# Fine-grained Genre Classification using Structural Learning Algorithms

**Blank for Review**

## Abstract

Machine Learning (ML) techniques, such as SVM, are commonly used in various text classification tasks, one of which is automatic genre identification. The use of ML normally relies on a list of labels. However, genre classes are often organised into hierarchies, e.g., covering the subgenres of fiction or newspaper texts. 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 SVM over the flat one are not statistically significant.

## 1 Introduction

Early computational attempts into automatic genre identification (AGI) can be traced to the middle of the 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 other NLP applications. In Information Retrieval one and the same topical query can relate to pages of different types. Given the large number of 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 NLP 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 aproliferation of studies on development of corpora of web genres and comparison of methods for in 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 standard 10-fold cross-validation) reach 84.1% on KI-04 and 96.5% accuracy on Santinis (Kanaris and Stamatatos, 2009). In our research (anonymized) we managed to produce 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, e.g., using the WEGA plugin (Stein and Meyer zu Eissen, 2008). One reason comes from the limited number of genres present in these two collections (eight genres in KI-04 and seven genres in Santinis), which do not cover many genres frequent on the Web, e.g., 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., FAQs in Santinis come from two sources, a website with FAQs on hurricanes and another one with tax advice. In the end, an SVM classifier built for FAQs relies on occasional properties of these two collections 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 using 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, see 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`).

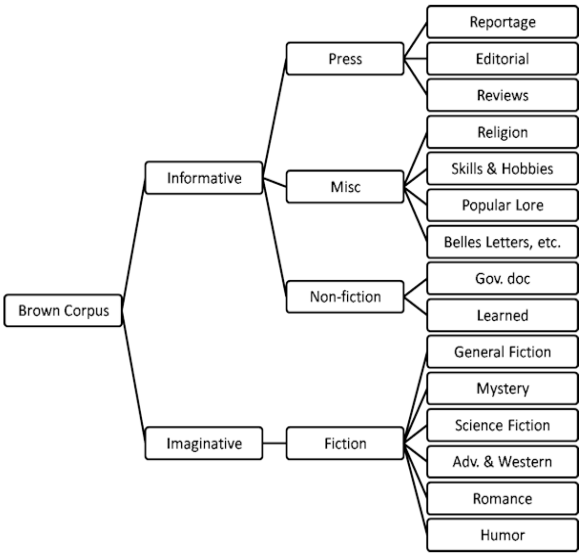

Figure 1: Hierarchy of Brown corpus.

This paper proceeds from this idea and 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 in Section 4. These experiments show that on three different genre-annotated corpora with their own genre hierarchies, structural SVMs can outperform classification accuracy over the non-structural standard SVM baseline. However, the improvement is only statistically significant on the Brown corpus.

## 2 Structural SVMs

Discriminative methods are often used for classification, with SVMs being a well-performing method in many tasks (Boser et al., 1992). When it is applied to text materials, linear support vector machines are normally used (Joachims, 1999), since NLP tasks usually result in a large number of features and documents.

Linear SVMs on a flat list of labels achieve high efficiency and accuracy as 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.

The formulation presented here can be taken as a special instance of the general structural learning framework in (Tsochantaridis et al., 2005). However, (Tsochantaridis et al., 2005) concentrate on more complicated label structures as for sequence alignment or sentence parsing. They proposed two formulations, a slack-rescaling and a margin-rescaling. They claimed that the slack-rescaling version has two advantages. Firstly, it is invariant to the scaling of the distance matrix; secondly, it may improve significant weights to output values that are more confusable with the true output. However, they did not provide empirical evidence in favour of slack-rescaling. In this experiment, we tried margin-rescaling, since it allows us to use the sequential dual method for large-scale implementation based on (Keerthi et al., 2008), which is not applicable to the slack-rescaling formulation, while for web page classification we will need fast processing. We also perform model calibration to adress the first disadvantage of margin-rescaling mentioned in (Tsochantaridis et al., 2005).

