Learning to classify documents according to genre

Aidan Finn and Nicholas Kushmerick
Smart Media Institute, Department of Computer Science, University College Dublin
{aidan.finn, nick}@ucd.ie

Abstract

Current document retrieval tools succeed in locating large numbers of documents relevant to a given query. While search results may be relevant according to the topic of the documents, it is more difficult to identify which of the relevant documents are most suitable for a particular user. Automatic genre analysis - that is, the ability to distinguish documents according to style - would be a useful tool for identifying documents that are most suitable for a particular user. We investigate the use of machine learning for automatic genre classification. We introduce the idea of domain transfer - genre classifiers should be reusable across multiple topics - which doesn't arise in standard text classification. We investigate different features for building genre classifiers and their ability to transfer across multiple topic domains. We also show how different feature-sets can be used in conjunction with each other to improve performance and reduce the number of documents that need to be labeled.

1 Introduction

There is a vast amount of information available to the casual user today mainly due to the proliferation of the world wide web. However, it has become difficult to find the information that is most appropriate to a given query. While users can usually find relevant information, it is increasingly difficult to isolate information that is suitable in terms of style or genre. Current search services take a "one size fits all approach", taking little account of the individual users needs and preferences. These techniques succeed in identifying relevant documents, but the large number of documents relevant to a given query can make it difficult to isolate those documents that are *most relevant* to that query. Achieving high recall while maintaining

precision is very challenging. The huge volume of information available means that new techniques are needed to filter the relevant documents and identify the information that best satisfies a users information need.

We explore the use of genre to address this issue. By genre we loosely mean the style of text in the document. A genre class is a class of documents that are of a similar type. This classification is based not on the topic of the document, but rather on the kind of text used. We have identified automatic genre analysis as an additional tool that can complement existing techniques and improve the results returned to a user. Genre information could be used to filter or re-rank documents deemed relevant. The relevance of a particular document to a given query is dependent on the particular user issuing the query. We believe that the genre or style of text in a document can provide valuable additional information when determining which documents are most relevant to a particular user's query.

Machine learning has been widely used to categorize documents according to topic. In automatic text classification, a machine learning algorithm is given a set of examples of documents of different topics and it uses these examples to learn to distinguish documents. We consider the use of machine learning techniques to automatically categorize documents according to genre.

The ability to identify the style of text used in a document would be a valuable service in any text retrieval system. For example, consider a query about "chaos theory". Different users will require documents which assume different levels of expertise, depending on the users technical background. It would be useful to be able to rank documents according to the level of technical detail with which they present their subject. Current information retrieval systems would be greatly enhanced by the ability to filter documents according to their genre class. A high school student may require documents that are introductory or tutorial in style, while a college professor may require scholarly research documents.

As another example, consider news filtering according to the topic of the article. Such a service would be improved by the ability to filter the news articles according to different genre classes. For example, consider a financial analyst who tracks daily news about companies in which she is interested. It would be useful to be able to further classify these documents as being subjective or objective. One class of documents would present the latest news about the various companies

of interest, while the other class would contain the opinions of various columnists and analysts about these companies. Depending on circumstances, the user may require documents of one class or the other.

Another genre class with useful application is the ability to identify whether a document is describing something in a positive or negative way. This could be used to improve a recommender system. Products could be recommended on the basis that they were given a positive review by a reviewer with similar interests to the target user.

Another application of review classification is filtering of newswire articles for financial analysis. Financial analysts must quickly digest large amounts of information when making investment decisions. A delay of a few seconds in identifying important information can result in significant gains or losses. The ability to automatically identify whether news about a company is positive or negative would be a valuable service in such a situation [10].

The ability to filter documents according to the level of technical information presented and the readability of the document would enable a system to personalize documents retrieved according to the user's educational background. With a suitable set of genre classes, a system with a dual category structure that allowed users to browse documents according to both topic and genre would be useful. Genre analysis can facilitate improved personalization by recommending documents that are written in a style that the user finds interesting or a style that is appropriate to the users needs. We consider genre to be complimentary to topic as a method of recommendation. The two used in conjunction with each other can improve the quality of a user's recommendations.

In this article we make the following contributions:

- To investigate the feasibility of *genre classification using machine learning*. We wish to investigate whether machine learning can successfully be applied to the task of genre classification.
- To investigate how well *different feature-sets* perform on the task of genre classification. Using two sample genre classification tasks, we perform experiments using three different feature-sets and investigate which features satisfy the criteria for building good genre classifiers.
- To investigate the issues involved in building genre classifiers with *good domain transfer*. The task of genre classification

requires additional methods of evaluation. We introduce the idea of domain transfer as an indication of the performance of a genre classifier across multiple topic domains. We evaluate each of the feature-sets for their ability to produce classifiers with good domain transfer.

- To investigate how we can apply active learning techniques to build classifiers that perform well with *small amounts of training data*.
- To investigate methods of *combining multiple feature-sets* to improve classifier performance.

2 Genre Classification

In our introduction we gave a general outline of what we mean by genre. Here we define our interpretation in more detail, give several examples and compare our definition with previous definitions from related research.

2.1 What is genre?

The term "genre" occurs frequently in popular culture. Music is divided into genres based on differences in style, e.g. blues, rock or jazz. Sample genres from popular fiction include science fiction, mystery and drama. Genres are often vague concepts with no clear boundaries and need not be disjoint. For a given subject area there are no fixed set of genre categories. Identifying a genre taxonomy is a subjective process and people may disagree about what constitutes a genre, or the criteria for membership of a particular genre.

The American heritage dictionary of the English language defines genre as "A category of artistic composition, as in music or literature, marked by a distinctive style, form or content". Webster's revised unabridged dictionary defines a genre as "class; form; style esp. in literature". Wordnet defines genre as "1: a kind of literary or artistic work 2: a style of expressing yourself in writing 3: a class of artistic endeavor having a characteristic form or technique".

