we present a method of using the hierarchy of labels to improve the classification accuracy. As a testbed for this approach we use the Brown Corpus as well as a range of other corpora, including the BNC, HGC and Syracuse. The results are not encouraging: apart from the Brown corpus, the improvements of our structural classifier over the flat one are not statistically significant. We discuss the relation between structural learning performance and the visual and distributional balance of the label hierarchy, suggesting that only balanced hierarchies might profit from structural learning.

1 Introduction

Automatic genre identification (AGI) can be traced to the mid-1990s (Karlgren and Cutting, 1994; Kessler et al., 1997), but this research became much more active in recent years, partly because of the explosive growth of the Web, and partly because of the importance of making genre distinctions in NLP applications. In Information Retrieval, given the large number of web pages on any given topic, it is often difficult for the users to find relevant pages that are in the right genre (Vidulin et al., 2007). As for other applications, the accuracy of many tasks, such as machine translation, POS tagging (Giesbrecht and Evert, 2009) or identification of discourse relations (Webber, 2009) relies of defining the language model suitable for the genre of a given text. For example, the accuracy of POS tagging reaching 96.9% on newspaper texts drops down to 85.7% on forums (Giesbrecht and Evert, 2009), i.e., every seventh word in forums is tagged incorrectly.

This interest in genres resulted in a proliferation of studies on corpus development of web genres and comparison of methods for AGI. The two corpora commonly used for this task are KI-04 (Meyer zu Eissen and Stein, 2004) and Santinis (Santini, 2007). The best results reported for these corpora (with 10-fold cross-validation) reach 84.1% on KI-04 and 96.5% accuracy on Santinis (Kanaris and Stamatatos, 2009). In our research (Sharoff et al., 2010) we produced even better results on these two benchmarks (85.8% and 97.1%, respectively). However, this impressive accuracy is not realistic *in vivo*, i.e., in classifying web pages retrieved as a result of actual queries. One reason comes from the limited number of genres present in these two collections (eight genres in KI-04 and seven in Santinis). As an example, only front pages of online newspapers are listed in Santinis, but not actual newspaper articles, so once an article is retrieved, it cannot be assigned to any class at all. Another reason why the high accuracy is not useful concerns the limited number of sources in each collection, e.g., all FAQs in Santinis come from either a website with FAQs on hurricanes or another one with tax advice. In the end, a classifier built for FAQs on this training data relies on a high topic-genre correlation in this particular collection and fails to spot any other FAQs.

There are other corpora, which are more diverse in the range of their genres, such as the fifteen genres of the Brown Corpus (Kučera and Francis, 1967) or the seventy genres of the BNC (Lee, 2001), but because of the number of genres in them and the diversity of documents within each genre, the accuracy of prior work on these collections is much less impressive. For example, Karlgren and Cutting (1994) using linear discriminant analysis achieve an accuracy of 52% without us-

Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 749–759,
Uppsala, Sweden, 11-16 July 2010.

ing cross-validation (the entire Brown Corpus was used as both the test set and training set), with the accuracy improving to 65% when the 15 genres are collapsed into 10, and to 73% with only 4 genres (Figure 1). This result suggests the importance of the hierarchy of genres. Firstly, making a decision on higher levels might be easier than on lower levels (fiction or non-fiction rather than science fiction or mystery). Secondly, we might be able to improve the accuracy on lower levels, by taking into account the relevant position of each node in the hierarchy (distinguishing between `reportage` or `editorial` becomes easier when we know they are safely under the category of `press`).

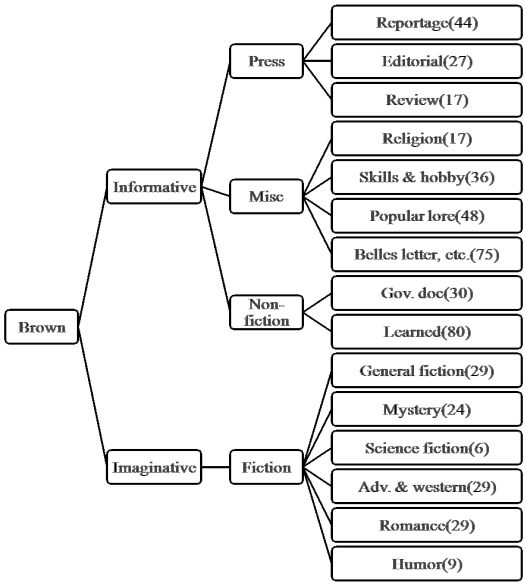

Figure 1: Hierarchy of Brown corpus.

This paper explores a way of using information on the hierarchy of labels for improving fine-grained genre classification. To the best of our knowledge, this is the first work presenting structural genre classification and distance measures for genres. In Section 2 we present a structural reformulation of Support Vector Machines (SVMs) that can take similarities between different genres into account. This formulation necessitates the development of distance measures between different genres in a hierarchy, of which we present three different types in Section 3, along with possible estimation procedures for these distances. We present experiments with these novel structural SVMs and distance measures on three different corpora in Section 4. Our experiments show that structural SVMs can outperform the non-structural standard. However, the improvement is only statistically significant on the Brown corpus. In Section 5 we investigate potential reasons for this, including the (im)balance of different genre hierarchies and problems with our distance measures.