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,

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

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 $y_i$ the 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,

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

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 structural distance measure $h(y_i, m)$ between two genre classes, leading to the following constraint:

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

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:

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

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).

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

Efficient methods can be derived by borrowing the sequential dual methods in (Keerthi et al., 2008) or other optimization techniques (Crammer and Singer, 2002). Due to space limitation, a detailed analysis on its algorithmic steps and complexity is not presented here.

## 3 Genre Distance Measures

The structural SVM as formulated in Section 2 requires a distance measure $h$ between two genre classes. We can derive such distance measures from the hierarchy of genre organisation 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 subsections 3.1 and 3.2 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, which is in turn based on the frequency of a node in a hierarchy. Whereas the information content or frequency 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 or frequency 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 (see Figure 1), one of the simplest ways to measure distance between two genre labels (= tree nodes) is the *path length* ($h(a, b)_{plen}$):

$$f(a, LCS(a, b)) + f(b, LCS(a, b)), \tag{6}$$

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` would be 3.

In addition to (6), the maximum height to their least common subsumer can be used to reduce the range of possible values ($h(a, b)_{pmax}$) defined as

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

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.

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

To convert it into a distance measure, we can follow the typical way of inverting it $h(a, b)_{plsk} = 1/s(a, b)$.

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

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

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 $h(a, b)_{pwupal} = 1 - s(a, b)$.

## 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 (which ensures that the frequency of a concept is always bigger than the frequency of any of 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 it 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 distance we use:

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

(Jiang and Conrath, 1997) directly define a distance based on three terms ($h(a, b)_{jng}$):

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

### 3.2.1 Information Content of Genre Labels

It is easy to use the measures based on path length for any genre hierarchy, whereas the notion of information content of a genre or a genre label is not as straightforward. There are 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). There are no general estimates for genre frequencies on, for example, all web documents and the genre frequency defined in this way will vary between corpus collections. Thus, the (relative) frequency of Belles Lettres in the Brown Corpus will probably be higher than on the Web. Therefore, we approximate genre frequencies from the document frequencies (dfs) in the training sets used in classification. Note that for balanced class distributions this information will not be helpful.

**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]

One problem in estimating genre label frequencies is that they may not correspond to single words or standard compound phrases. To avoid the need to analyse compound phrases, we just counted the frequencies of their components. In addition, Any abbreviations such as ("newsp" for BNC genre labels) have been manually expanded. Stop words and function words in each genre label have been removed and all words are stemmed.

## 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 typologies of hierarchies. For example, the Brown corpus hierarchy is a relatively well balanced tree, see Figure 1, whereas the BNC hierarchy has quite short as well as long branches.

For estimating the frequencies of genre labels, we use different estimations based on the Brown Corpus, the BNC or the Web where the latter is approximated by Google hit counts. As described in Section 3.2.1, we also use the document frequency of a genre type in the training corpus as a different genre frequency measure.

### 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). With proper normalization we can get a structural accuracy rate (S-Acc).

Formally, given the classification confusion matrix $M$, with $M_{ab} = \sum_i (y_i = a) \wedge (\bar{y}_i = b)$, where $y_i$ and $\bar{y}_i$ are the true label and the predicted label of sample $i$, respectively, 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. 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 $Q_{ab} = (k - 1)h_{ab}/d_a$, $a \neq b$. 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, relying on either the document frequency (*df*) of each genre in the training corpus or word/4-gram frequency estimates of genre labels (denoted as *word* and *gram*) in *brown*, *bnc* or the Web as estimated by Google search *gg*. As a baseline we use the accuracy achieved by a standard "flat" SVM.

We use 10-fold random 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 approach their best performance.

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

For $N$ distance measures, we have $N^2 - N$ statistical results. We then can also count how many times a measure wins, and subtract the value by the number of times it is worse than other measures.

[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*.

### 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, they are 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).