Swales [16] gives a working definition of genre. A genre is defined as a class of communicative events where there is some shared set of communicative purposes. This is a loose definition and any particular instance of a genre may vary in how closely it matches the definition. However instances of a genre will have some similarity in form or function.

Karlgren [5] distinguishes between a style and a genre. A style is a consistent and distinguishable tendency to make certain linguistic choices. A genre is a grouping of documents that are stylistically consistent and intuitive to accomplished readers of the communication channel in question.

From the different definitions we see that there is no definitive agreement on what is meant by genre. However, the common thread among these definitions is that genre relates to style. The genre of a document reflects a certain style rather than being related to the content. In general this is what we mean when we refer to the genre of a document: the genre describes something about what kind of document it is rather than what topic the document is about.

Genre is often regarded as orthogonal to topic. Documents that are about the same topic can be from different genres. Similarly, documents from the same genre can be about different topics. Thus we must separate the identification of the topic and genre of a document and try to build classifiers that are topic-independent. This contrasts with the aim of other text classification tasks, thus the standard methods of evaluating text classifiers are not completely appropriate. This suggests the notion of domain transfer - whether genre classifiers trained on documents about one topic can successfully be applied to documents about other topics.

We explicitly distinguish between the topic and style of the document. While assuming that genres are stylistically different, we investigate the effect of topic on our ability to distinguish genres. When we evaluate our genre classifiers, we measure how well they perform across multiple topic domains. In order for genre classification techniques to be generally useful, it must be easy to build genre classifiers. There are two aspects to this. The first is that of domain transfer: classifiers should be generally applicable across multiple topics. The second is that of learning with small amounts of training data. When building genre classifiers, we want to achieve good performance with a small number of examples of the genre class.

Genres depend on context and whether or not a particular genre class is useful or not depends on how useful it is for distinguishing documents from the users point-of-view. Therefore genres should be defined with some useful user-function in mind. In the context of the Web, where most searches are based on the content of the document, useful genre classes are those that allow a user to usefully distinguish between documents about similar topics.

To summarize, we view a genre as a class of documents that arises naturally from the study of the language style and text used in the document collection. Genre is an abstraction based on a natural grouping of documents written in a similar style and is orthogonal to topic. It refers to the style of text used in the document. A genre class is a set of documents written in a similar style which serves some useful discriminatory function for users of the document collection.

2.2 Sample Genre Classes

We focused on two sample genres which we use for our automatic genre classification experiments. These were two genres that we identified as functionally useful for web users. The first is whether a news article is subjective i.e. it presents the opinion of its author, or objective. The second is whether a review is positive or negative.

The first genre class we investigate is whether a document is subjective or objective. This is a common distinction in newspaper articles and other media. Many news articles report some significant event objectively. Other articles, which often take the form of columns or editorials, offer the author's opinion.

Consider the example of financial news. Financial news sites publish many articles each day. Articles of genre class fact may be reporting the latest stock prices and various events that are likely to influence the stock price of a particular company. Articles of genre class opinion may give the opinions of various financial analysts as to the implications of the events of the day for future stock prices. Different users at different times may be better served by articles from one genre or the other. It would be a useful service for the user to be able to filter or retrieve documents from each of these genre classes.

Our second sample genre class is classifying reviews as being either positive or negative. The ability to automatically recognize the tone of a review could have application in collaborative recommendation systems. For example, if a particular movie critic who generally has similar tastes to a user gives a film a positive review, then that film could be recommended to the user. Films could be recommended on the basis of how they are reviewed by critics that are known to have similar tastes to a particular user.

	Fact	Opinion
Football	Liverpool have revealed they have agreed a fee with Leeds United for striker Robbie Fowler - just hours after caretaker boss Phil Thompson had said that contract talks with the player were imminent.	The departure of Robbie Fowler from Liverpool saddens me but does not surprise me. What did come as a shock, though, was that the club should agree terms with Leeds, one of their chief rivals for the Championship.
Politics	Al Gore picked up votes Thursday in Broward County as election officials spent Thanksgiving weekend reviewing questionable presidential ballots.	Democrats are desperate and afraid. The reality that their nominee for President has a compulsive tendency to make things up to make himself look good is sinking in.
Finance	In a move that sent Enron shares higher after days of double-digit declines, Dynegy confirmed Tuesday that it is in talks to renegotiate its $9 billion deal to buy its rival.	The collapse of Enron is hard to believe, and even harder to understand. But in retrospect, there are some valuable lessons in the whole mess.

Table 1: Examples of objective and subjective articles from three topic domains

	Positive	Negative
Movie	Almost Famous:	Vanilla Sky:

	Cameron Crowe's first film since "Jerry Maguire" is so engaging, entertaining and authentic that it's destined to become a rock-era classic. Set in 1973, this slightly fictionalized, semi-autobiographical, coming-of-age story revolves around a baby-faced 15 year old prodigy whose intelligence and enthusiasm land him an assignment from "Rolling Stone" magazine to interview Stillwater, an up-and-coming band.	Presumably Cameron Crowe and Tom Cruise have some admiration for "Abre Los Ojos" the 1998 Spanish thriller from Alejandro Amenabar; why else would they have chosen to do an English-language remake? "Vanilla Sky", however shows that respect for ones source material isn't enough. It's a misbegotten venture that transforms a flawed but intriguing original into an elephantine, pretentious mess.
Restaurant	Though the New American menu at this neighbourhood treasure near Capitol Hill is ever changing, it's always beautifully conceived and prepared and based on mostly organic ingredients; the bistro dishes, paired with a fabulous, descriptive wine list, are served in an offbeat atmosphere.	Hidden in the back of a shopping mall near Emory, this Chinese eatery is so isolated that diners sometimes feel as if they're having a private meal out; the decor isn't much to look at and the foods nothing special but it's decent.