2 Structural SVMs

Discriminative methods are often used for classification, with SVMs being a well-performing method in many tasks (Boser et al., 1992; Joachims, 1999). Linear SVMs on a flat list of labels achieve high efficiency and accuracy in text classification when compared to nonlinear SVMs or other state-of-the-art methods. As for structural output learning, a few SVM-based objective functions have been proposed, including margin formulation for hierarchical learning (Dekel et al., 2004) or general structural learning (Joachims et al., 2009; Tsochantaridis et al., 2005). But many implementations are not publicly available, and their scalability to real-life text classification tasks is unknown. Also they have not been applied to genre classification.

Our formulation can be taken as a special instance of the structural learning framework in (Tsochantaridis et al., 2005). However, they concentrate on more complicated label structures as for sequence alignment or parsing. They proposed two formulations, slack-rescaling and margin-rescaling, claiming that margin-rescaling has two disadvantages. First, it potentially gives significant weight to output values that might not be easily confused with the target values, because every increase in the loss increases the required margin. However, they did not provide empirical evidence for this claim. Second, margin rescaling is not necessarily invariant to the scaling of the distance matrix. We still used margin-rescaling because it allows us to use the sequential dual method for large-scale implementation (Keerthi et al., 2008), which is not applicable to the slack-rescaling formulation. For web page classification we will need fast processing. In addition, we performed model calibration to address the second disadvantage (distance matrix invariance).

Let $\mathbf{x}$ be a document and $\mathbf{w}_m$ a weight vector associated with the genre class m in a corpus with k genres at the most fine-grained level. The predicted class is the class achieving the maximum inner product between $\mathbf{x}$ and the weight vector for the class, denoted as,

$$\arg\max_m \ \mathbf{w}_m^T \mathbf{x}, \forall m. \qquad (1)$$

Accurate prediction requires that when a document vector is multiplied with the weight vector associated with its own class, the resulting inner product should be larger than its inner products with a weight vector for any other genre class m. This helps us to define criteria for weight vectors. Let $\mathbf{x}_i$ be the $i-$th training document, and y_i its genre label. For its weight vector $\mathbf{w}_{y_i}$, the inner product $\mathbf{w}_{y_i}^T \mathbf{x}_i$ should be larger than all other products $\mathbf{w}_m^T \mathbf{x}_i$, that is,

$$\mathbf{w}_{y_i}^T \mathbf{x}_i - \mathbf{w}_m^T \mathbf{x}_i \geq 0, \forall m. \quad (2)$$

To strengthen the constraints, the zero value on the right hand side of the inequality for the flat SVM can be replaced by a positive value, corresponding to a distance measure $h(y_i, m)$ between two genre classes, leading to the following constraint:

$$\mathbf{w}_{y_i}^T \mathbf{x}_i - \mathbf{w}_m^T \mathbf{x}_i \geq h(y_i, m), \forall m. \quad (3)$$

To allow feasible models, in real scenarios such constraints can be violated, but the degree of violation is expected to be small. For each document, the maximum violation in the k constraints is of interest, as given by the following loss term:

$$Loss_i = \max_m \{h(y_i, m) - \mathbf{w}_{y_i}^T \mathbf{x}_i + \mathbf{w}_m^T \mathbf{x}_i\}. \quad (4)$$

Adding up all loss terms over all training documents, and further introducing a term to penalize large values in the weight vectors, we have the following objective function (C is a user-specified nonnegative parameter).

$$\min_{m,i} : \frac{1}{2} \sum_{m=1}^{k} \mathbf{w}_m^T \mathbf{w}_m + C \sum_{i=1}^{p} Loss_i. \quad (5)$$

Efficient methods can be derived by borrowing the sequential dual methods in (Keerthi et al., 2008) or other optimization techniques (Crammer and Singer, 2002).

3 Genre Distance Measures

The structural SVM (Section 2) requires a distance measure h between two genres. We can derive such distance measures from the genre hierarchy in a way similar to word similarity measures that were invented for lexical hierarchies such as WordNet (see (Pedersen et al., 2007) for an overview). In the following, we will first shortly summarise path-based and information-based measures for similarity. However, information-based measures are based on the information content of a node in a hierarchy. Whereas the information content of a word or concept in a lexical hierarchy has been well-defined (Resnik, 1995), it is less clear how to estimate the information content of a genre label. We will therefore discuss several different ways of estimating information content of nodes in a genre hierarchy.

3.1 Distance Measures based on Path Length

If genre labels are organised into a tree (Figure 1), one of the simplest ways to measure distance between two genre labels (= tree nodes) is ***path length*** ($h(a,b)_{plen}$):

$$f(a, LCS(a,b)) + f(b, LCS(a,b)), \quad (6)$$

where a and b are two nodes in the tree, $LCS(a,b)$ is their Least Common Subsumer, and $f(a, LCS(a,b))$ is the number of levels passed through when traversing from a to the ancestral node $LCS(a,b)$. In other words, the distance counts the number of edges traversed from nodes a to b in the tree. For example, the distance between `Learned` and `Misc` in Figure 1 would be 3.