### 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 |

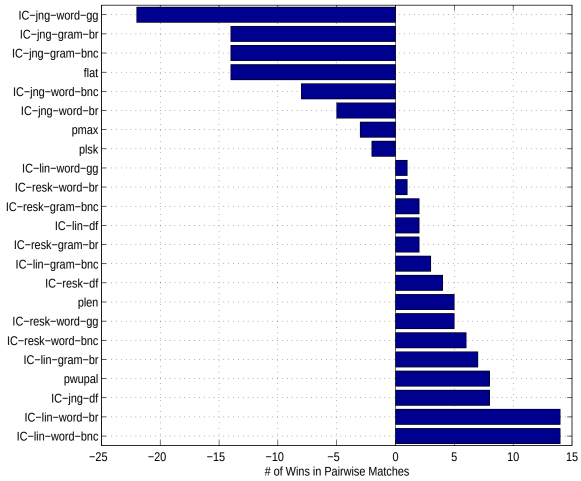

Figure 2: Wins in Brown Corpus (10 classes)

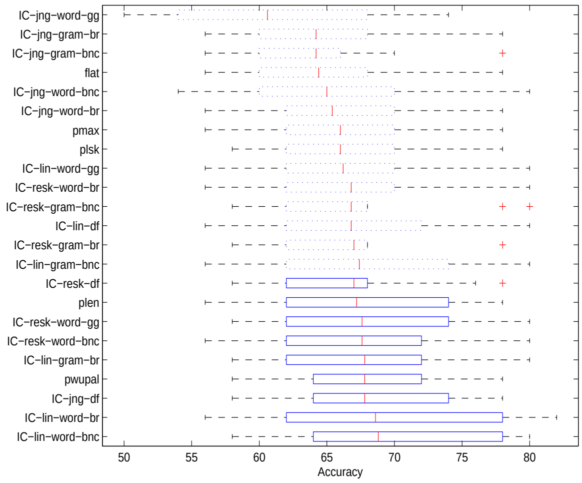

Figure 3: Accuracy on Brown Corpus.

For a full comparison, Figure 2 shows the number of wins in the pairwise Mcnemar comparisons of testing set predictions when all algorithms are applied to the 10-genre hierarchy of the Brown corpus. It can be noticed from the bottom part of the graph that the Lin's information content measures become the winner more frequently, as compared to the Flat SVM (denoted as 'flat').
Figure 3 (the box plot is adapted to indicate the mean result in the box) further provides the box plots of accuracy scores. The dashed boxes indicate that the distance measures perform significantly worse than the *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.

**Results on 15-genre Brown Corpus** We perform experiments on the 15 genres on the end level of Brown corpus. The distribution of end-level genres under *fiction* is quite imbalanced, containing only 6 *science fiction* and 9 *humor* fictions. The largest genre group of the corpus is *learned*, containing 80 educational and academic documents. The uneven nature and the increase of genre classes lead to reduced classification performance. In our experiment, the flat SVM achieves an accuracy of 52.40%, and the structural SVM using Lin's information content measure achieves 55.40%. The structural SVMs using *IC-lin-gram-bnx,plen,IC-resk-word-br* are significantly better

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 |

Table 3: HGC Corpus. Accuracy and structural Accuracy are reported.

| Method | no-struct (=standard accuracy) | IC-lin-word-gg | plen | pmax |
|---|---|---|---|---|
| Flat SVM | 0.6910 | 0.7003 | 0.7104 | 0.7113 |
| SSVM(IC-lin-word-gg) | 0.6910 | 0.7020 | 0.7139 | 0.7151 |

than the flat SVM at the 5% level. 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 "large" mistakes. Therefore, Table 2 shows a cross comparison of structural accuracy. Each row shows how accurate the prediction of the corresponding method is under the structural accuracy 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 based on ngram counting and the plen measure perform well under any structural accuracy evaluation measure and outperform flat SVMs.

As can be seen in the graphs above, our structural SVMs do not benefit much from the Jiang's information content measure. To see how differently it generates distances, 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:

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

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 4 shows several distance matrices. 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 distance entries, has a much lower alignment score with the *plen* matrix, and doesn't reveal apparent blocks.