Table 2: Examples of positive and negative reviews from two topic domains

Tables 1 and 2 show a selection of document extracts from our document collection. A human reader can recognize a subtle difference in style between extracts from subjective and objective articles and similarly between the positive and negative reviews. We investigate techniques for automating this classification.

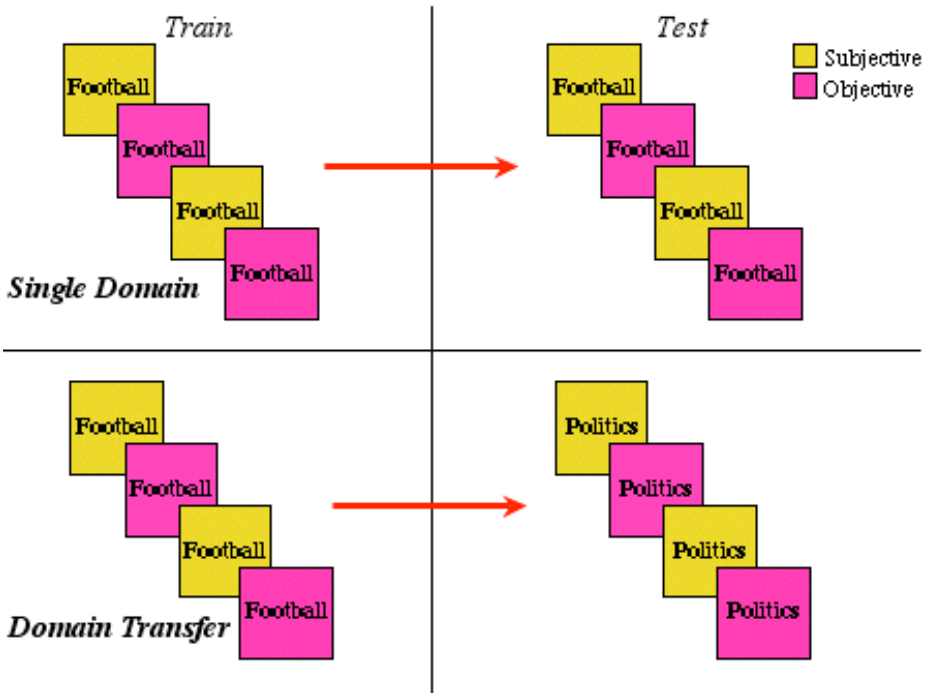

Figure 1: Genre classification for domain transfer

Our aim in constructing the classifier is to maximize accuracy both on a single topic and across topics. To this end we use datasets from three topic domains: football, politics and finance, for the subjectivity classification task and documents from two topic domains: movie reviews and restaurant reviews for the review classification task. We are interested in how well a classifier trained on documents from one domain performs in another. We identify the different topics in order to determine how well the classifier performs across multiple topics for the same genre classification task. If genre is orthogonal to topic we should be able to build classifiers that perform well on topics other than the one used to build the classifier. For example, a classifier built to recognize whether a document is subjective or objective by training it on documents about football should ideally be able to recognize

subjective documents that are about topics other than football such as finance or politics (Figure 1).

The practical effort involved in building a genre classifier is considerable. A human must label a number of examples to train the classifier. It is essential to minimize this human effort. Because of this we aim to build genre classifiers with good domain transfer. Because of the amount of human effort involved in constructing a genre classifier, it should be reusable across multiple topic domains. If it has to be retrained every time it is to be used in a new topic domain, the amount of work required to maintain it will be considerable and in a high volume digital library scenario could be prohibitive.

2.3 Related work

One of the two sample genres we study in our experiments is subjective vs. objective news articles. Wiebe [18] defines subjectivity classification as distinguishing sentences used to present opinions and evaluations from sentences used to objectively present factual information. She investigates subjectivity classification at the sentence level and concludes that the presence and type of adjectives in a sentence is indicative of whether the sentence is subjective or objective. We seek to perform subjectivity classification at the document level.

Tong [17] describes a system that focuses on tracking various entities and the opinions being expressed about them. The opinions are tracked by monitoring online public discussion forums. In particular, they monitor online discussion about movies and determine the level of "buzz" associated with specific movies as they move from announcement of release, through opening weekend and on to extended distribution. Opinions are extracted using sentiment models. These are patterns that capture the way people talk about movies and use a set of custom lexicons that cover personal emotions, movie features and language tone. These are also used to model the tone of the opinion. It appears that they use heuristics to identify positive and negative opinions being expressed about particular movies. Our positive vs. negative review task seeks to automate this classification process.

Stamatatos et al. [15] recognize the need for classifiers that can easily transfer to new topic domains, without explicitly mentioning domain transfer. However they do not elaborate on how to evaluate transfer. Their notion of genre is similar to ours.

Their feature-set is the most frequently occurring words of the entire written language and they show that the frequency of occurrence of the most frequent punctuation marks contains very useful stylistic information that can enhance the performance of an automatic text genre classifier. This approach is domain and language independent and requires minimal computation. They do not perform any experiments to measure the performance of their classifier when it is transferred to new topic domains.

This work is closely related to ours. They identify the need for domain transfer but do not develop this idea any further. Their definition of text genre is similar to ours and two of the genre classes they identify are similar to our subjectivity classification task. The features they use, namely stop-words and punctuation, are similar to our text statistics feature-set.

Kessler et al. [8] argue that genre detection based on surface cues is as successful as detection based on deeper structural properties. Argamon et al. [1] consider two types of features: lexical and pseudo syntactic. They compare the performance of function words against part-of-speech trigrams for distinguishing between different sets of news articles.