As an alternative, the ***maximum path length*** $h(a,b)_{pmax}$ to their least common subsumer can be used to reduce the range of possible values:

$$\max\{f(a, LCS(a,b)), f(b, LCS(a,b))\}. \quad (7)$$

The ***Leacock & Chodorow*** similarity measure (Leacock and Chodorow, 1998) normalizes the path length measure (6) by the maximum number of nodes D when traversing down from the root.

$$s(a,b)_{plsk} = -log((h(a,b)_{plen} + 1)/2D). \quad (8)$$

To convert it into a distance measure, we can invert it $h(a,b)_{plsk} = 1/s(a,b)_{plsk}$.

Other path-length based measures include the ***Wu & Palmer*** Similarity (Wu and Palmer, 1994).

$$s(a,b)_{pwupal} = \frac{2f(R, LCS(a,b))}{(f(R,a) + f(R,b))}, \quad (9)$$

where R describes the hierarchy's root node. Here similarity is proportional to the shared path from the root to the least common subsumer of two nodes. Since the Wu & Palmer similarity is always between [0 1), we can convert it into a distance measure by $h(a,b)_{pwupal} = 1 - s(a,b)_{pwupal}$.

3.2 Distance Measures based on Information Content

Path-based distance measures work relatively well on balanced hierarchies such as the one in Figure 1 but fail to treat hierarchies with different levels of granularity well. For lexical hierarchies, as a result, several distance measures based on *information content* have been suggested where the information content of a concept c in a hierarchy is measured by (Resnik, 1995)

$$IC(c) = -log(\frac{freq(c)}{freq(root)}). \quad (10)$$

The frequency $freq$ of a concept c is the sum of the frequency of the node c itself and the frequencies of all its subnodes. Since the root may be a dummy concept, its frequency is simply the sum of the frequencies of all its subnodes. The similarity between two nodes can then be defined as the information content of their least common subsumer:

$$s(a,b)_{resk} = IC(LCS(a,b)). \quad (11)$$

If two nodes just share the root as their subsumer, their similarity will be zero. To convert 11 into a distance measure, it is possible to add a constant 1 to it before inverting it, as given by

$$h(a,b)_{resk} = 1/(s(a,b)_{resk} + 1). \quad (12)$$

Several other similarity measures have been proposed based on the Resnik similarity such as the one by (Lin, 1998):

$$s(a,b)_{lin} = \frac{2IC(LCS(a,b))}{IC(a) + IC(b)}. \quad (13)$$

Again to avoid the effect of zero similarity when defining the ***Lin's distance*** we use:

$$h(a,b)_{lin} = 1/(s(a,b)_{lin} + 1). \quad (14)$$

(Jiang and Conrath, 1997) directly define ***Jiang's distance*** ($h(a,b)_{jng}$):

$$IC(a) + IC(b) - 2IC(LCS(a,b)). \quad (15)$$

3.2.1 Information Content of Genre Labels

The notion of information content of a genre is not straightforward. We use two ways of measuring the frequency $freq$ of a genre, depending on its interpretation.

Genre Frequency based on Document Occurrence. We can interpret the "frequency" of a genre node simply as the number of all documents belonging to that genre (including any of its subgenres). Unfortunately, there are no estimates for genre frequencies on, for example, a representative sample of web documents. Therefore, we approximate genre frequencies from the document frequencies (*dfs*) in the training sets used in classification. Note that (i) for balanced class distributions this information will not be helpful and (ii) that this is a relatively poor substitute for an estimation on an independent, representative corpus.

Genre Frequency based on Genre Labels. We can also use the labels/names of the genre nodes as the unit of frequency estimation. Then, the frequency of a genre node is the occurrence frequency of its label in a corpus plus the occurrence frequencies of the labels of all its subnodes. Note that there is no direct correspondence between this measure and the document frequency of a genre: measuring the number of times the potential genre label *poem* occurs in a corpus is not in any way equivalent to the number of poems in that corpus. However, the measure is still structurally aware as frequencies of labels of subnodes are included, i.e. a higher level genre label will have higher frequency (and lower information content) than a lower level genre label.[1]

For label frequency estimation, we manually expand any label abbreviations (such as "newsp" for BNC genre labels), delete stop words and function words and then use two search methods. For the search method $word$ we simply search the frequency of the genre label in a corpus, using three different corpora (the BNC, Brown and Google web search). As for the BNC and Brown corpus some labels are very rarely mentioned, we for these two corpora use also a search method $gram$ where all character 5-grams within the genre label are searched for and their frequencies aggregated.