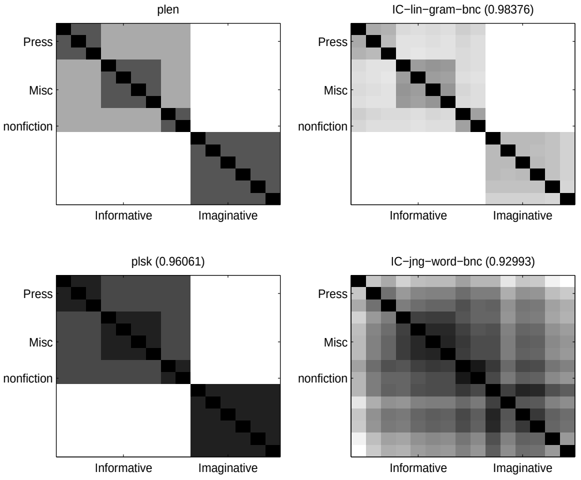

Figure 4: Distance Matrices on Brown. Values in bracket is the alignment with the *plen* matrix

These distance diagrams also show the high closeness between the best performing IC measure and the simple path length based measure. In fact since the Brown hierarchy is quite balanced and compact, path length based measures, like plen, pwupal, pmax, perform quite competitively. But for genre hierarchies that are more skewed and less compact, path length based measures may not always so effective.

### 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. The accuracy for HGC is just the same as flat SVM, with marginally better structural accuracy Table 3.

Table 4: BNC Corpus.

| Method | IC-lin-gram-br (S-Accuracy) | no-struct (=standard accuracy) |
|---|---|---|
| Flat SVM | 74.88 | 73.60 |
| SSVM(IC-lin-gram-br) | 75.24 | 73.84 |

Table 5: Syracuse Corpus.

| Method | IC-lin-word-br (S-Accuracy) | no-struct (=standard accuracy) |
|---|---|---|
| Flat SVM | 53.30 | 51.62 |
| SSVM(IC-lin-word-br) | 53.70 | 51.66 |

For the BNC, the accuracy of SSVM is also just comparable to flat SVM (Table 4).

The Syracuse corpus is a large recently developed 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. Table 5 shows one comparison between flat SVM and SSVM.

## 5 Discussion and Conclusion

In this paper, we have evaluated structural learning approaches to genre classification using several genre distance measures derived from measures used in word similarity. 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 using a single distance measure. The measures we evaluated perform competitively and may complement each other, so ensemble learning approaches can be use to combine results in a voting schema.

When comparing distance measures, we found that Lin's information content distance measures perform quite consistently well. The simple path length-based measures and Resnik's information content based one also work quite competitively. Analysis shows Jiang's information content based distance measure seems to be less helpful for structural learning.

Given that structural learning helps in topical classification tasks (Tsochantaridis et al., 2005; Dekel et al., 2004), the lack of success on genres is surprising. We think that it can be partly attributed to the structure of genre hierarchies in existing collections. The best results were achieved on the Brown corpus, containing a well balanced genre tree. The genres in HGC are 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 to 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*. For a full assessment of hierarchical learning for genre classification, the field of genre studies needs a representative testbed similar to the Reuters or 20 Newsgroups datasets used in topic-based IR with a clearly specified genre hierarchy.

We are also interested in other formulations for structural SVMs and their large-scale implementation. It may also be interesting to test corpus-based distributional distance measures between genres. As current genre hierarchies might be underspecified, a related task is their refinement or unsupervised generation of new hierarchies, using information theoretic or data driven approaches.

## 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. 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*, 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 multiclass 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.

Stein, B. and Meyer zu Eissen, S. (2008). Retrieval models for genre classification. *Scandinavian Journal of Information Systems (SJIS)*, 20(1):91–117.

Stubbe, A. and Ringlstetter, C. (2007). Recognizing genres. In *Abstract Proceedings of the Colloqium "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.