Roussinov et al. [14] view genre as a group of documents with similar form, topic or purpose, "a distinctive type of communicative action, characterized by a socially recognized communicative purpose and common aspects of form". This is a more general view of genre where genre is a grouping of similar documents. Some genres are defined in terms of purpose or function, others in terms of physical form while most documents combine the two.

They attempt to identify genres that web users frequently face and propose a collection of genres that are better suited for certain types of information need. To this end, they performed a user survey to see *1)* what is the purpose for which users search the web, and *2)* whether there was a relation between the purpose of a respondents search and the genre of document retrieved. This results in a proposed set of genres, along with a set of features for each genre and a user interface for genre based searching.

Dewdnew et al. [3] take the view that genre of a document is the format style. Genre is defined as a "label which denotes a set of conventions in the way in which information is presented". The conventions cover both formatting and the style of language used. They use two feature-sets: a set based on words (traditional bag-of-words)

and a set of presentation features which represent stylistic information about the document. The presentation features do consistently better than the word frequency features and combining the feature-sets gives a slight improvement. They conclude that linguistic and format features alone can be used successfully for sorting documents into different genres.

Rauber and Muller-Koller [13] argue that in a traditional library, non content-based information such as age of a document and whether it looks frequently used are important distinguishing features and present a method of automatic analysis based on various surface level features of the document. The approach uses a self-organizing map (SOM) [9] to cluster the documents according to structural and stylistic similarities. This information is then used to graphically represent documents. In this approach the genres are identified from clusters of documents that occur in the SOM rather than being defined in advance.

Karlgren [4, 6, 7] has done several experiments in genre classification. In [4] he shows that the texts that were judged relevant to a set of TREC queries differ systematically (in terms of style) from the texts that were not relevant.

In [6], Karlgren et al. use topical clustering in conjunction with stylistics based genre prediction to build an interactive information retrieval engine and to facilitate multi-dimensional presentation of search results. They built a genre palette by interviewing users and identifying several genre classes that are useful for web filtering.

The system was evaluated by users given particular search tasks. The subjects did not do well on the search tasks, but all but one reported that they liked the genre enhanced search interface. Subjects used the genres in the search interface to filter the search results. The search interface described is an example of how genre classification can usefully aid information retrieval.

3 Automated Genre Classification

The Machine Learning approach to document classification takes a set of pre-classified examples and uses these to induce a model which can be used to classify future instances. The classifier model is automatically induced by examination of the training examples. The human effort in this process is in assembling the labeled examples and choosing a representation for the training examples. A human must initially decide what features will be used to describe the training

examples, and represent the training documents with respect to these features.

When using Machine Learning algorithms, we first identify the concept to be learned. In our case this is the particular genre class we are attempting to classify. The output of the learning algorithm is a concept description that should ideally be both intelligible and operational. The concept description should be intelligible in the sense that it can be understood, discussed, disputed and interrogated by humans. It should also be operational in the sense that we can practically apply it to future examples.

The type of learning we are interested in is classification learning. In this learning scheme, the learner takes a set of labeled pre-classified examples. The learner is then expected to induce ways of classifying unseen examples based on the pre-classified examples given. This form of learning is supervised in that the training examples are provided and labeled by a human overseer.

The training data is a set of instances. Each instance is a single example of the concept to be learned. Instances are characterized by a set of attributes where each attribute measures a certain aspect of the concept being described. Attributes can be nominal or continuous. Continuous attributes represent some numerical value that can be measured. Nominal attributes are categorical. They assign the attribute to membership of a particular category.

Representing the classification problem as a set of instances is a restrictive way of formulating the learning problem. Each instance is characterized by values of a set of predetermined attributes. The selection of these attributes can affect the quality of the classifier produced by the learning algorithm. As part of our experiments, we are interested in identifying attributes which perform well on the genre classification task, can be easily extracted automatically and are useful across multiple topics. We use C4.5 [12] as our main learning algorithm. C4.5 is a machine learning algorithm that induces a decision tree from labeled examples and can easily be converted to a set of rules for a human to analyze.

We identify three different sets of features and investigate the utility of each of these for genre classification. Furthermore we attempt to identify the features which will lead to classifiers that perform well across multiple topic domains and can easily be built automatically. We use two sample genre tasks to test the utility of three sets of features for the purpose of automatic genre classification.

We emphasize the ability to transfer to new topic domains when building our classifiers and we evaluate different feature-sets for performance across multiple topic domains.

In addition to building classifiers that will transfer easily to new domains, we wish to minimize the effort involved in building a genre classifier. We wish to achieve good performance, that is, prediction accuracy as a function of amount of training data, with a minimum amount of labeled data. To this end we examine the learning rates of our classifiers and investigate methods of improving this learning rate using active learning techniques.

The three feature-sets investigated can be thought of as three independent views of the dataset. We investigate methods of combining the models built using each feature-set to improve classifier performance.

4 Features

We have explored three different ways to encode a document as a vector of features.

4.1 Bag-of-words

The first approach represented each document as a bag-of-words (BOW), a standard approach in text classification. A document is encoded as a feature-vector, with each element in the vector indicating the presence or absence of a word in the document. We wish to determine how well a standard keyword based learner performs on this task. This approach led to feature-vectors that are large and sparse. We used stemming [11] and stop-word removal to reduce the size of the feature vector for our document collection.

This approach to document representation works well for standard text classification where the target of classification is the topic of the document. In the case of genre classification however, the target concept is often independent of the topic of the document, so this approach may not perform as well.