3.3 Terminology

Algorithms are prefixed by the kind of distance measure they employ — IC for Information content and p for path-based). If the measure is infor-

[1] Obviously when using this measure we rely on genre labels which are meaningful in the sense that lower level labels were chosen to be more specific and therefore probably rarer terms in a corpus. The measure could not possibly be useful on a genre hierarchy that would give random names to its genres such as *genre 1*.

mation content based the specific measure is mentioned next, such as lin. The way for measuring genre frequency is indicated last with df for measuring via document frequency and $word/gram$ when measured via frequency of genre labels. If frequencies of genre labels are used, the corpus for counting the occurrence of genre labels is also indicated via *brown*, *bnc* or the Web as estimated by Google hit counts gg. Standard non-structural SVMs are indicated by $flat$.

4 Experiments

4.1 Datasets

We use four genre-annotated corpora for genre classification: the Brown Corpus (Kučera and Francis, 1967), BNC (Lee, 2001), HGC (Stubbe and Ringlstetter, 2007) and Syracuse (Crowston et al., 2009). They have a wide variety of genre labels (from 15 in the Brown corpus to 32 genres in HGC to 70 in the BNC to 292 in Syracuse), and different types of hierarchies.

4.2 Evaluation Measures

We use standard classification accuracy (Acc) on the most fine-grained level of target categories in the genre hierarchy.

In addition, given a structural distance H, misclassifications can be weighted based on the distance measure. This allows us to penalize incorrect predictions which are further away in the hierarchy (such as between government documents and westerns) more than "close" mismatches (such as between science fiction and westerns). Formally, given the classification confusion matrix M then each M_{ab} for $a \neq b$ contains the number of class a documents that are misclassified into class b. To achieve proper normalization in giving weights to misclassified entries, we can redistribute a total weight $k-1$ to each row of H proportionally to its values, where k is the number of genres. That is, given g the row summation of H, we define a weight matrix Q by normalizing the rows of H in a way given by $Q_{ab} = (k-1)h_{ab}/g_a,\ a \neq b$. We further assign a unit value to the diagonal of Q. Then it is possible to construct a structurally-aware measure (S-Acc):

$$\text{S-Acc} = \sum_a M_{aa} / \sum_{a,b} M_{ab} Q_{ab}. \qquad (16)$$

4.3 Experimental Setup

We compare structural SVMs using all path-based and information-content based measures (see also Section 3.3). As a baseline we use the accuracy achieved by a standard "flat" SVM.

We use 10-fold (randomised) cross validation throughout. In each fold, for each genre class 10% of documents are used for testing. For the remaining 90%, a portion of 10% are sampled for parameter tuning, leaving 80% for training. In each round the validation set is used to help determine the best C associated with Equation (5) based on the validation accuracy from the candidate list 0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1. Note via this experiment setup, all methods are tuned to their best performance.

For any algorithm comparison, we use a McNemar test with the significance level of 5% as recommended by (Dietterich, 1998).

4.4 Features

The features used for genre classification are character 4-grams for all algorithms, i.e. each document is represented by a binary vector indicating the existence of each character 4-gram. We used character n-grams because they are very easy to extract, language-independent (no need to rely on parsing or even stemming), and they are known to have the best performance in genre classification tasks (Kanaris and Stamatatos, 2009; Sharoff et al., 2010).

4.5 Brown Corpus Results

The Brown Corpus has 500 documents and is organized in a hierarchy with a depth of 3. It contains 15 end-level genres. In one experiment in (Karlgren and Cutting, 1994) the subgenres under *fiction* are grouped together, leading to 10 genres to classify.

Results on 10-genre Brown Corpus. A standard flat SVM achieves an accuracy of 64.4% whereas the best structural SVM based on Lin's information content distance measure (IC-lin-word-bnc) achieves 68.8% accuracy, significantly better at the 1% level. The result is also significantly better than prior work on the Brown corpus in (Karlgren and Cutting, 1994) (who use the whole corpus as test as well as training data). Table 1 summarizes the best performing measures that all outperform the flat SVM at the 1% level.

Table 1: Brown 10-genre Classification Results.

Method	Accuracy
Karlgren and Cutting, 1994	65 (Training)
Flat SVM	64.40
SSVM(IC-lin-word-bnc)	68.80
SSVM(IC-lin-word-br)	68.60
SSVM(IC-lin-gram-br)	67.80

Figure 2 provides the box plots of accuracy scores. The dashed boxes indicate that the distance measures perform significantly worse than the best performing *IC-lin-word-bnc* at the bottom. The solid boxes indicate the corresponding measures are statistically comparable to the *IC-lin-word-bnc* in terms of the mean accuracy they can achieve.

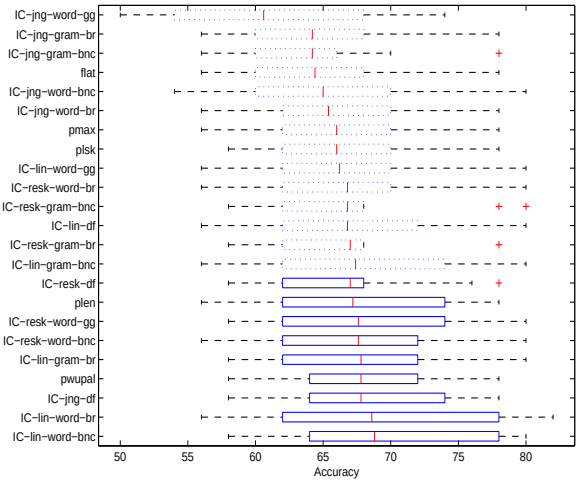

Figure 2: Accuracy on Brown Corpus (10 genres).