It is not obvious whether certain keywords would be indicative of the genre of the document. We are interested in investigating how well this standard text classification approach works on the genre classification tasks. We expect that a classifier built using this feature-set may perform well in a single topic domain, but not very well when domain transfer is evaluated. By topic domain we mean a group of

documents that can be regarded as being about the same general subject or topic. For example, for the subjectivity classification task, we have three topic domains: football, politics and finance. For the review classification task we have two topic domains: restaurant reviews and movie reviews. The reason we identify different topic domains is that a text genre class may occur across multiple topic domains. We wish to evaluate the domain transfer of a genre classifier. For example, if a classifier is trained for the subjectivity classification task using documents from the football domain, how well does it perform when this classifier is transferred to the new domain of politics?

It is common in text classification, where the aim it to classify documents by content, to use a binary representation for the feature vector rather encoding the frequencies of the words occurrences. It is also common to filter out commonly occurring words as they do not usefully distinguish between topics. We are interested in measuring domain transfer so we choose the binary vector representation.

4.2 Part-of-Speech statistics

The second approach uses the output of Brill's part-of-speech (POS) tagger [2] as the basis for its features. It was anticipated that the POS statistics would reflect the style of the language sufficiently for our learning algorithm to distinguish between different genre classes. A document is represented as a vector of 36 POS features, one for each POS tag, expressed as a percentage of the total number of words for the document. The POS features are listed in table 3.

Tag	Description	Tag	Description
CC	Coordinating conjunction	PP$	Possessive pronoun
CD	Cardinal number	RB	Adverb
DT	Determiner	RBR	Adverb, comparative
EX	Existential there	RBS	Adverb, superlative
FW	Foreign word	RP	Particle
IN	Preposition or subordinating conjunction	SYM	Symbol
JJ	Adjective	TO	to
JJR	Adjective, comparative	UH	Interjection
JJS	Adjective, superlative	VB	Verb, base form
LS	List item marker	VBD	Verb, past tense
MD	Modal	VBG	Verb, gerund or present participle
NN	Noun, singular or mass	VBN	Verb, past participle
NNS	Noun, plural	VBP	Verb, non-3rd person singular present

NP	Proper noun, singular	VBZ	Verb, 3rd person singular present
NPS	Proper noun, plural	WDT	Wh-determiner
PDT	pre-determiner	WP	Wh-pronoun
POS	Possessive ending	WP$	Possessive wh-pronoun
PP	Personal pronoun	WRB	Wh-adverb

Table 3: Part-of-Speech features

This approach uses a part-of-speech representation of the documents rather than the actual words occurring in the document. It was hoped that this would give a representation that was still capable of discriminating genre but was independent of the subject of the document. The POS representation does not reflect the topic of the document, but rather the type of text used in the document.

We hope that POS features can be used to differentiate genres in a domain independent way. If the POS feature-set is capable of differentiating genre class, we would expect that it would do so in a domain independent manner as it doesn't have any information about the topic of the document.

4.3 Text Statistics

Our third approach is to use a set of shallow text statistics (TS). Many of these features were selected because they been shown to have discriminatory value between genre classes in the related literature. This feature set includes average sentence length, the distribution of long words, average word length. Additional features are based on the frequency of occurrence of various function words and punctuation symbols. Table 4 lists the features used.

Feature type	Features
Document level statistics	sentence length, number of words, word length
Frequency counts of	because been being beneath can can't certainly completely could couldn't did didn't do does

various function words	doesn't doing don't done downstairs each early enormously entirely every extremely few fully furthermore greatly had hadn't has hasn't haven't having he her herself highly him himself his how however intensely is isn't it its itself large little many may me might mighten mine mostly much musn't must my nearly our perfectly probably several shall she should shouldn't since some strongly that their them themselves therefore these they this thoroughly those tonight totally us utterly very was wasn't we were weren't what whatever when whenever where wherever whether which whichever while who whoever whom whomever whose why will won't would wouldn't you your
Frequency counts of various punctuation symbols	! " $ % & ' ( ) * + , - . : ; = ?

Table 4: Text statistic features

5 Experiments

We have evaluated the three feature-sets using two real-world genre classification tasks.

5.1 Evaluation

We evaluate our classifiers using two measures: accuracy of the classifier in a single topic domain and accuracy when trained on one topic domain but tested on another.

5.1.1 Single Domain Accuracy

Single domain accuracy measures the accuracy of the classifier when it is trained and tested on instances from the same topic domain. This measure indicates the classifier's ability to learn the classification task in the topic domain at hand.

Accuracy is defined as the percentage of the classifier's predictions that are actually correct as measured against the known classes of the test examples. Accuracy is measured using ten-fold cross-validation.

5.1.2 Domain Transfer Accuracy

Note that single-topic accuracy give us no indication of how well our genre classifier will perform on documents from other topic domains. We introduce a new evaluation measure, domain transfer, which indicates the classifier's performance on documents from other topic domains.

We measure domain transfer in an attempt to measure the classifier's ability to generalize to new domains. For example, a genre classifier built using documents about football should be able to recognize documents about politics from the same genre. Domain transfer is essential in a high volume digital library scenario as it may be prohibitively expensive to train a separate genre classifier for every topic domain.

We use the domain transfer measure as an indicator of the classifier's generality. It also gives us an indication of how much the genre classification task in question is topic dependent or topic independent.

Domain transfer is evaluated by training the classifier on one topic domain and testing it on another topic domain. In addition to measuring the domain transfer accuracy, we can calculate the domain transfer rate. This measures how much the classifier's performance degrades when the classifier is evaluated on new topic domains. A classifier that performs equally well in the transfer condition as in a single domain would achieve a transfer score of 1. A classifier whose performance degrades when transferred to new topic domains would achieve a transfer score of less than 1.

Consider a classification task consisting of a learning algorithm C and a set of features F. Let $D_1, D_2, ..., D_n$ be a set of topic domains. Let ${}_{D}A_{D}$ be the performance of C when evaluated using ten-fold cross-

validation in domain D. Let ${}_{D_1}A_{D_2}$ denote the performance of classification scheme C when trained in domain D_1 and tested in domain D_2 . We will use accuracy as our measure of performance. We define the domain transfer rate for classification scheme C as

$$ {}_{C}DT_{F} = \frac{1}{n(n\text{-}1)} \sum_{i=1}^{n} \sum_{j=1; j \neq i}^{n} \left(\frac{{}_{D_i}A_{D_j}}{{}_{D_i}A_{D_i}} \right) \quad (1) $$

We evaluate the quality of a genre classifier using both single domain accuracy and domain transfer accuracy. Ideally we would hope to get high single domain accuracy and a high domain transfer rate (see figure 2). A classifier with high accuracy and low transfer may be useful in some situations but not in others.

5.2 Experimental Setup and Document Corpora

Table 5 shows the number of documents in the corpus used for the subjectivity classification experiment. We identified three topic domains, football, politics and finance. For each of these topics, we identified a number of web sites that specialize in news from the particular topic domain. Articles were then automatically spidered from these sites over period of several weeks. The documents were then classified by hand by the author as being either subjective or objective.

Topic	Opinion	Fact	Total
Football	174	177	351
Politics	144	145	289
Finance	56	100	156

Table 5: Corpus details for the subjectivity classification experiment

Table 6 shows the details of our document collection for the review experiment. The collection of review datasets was somewhat easier than the collection of the subjectivity datasets. The reason is that the classification of a document could be extracted automatically using a

wrapper for the particular site. For example, most movie reviews come with a recommendation mark. A review that awards a film 4 stars could be considered a positive review, while a review that awards a film 1 star could be considered a negative review. Thus we automatically extract the classification of a particular review, negating the need to manually classify each document.

The Movie reviews were downloaded from the Movie Review Query Engine[1] . This site is a search engine for movie reviews. It extracts movie reviews from a wide range of sites. If the review contains a mark for the film, the mark is also extracted. We wrote a wrapper to extract a large number of movie reviews and their corresponding marks from this site. The marks from various sites were normalized by converting them to a percentage and then we used documents with high percentages as examples of positive reviews and vice versa. Marks below 41 were considered negative while marks of 100% were considered positive. Reviews with marks in the range 41-99 were ignored as many of them would require a human to label them as positive or negative.

The restaurant reviews were gathered from the Zagat survey site[2] . This is a site that hosts a survey of restaurants from the U.S.A. and Europe. Users of the site submit their comments about a particular restaurant and assign marks in three categories (food, decor and service). The marks for these categories are between 1 and 30 and are the average for all the users that have provided feedback on that particular restaurant. The reviews themselves consist of an amalgamation of different users comments about the restaurant. We averaged the marks for the three categories to get a mark for each restaurant. Restaurants that got an average mark below 15 were considered negative while those getting marks above 23 were considered positive.

Topic	Positive	Negative	Total
Movie	386	337	723
Restaurant	300	331	631

[1]http://www.mrqe.com
[2]http://www.zagat.com

Table 6: Corpus details for the review classification experiment

5.3 Evaluation Methods

The standard method of evaluating Machine Learning classifiers is to use cross validation. We believe that for the task of genre classification, this alone is not sufficient. An extra method of evaluation is needed. In order to test whether genre classes are orthogonal to topic we need to measure the classifiers performance across topics as well as across genres. We use standard cross validation to measure accuracy[3] within a single topic domain. We also propose a domain transfer measure to evaluate the classifier's ability to generalize across topic domains.

Figure 1 shows what we mean by domain transfer. In the single domain case, the classifier is trained and tested on documents from the same topic domain. In the domain transfer case, the classifier is trained on documents from one topic domain and tested on documents from another topic domain.

Usually text classification is applied to tasks where topic specific rules are an advantage. In order to scale with large numbers of topics, this is not the case for genre classification. In the case of genre classification, topic specific rules reduce the generality of the genre classifier. In addition to evaluating the genre classifier's performance in a single topic domain, we also need to evaluate its performance across multiple topic domains.