Results on 15-genre Brown Corpus. We perform experiments on all 15 genres on the end level of the Brown corpus. The increase of genre classes leads to reduced classification performance. In our experiment, the flat SVM achieves an accuracy of 52.40%, and the structural SVM using path length measure achieves 55.40%, a difference significant at the 5% level. The structural SVMs using information content measures *IC-lin-gram-bnc and IC-resk-word-br* also perform equally well. In addition, we improve on the training accuracy of 52% reported in (Karlgren and Cutting, 1994).

We are also interested in structural accuracy (S-Acc) to see whether the structural SVMs make fewer "big" mistakes. Table 2 shows a cross comparison of structural accuracy. Each row shows how accurate the corresponding method is under the structural accuracy criteria given in the column. The 'no-struct' column corresponds to vanilla accuracy. It is natural to expect each diagonal entry of the numeric table to be the highest, since the respective method is optimised for its own structural distance. However, in our case, Lin's information content measure and the plen measure perform well under any structural accuracy evaluation measure and outperform flat SVMs.

4.6 Other Corpora

In spite of the promising results on the Brown Corpus, structural SVMs on other corpora (BNC, HGC, Syracuse) did not show considerable improvement.

HGC contains 1330 documents divided into 32 approximately equally frequent classes. Its hierarchy has just two levels. Standard accuracy for the best performing structural methods on HGC is just the same as for flat SVM (69.1%), with marginally better structural accuracy (for example, 71.39 vs. 71.04%, using a path-length based structural accuracy). The BNC corpus contains 70 genres and 4053 documents. The number of documents per class ranges from 2 to 501. The accuracy of SSVM is also just comparable to flat SVM (73.6%). The Syracuse corpus is a recently developed large collection of 3027 annotated webpages divided into 292 genres (Crowston et al., 2009). Focusing only on genres containing 15 or more examples, we arrived at a corpus of 2293 samples and 52 genres. Accuracy for flat (53.3%) and structural SVMs (53.7%) are again comparable.

5 Discussion

Given that structural learning can help in topical classification tasks (Tsochantaridis et al., 2005; Dekel et al., 2004), the lack of success on genres is surprising. We now discuss potential reasons for this lack of success.

5.1 Tree Depth and Balance

Our best results were achieved on the Brown corpus, whose genre tree has at least three attractive properties. Firstly, it has a depth greater than 2, i.e. several levels are distinguished. Secondly, it seems visually balanced: branches from root to leaves (or terminals) are of pretty much equal length; branching factors are similar, for example ranging between 2 and 6 for the last level of branching. Thirdly, the number of examples at

Table 2: Structural Accuracy on Brown 15-genre Classification.

Method	no-struct (=typical accuracy)	IC-lin-gram-bnc	plen	IC-resk-word-br	IC-jng-word-gg
flat	52.40	55.34	60.60	58.91	52.19
IC-lin-gram-bnc	55.00	58.15	63.59	61.83	53.85
plen	**55.40**	**58.74**	**64.51**	**62.61**	**54.27**
IC-resk-word-br	55.00	58.24	63.96	62.08	54.08
IC-jng-word-gg	46.00	49.00	54.89	53.01	52.58

each leaf node is roughly comparable (distributional balance).

The other hierarchies violate these properties to a large extent. Thus, the genres in HGC are almost represented by a flat list with just one extra level over 32 categories. Similarly, the vast majority of genres in the Syracuse corpus are also organised in two levels only. Such flat hierarchies do not offer much scope to improve over a completely flat list. There are considerably more levels in the BNC for some branches, e.g., *written/national/broadsheet/arts*, but many other genres are still only specified to the second level of its hierarchy, e.g., *written/adverts*. In addition, the BNC is also distributionally imbalanced, i.e. the number of documents per class varies from 2 to 501 documents.

To test our hypothesis, we tried to skew the Brown genre tree in two ways. First, we kept the tree relatively balanced visually and distributionally but flattened it by removing the second layer *Press, Misc, Non-Fiction, Fiction* from the hierarchy, leaving a tree with only two layers. Second, we skewed the visual and distributional balance of the tree by collapsing its three leaf-level genres under *Press*, and the two under *non-fiction*, leading to 12 genres to classify (cf. Figure 1).