[3]Other measures could be used such as precision, recall or F-measure

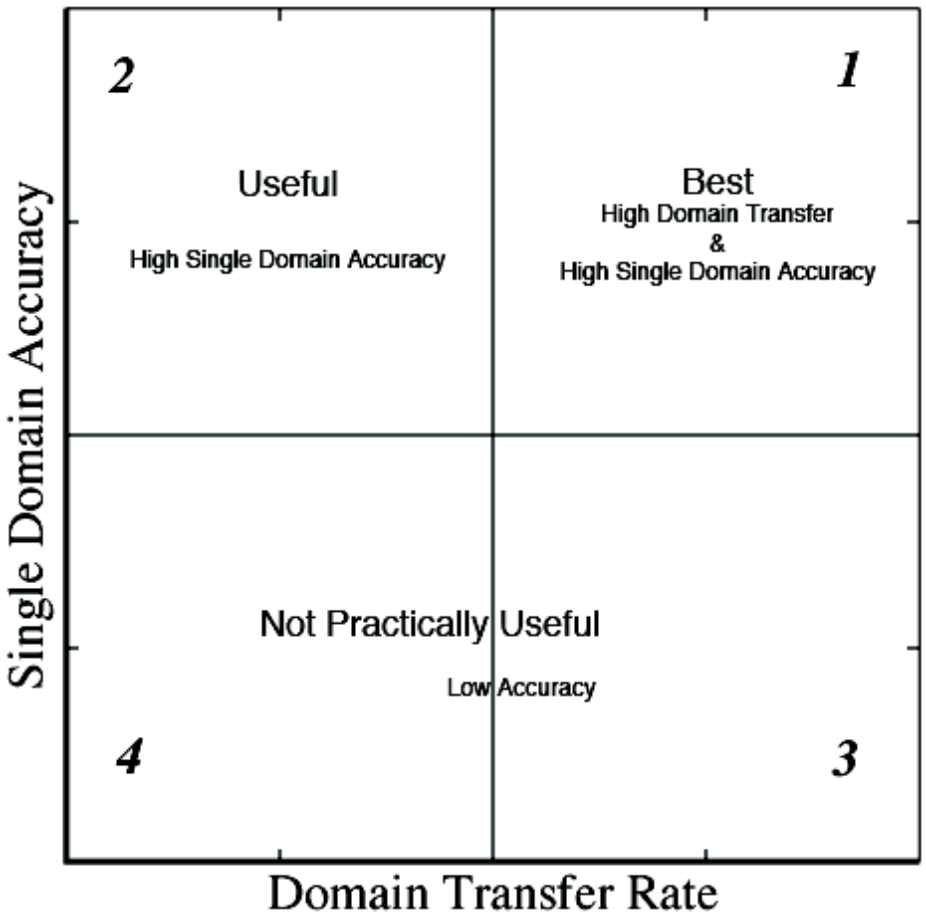

Figure 2: Desirable performance characteristics of a genre classifier. A useful genre classifier should have both high single domain accuracy and high domain accuracy(region 1). Performance in region 2 can be useful but performance in region 3 or 4 is undesirable.

Figure 2 shows single domain accuracy plotted against domain transfer rate, with areas of the graph labeled in order of desirability. The most desirable classifiers would be in region 1 of the graph. These classifiers have both high single domain accuracy and high transfer rate. Next most desirable are classifiers occurring in region 2 of the graph. These classifiers have high single domain accuracy but poor domain transfer. Classifiers occurring in regions 3 and 4 are undesirable because they have poor levels of single domain accuracy and even if they have good transfer rates, they are not useful in practice.

5.4 Choice of learning algorithm

We chose C4.5 as our learning algorithm. The main reason for choosing C4.5 is that it generated a model that can easily be converted into a set of rules that can be examined by a human observer. We performed some initial experiments using different learning algorithms but non of them were substantially superior to C4.5 on these tasks.

5.5 Single Domain Experiments

	BOW	POS	TS	MVE
Football	86.7	82.4	86	88.6
Politics	85.3	83.1	80.3	90.8
Finance	89.7	88.6	83.3	92
Average	87.2	84.7	83.2	90.5

Table 7: Single domain accuracy for subjectivity classification

	BOW	POS	TS	MVE
Movie	76.8	59.6	59	74.1
Restaurant	88.5	62.9	94.1	83.4
Average	82.7	61.3	76.6	78.8

Table 8: Single domain accuracy for review classification

Table 7 shows the single domain experiments for the subjectivity classification task (Note: the MVE approach shown is described in section 6.1). The BOW feature-set performs best in all three topic domains. The POS feature-set is second best on average, although the

difference between it and the TS feature-set is insignificant. All three feature-sets achieve good accuracy on this classification task, indicating that any of these feature-sets alone is sufficient for building classifiers within a single topic domain. However BOW is the best performing feature-set on this task within topic domains. This indicates that there are keywords within each topic domain that indicate the subjectivity of a document.

Table 8 shows the single domain results for the review classification experiment. In both domains, the BOW approach performs significantly better than the POS approach. On average, the BOW approach achieves accuracy of 82.7%. This is a good level of accuracy for this classification task. The POS approach performs poorly in comparison (61.3% on average).

The BOW approach is capable of achieving good levels of performance when attempting to classify reviews as positive or negative in a single topic domain. The POS approach performs poorly on this classification task, even in a single topic domain. The TS approach performs well in the restaurant domain (94.1) but poorly on the movie domain (59). Thus while it's average performance is good, it does not perform consistently well in each domain.

5.6 Domain Transfer Experiments

Train	Test	BOW	POS	TS	MVE
Football	Politics	58.5	74	63.7	72.3
Football	Finance	61.5	78.8	75.6	80.8
Politics	Football	76.9	70.7	64.1	76.6
Politics	Finance	66.7	90.4	66.7	75.6
Finance	Football	76.9	73.2	70.7	81.5
Finance	Politics	63	83.7	66.1	76.9
Average		67.3	78.5	67.8	77.3

Table 9: Domain transfer for subjectivity classification

Train	Test	BOW	POS	TS	MVE
Movie	Rest	40.1	44.4	50.4	45.3
Rest	Movie	55.5	49.8	44.3	52.9
Average		47.8	47.1	47.35	49.1

Table 10: Domain transfer for review classification

Table 9 shows domain transfer results for the subjectivity classification task. In this case POS feature-set performs best (78.5), while the BOW feature-set performs worst (63.7). So, BOW goes from being best when evaluated in a single topic domain to worst when evaluated across multiple topic domains.

This indicates that while keywords can be used to identify subjective documents, a model built using these features is more closely tied to the document collection used for training. Intuitively we would expect that the classifier built using the POS statistics as features would have a more generalizable model of what constitutes genre than one built using keywords or domain-specific hand-crafted features.

Table 10 shows the domain transfer results for the review classification experiment. On average, each feature-set performs to a similar level with there being less than 1% between them. Each feature-set achieves average accuracy of around 47%. This level of performance is no better than that achievable by a simple majority classifier.

The single domain experiment on this classification task showed that BOW can achieve high levels of accuracy on this classification task in a single topic domain. However the domain transfer experiment shows that the BOW approach fails when the transfer approach is evaluated. The BOW features which indicate a positive movie review are not transferable to the restaurant domain and vice versa. The POS approach fails in both the single domain and domain transfer experiments.