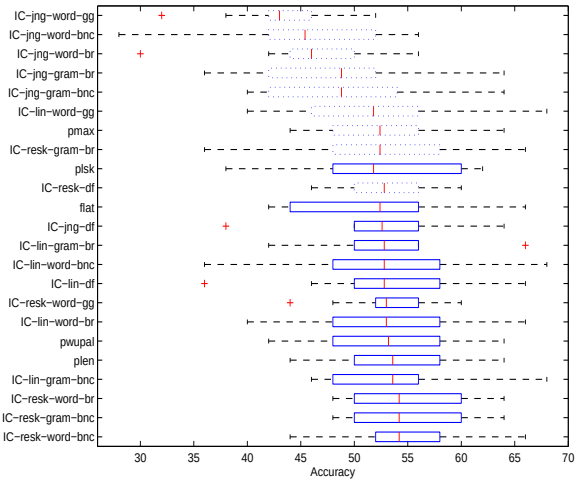

Figure 3: Accuracy on flattened Brown Corpus (15 genres).

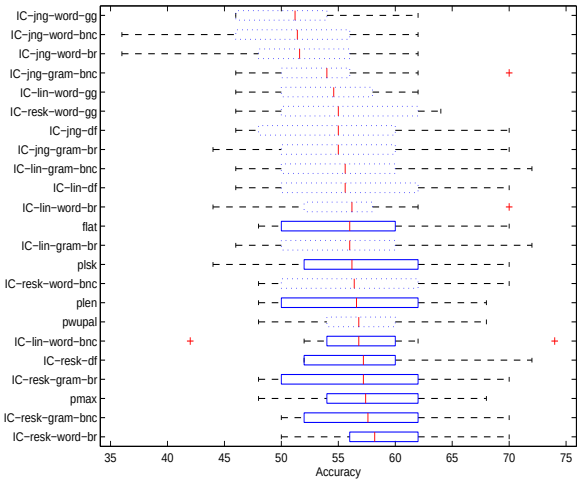

Figure 4: Accuracy on skewed Brown Corpus (12 genres).

As expected, the structural methods on either skewed or flattened hierarchies are not significantly better than the flat SVM. For the flattened hierarchy of 15 leaf genres the maximal accuracy is 54.2% vs. 52.4% for the flat SVM (Figure 3), a non-significant improvement. Similarly, the maximal accuracy on the skewed 12-genre hierarchy is 58.2% vs. 56% (see also Figure 4), again a not significant improvement.

To measure the degree of balance of a tree, we introduce two tree balance scores based on entropy. First, for both measures we extend all branches to the maximum depth of the tree. Then level by level we calculate an entropy score, either according to how many tree nodes at the next level belong to a node at this level (denoted as **vb: visual balance**), or according to how many end level documents belong to a node at this level (denoted as **db: distribution balance**). To make trees with different numbers of internal nodes and leaves more comparable, the entropy score at each level is normalized by the maximal entropy achieved by a tree with uniform distribution of nodes/documents, which is simply $-log(1/N)$, where N denotes the number of nodes at the corre-

sponding level. Finally, the entropy scores for all levels are averaged. It can be shown that any perfect N-ary tree will have the largest visual balance score of 1. If in addition its nodes at each level contain the same number of documents, the distribution balance score will reach the maximum, too.

Table 3 shows the balance scores for all the corpora we use. The first two rows for the Brown corpus have both large visual balance and distribution balance scores. As shown earlier, for those two setups the structural SVMs perform better than the flat approach. In contrast, for the tree hierarchies of Brown that we deformed or flattened, and also BNC and Syracuse, either or both of the two balance scores tend to be lower, and no improvement has been obtained over the flat approach. This may indicate that a further exploration of the relation between tree balance and the performance of structural SVMs is warranted. However, high visual balance and distribution scores do not necessarily imply high performance of structural SVMs, as very flat trees are also visually very balanced. As an example, HGC has a high visual balance score due to a shallow hierarchy and a high distributional balance score due to a roughly equal number of documents contained in each genre. However, HGC did not benefit from structural learning as it is also a very shallow hierarchy; therefore we think that a third variable depth also needs to be taken into account.

A similar observation on the importance of well-balanced hierarchies comes from a recent Pascal challenge on large scale hierarchical text classification,[2] which shows that some flat approaches perform competitively in topic classification with imbalanced hierarchies. However, the participants do not explore explicitly the relation between tree balance and performance.

Other methods for measuring tree balance (some of which are related to ours) are used in the field of phylogenetic research (Shao and Sokal, 1990) but they are only applicable to visual balance. In addition, the methods they used often provide conflicting results on which trees are considered as balanced (Shao and Sokal, 1990).

Table 3: Tree Balance Scores

Corpus	depth	vb	db
Brown (10 genres)	3	**0.9115**	**0.9024**
Brown (15 genres)	3	**0.9186**	**0.9083**
Brown (15, flattened)	2	0.9855	0.8742
Brown (12, skewed)	3	0.8747	0.8947
HGC (32)	2	**0.9562**	**0.9570**
BNC (70)	4	0.9536	0.8039
Syracuse (52)	3	0.9404	0.8634

5.2 Distance Measures

We also scrutinise our distance measures as these are crucial for the structural approach. We notice that simple path length based measures perform well overall; again for the Brown corpus this is probably due to its balanced hierarchy which makes path length appropriate. There are other probable reasons why information content based measures do not perform better than path-length based ones. When measured via document frequency in a corpus we do not have sufficiently large, representative genre-annotated corpora to hand. When measured via genre label frequency, we run into at least two problems. Firstly, as mentioned in Section 3.2.1 genre label frequency does not have to correspond to class frequency of documents. Secondly, the labels used are often abbreviations (e.g. *W_institut_doc, W_newsp_brdsht_nat_social* in BNC Corpus), underspecified (*other, misc, unclassified*) or a collection of phrases (e.g. *belles letters, etc.* in Brown). This made search for frequency very approximate and also loosens the link between label and content.

We investigated in more depth how well the different distance measures are aligned. We adapt the alignment measure between kernels (Cristianini et al., 2002), to investigate how close the distance matrices are. For two distance matrices H_1 and H_2, their alignment $A(H_1, H_2)$ is defined as:

$$\frac{< H_1, H_2 >_F}{\sqrt{< H_1, H_1 >_F, < H_2, H_2 >_F}}, \quad (17)$$

where $< H_1, H_2 >_F = \sum_{i,j}^{k} H_1(g_i, g_j) H_2(g_i, g_j)$ which is the total sum of the entry-wise products between the two distance matrices. Figure 5 shows several distance matrices on the (original) 15 genre Brown corpus. The *plen* matrix has clear blocks for the super genres *press, informative, imaginative*, etc. The *IC-lin-gram-bnc* matrix refines distances in the blocks, due to the introduction of information content. It keeps an alignment score that is over 0.99 (the maximum is 1.00) toward the *plen* matrix, and still has visible block patterns. However, the *IC-jng-word-bnc* significantly adjusts the

[2] http://lshtc.iit.demokritos.gr/

distance entries, has a much lower alignment score with the *plen* matrix, and doesn't reveal apparent blocks. This partially explains the bad performance of the Jiang distance measure on the Brown corpus (see Section 4). The diagrams also show the high closeness between the best performing IC measure and the simple path length based measure.

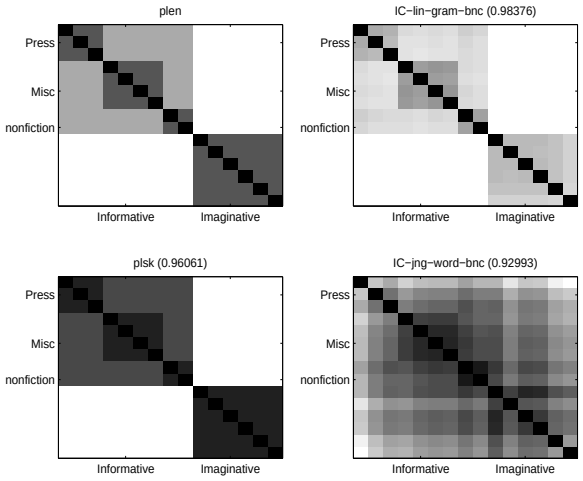

Figure 5: Distance Matrices on Brown. Values in bracket is the alignment with the *plen* matrix

An alternative to structural distance measures would be distance measures between the genres based on pairwise cosine similarities between them. To assess this, we aggregated all character 4-gram training vectors of each genre and calculated standard cosine similarities. Note that these similarities are based on the documents only and do not make use of the Brown hierarchy at all. After converting the similarities to distance, we plug the distance matrix into our structural SVM. However, accuracy on the Brown corpus (15 genres) was almost the same as for a flat SVM. Inspecting the distance matrix visually, we determined that the cosine similarity could clearly distinguish between Fiction and Non-Fiction texts but not between any other genres. This also indicates that the genre structural hierarchy clearly gives information not present in the simple character 4-gram features we use. For a more detailed discussion of the problems of the currently prevalently used character n-grams as features for genre classification, we refer the reader to (Sharoff et al., 2010).

6 Conclusions

In this paper, we have evaluated structural learning approaches to genre classification using several different genre distance measures. Although we were able to improve on non-structural approaches for the Brown corpus, we found it hard to improve over flat SVMs on other corpora. As potential reasons for this negative result, we suggest that current genre hierarchies are either not of sufficient depth or are visually or distributionally imbalanced. We think further investigation into the relationship between hierarchy balance and structural learning is warranted. Further investigation is also needed into the appropriateness of n-gram features for genre identification as well as good measures of genre distance.

In the future, an important task would be the refinement or unsupervised generation of new hierarchies, using information theoretic or data-driven approaches. For a full assessment of hierarchical learning for genre classification, the field of genre studies needs a testbed similar to the Reuters or 20 Newsgroups datasets used in topic-based IR with a balanced genre hierarchy and a representative corpus of reliably annotated webpages.

With regard to algorithms, we are also interested in other formulations for structural SVMs and their large-scale implementation as well as the combination of different distance measures, for example in ensemble learning.

Acknowledgements

We would like to thank the authors of each corpus collection, who invested a lot of effort into producing them. We are also grateful to Google Inc. for supporting this research via their Google Research Awards programme.

References

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *COLT '92: Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, New York, NY, USA. ACM.

Crammer, K. and Singer, Y. (2002). On the algorithmic implementation of multiclass kernel-based vector machines. *J. Mach. Learn. Res.*, 2:265–292.

Cristianini, N., Shawe-Taylor, J., and Kandola, J. (2002). On kernel target alignment. In *Proceedings of the Neural Information Process-*

ing Systems, NIPS'01, pages 367–373. MIT Press.

Crowston, K., Kwasnik, B., and Rubleske, J. (2009). Problems in the use-centered development of a taxonomy of web genres. In Mehler, A., Sharoff, S., and Santini, M., editors, *Genres on the Web: Computational Models and Empirical Studies*. Springer, Berlin/New York.

Dekel, O., Keshet, J., and Singer, Y. (2004). Large margin hierarchical classification. In *ICML '04: Proceedings of the twenty-first international conference on Machine learning*, page 27, New York, NY, USA. ACM.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10:1895–1923.

Giesbrecht, E. and Evert, S. (2009). Part-of-Speech (POS) Tagging - a solved task? An evaluation of POS taggers for the Web as corpus. In *Proceedings of the Fifth Web as Corpus Workshop (WAC5)*, pages 27–35, Donostia-San Sebastián.

Jiang, J. J. and Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. *CoRR*, cmp-lg/9709008.

Joachims, T. (1999). Making large-scale SVM learning practical. In Schölkopf, B., Burges, C., and Smola, A., editors, *Advances in Kernel Methods – Support Vector Learning*, pages 41–56. MIT Press.

Joachims, T., Finley, T., and Yu, C.-N. (2009). Cutting-plane training of structural svms. *Machine Learning*, 77(1):27–59.

Kanaris, I. and Stamatatos, E. (2009). Learning to recognize webpage genres. *Information Processing and Management*, 45:499–512.

Karlgren, J. and Cutting, D. (1994). Recognizing text genres with simple metrics using discriminant analysis. In *Proc. of the 15th. International Conference on Computational Linguistics (*COLING 94*)*, pages 1071 – 1075, Kyoto, Japan.

Keerthi, S. S., Sundararajan, S., Chang, K.-W., Hsieh, C.-J., and Lin, C.-J. (2008). A sequential dual method for large scale multi-class linear svms. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 408–416, New York, NY, USA. ACM.

Kessler, B., Nunberg, G., and Schütze, H. (1997). Automatic detection of text genre. In *Proceedings of the 35^{th} ACL/8^{th} EACL*, pages 32–38.

Kučera, H. and Francis, W. N. (1967). *Computational analysis of present-day American English*. Brown University Press, Providence.

Leacock, C. and Chodorow, M. (1998). *Combining local context and WordNet similarity for word sense identification*, pages 305–332. In C. Fellbaum (Ed.), MIT Press.

Lee, D. (2001). Genres, registers, text types, domains, and styles: clarifying the concepts and navigating a path through the BNC jungle. *Language Learning and Technology*, 5(3):37–72.

Lin, D. (1998). An information-theoretic definition of similarity. In *ICML '98: Proceedings of the Fifteenth International Conference on Machine Learning*, pages 296–304, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Meyer zu Eissen, S. and Stein, B. (2004). Genre classification of web pages. In *Proceedings of the 27th German Conference on Artificial Intelligence*, Ulm, Germany.

Pedersen, T., Pakhomov, S. V. S., Patwardhan, S., and Chute, C. G. (2007). Measures of semantic similarity and relatedness in the biomedical domain. *J. of Biomedical Informatics*, 40(3):288–299.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *IJCAI'95: Proceedings of the 14th international joint conference on Artificial intelligence*, pages 448–453, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Santini, M. (2007). *Automatic Identification of Genre in Web Pages*. PhD thesis, University of Brighton.

Shao, K.-T. and Sokal, R. R. (1990). Tree balance. *Systematic Zoology*, 39(3):266–276.

Sharoff, S., Wu, Z., and Markert, K. (2010). The Web library of Babel: evaluating genre collections. In *Proc. of the Seventh Language Resources and Evaluation Conference, LREC 2010*, Malta.

Stubbe, A. and Ringlstetter, C. (2007). Recognizing genres. In Santini, M. and Sharoff, S., editors, *Proc. Towards a Reference Corpus of Web Genres*.

Tsochantaridis, I., Joachims, T., Hofmann, T., and Altun, Y. (2005). Large margin methods for structured and interdependent output variables. *J. Mach. Learn. Res.*, 6:1453–1484.

Vidulin, V., Luštrek, M., and Gams, M. (2007). Using genres to improve search engines. In *Proc. Towards Genre-Enabled Search Engines: The Impact of NLP. RANLP-07*.

Webber, B. (2009). Genre distinctions for discourse in the Penn TreeBank. In *Proc the 47th Annual Meeting of the ACL*, pages 674–682.

Wu, Z. and Palmer, M. (1994). Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138, Morristown, NJ, USA. Association for Computational Linguistics.