We conclude that the POS approach is not suitable for the task of classifying reviews as being either positive or negative. The BOW approach can achieve good performance in a single topic domain but cannot transfer to new topic domains. Even though the traditional means of evaluating a classifier indicate that the BOW achieves good performance, our experiments indicate that it performs poorly when we our extra domain transfer condition is evaluated.

5.7 Discussion

Our experiments show that it is possible to build genre classifiers that perform well within a single topic domain. However, single domain performance can be deceiving. When we further evaluate the classifiers for domain transfer performance, it becomes clear that good domain transfer is more difficult to achieve.

The review classification task is more difficult than the subjectivity classification task. All feature-sets achieved good single domain accuracy on the latter task, while the POS feature-set also achieved good domain transfer. On the review classification task, the BOW approach achieved good single domain accuracy, but none of the feature-sets achieved good domain transfer.

From examination of the dataset, reviews from the movie domain are easily recognizable by a human reader as being either positive or negative. It is more difficult to discern the category for many of the restaurant reviews. Recall that the reviews were classified automatically, based on scores extracted from the source website. The restaurant reviews consisted of an amalgamation of user comments about particular restaurant. For many of these reviews it is difficult for a reader to decide whether they are positive or negative. Because they combine different user comments, the style of the restaurant reviews is different from the style of the movie reviews which are written by individual authors. This may account for some poor performance when domain transfer was evaluated for the review classification task.

It is also clear that no one feature-set is suitable for both genre classification tasks. The BOW feature-set performs well in a single topic domain, while the POS feature-set performs best on the subjectivity classification task when we evaluate domain transfer.

5.8 Models generated

We can examine the models generated by C4.5 to see what features the classifier is using to make its prediction. This may give us some insight into the classification task in question and give us confidence in the model being generated. If the model gives us rules that seem intuitively related to the genre classification task, this gives us confidence in the validity of the model. The model may also give us insight into the genre class under investigation and which features differentiate genre but are not obvious from inspection of the data.

The root node of a C4.5 decision tree is the attribute that was deemed most informative with respect to discriminating between the target classes. For the subjectivity classification task, the BOW approach generated root nodes based on the words 'columnist', 'column' and 'column' for the football, politics and finance domains respectively. The presence of these words is strongly indicative of a document being subjective. It is easy to see that documents containing these words are likely to be subjective in style as they are probably written by a particular columnist, giving their opinion. For the review classification task, the BOW approach generated root nodes based on the words 'jolie' and 'romantic' for the movie and restaurant domains respectively. The word 'jolie' occurring in a movie review means it is likely to negative, while the word 'romantic' occurring in a restaurant review means it is likely to be positive.

This corresponds the fact that the movie 'Tomb Raider' starring Angelina Jolie was released around the time the dataset was collected. One can imagine that this is not the kind of movie that would appeal to film critics and would be likely to garner negative reviews. It also seems unlikely that this attribute would have any discriminatory value in the restaurant review domain.

It also seems reasonable that the word 'romantic' used in relation to a restaurant is likely to indicate a positive review. However this attribute may in fact penalize the classifier when transferred to the movie domain as it seems plausible that movie reviews containing the word 'romantic' are more likely to be negative rather than positive.

For the subjectivity classification task, the POS approach generates trees with root nodes DT, RB and RB for the football, politics and finance domains respectively. DT refers to the distribution of determiners (e.g. as, all, any, each, the, these, those). RB refers to adverbs (e.g. maddeningly, swiftly, prominently, predominately). Subjective documents tend to have relatively more determiners and

adverbs. On the review classification task, the POS approach failed to accurately discriminate between positive and negative reviews.

The TS approach generates trees with root nodes based on the number of words in the document for the football and politics domains and the distribution of the word ‘can’ for the finance domain. Shorter documents are more likely to be objective. It seems likely that objective documents will often be much shorter than subjective documents as they just report some item of news, without any discussion of the event involved. It is not clear how the distribution of the word ’can’ is indicative of the subjectivity of a document. On the review classification task, the TS approach did not perform well in the movie domain (59), but performed surprisingly well on the restaurant domain (94.1). In this case the root node of the generated tree is the number of long words in the document. Reviews containing a small number of long words are more likely to be negative.

6 Combining multiple views

We have investigated the use of three different feature-sets for the task of genre classification and attempted to determine which features are better for building general, transferable classifiers. Our experiments have shown that the utility of each feature-set depends on the genre classification task at hand. We seek to automate as much as possible the process of building a genre classifier. None of the feature-sets are obviously generally superior to the others and it is undesirable to have to determine the best feature-set for every new genre classification task. To further automate the construction of genre classifiers, we investigate methods of improving performance by combining feature-sets.

6.1 An ensemble learner

We can treat the three different feature-sets as different independent views of the data. We can build a meta-classifier that combines evidence from classifiers built using each feature-set to make it’s prediction.

There are several methods of combining classifiers to make predictions. Bagging combines the predictions of several separate models. The models are built using different subsets of the training data and each model votes on the final prediction. Boosting is similar to bagging except that the votes of each model are weighted according to some scheme such as the models success on the training data.

While bagging and boosting combine models of the same type, stacking combines models built using different learning models.

Our approach differs from these in that we will combine models based on our different feature-sets. This multi-view ensemble learning approach builds a model based on each of the three feature-sets. A majority vote is taken to classify a new instance.

The results achieved by the ensemble learner are encouraging. For the subjectivity classification task the results (Table 7) achieved by this approach (MVE) are better that those achieved by any of the individual feature-sets. The domain transfer (Table 9) is almost as good as that achieved by POS, and significantly better that that achieved by the other feature-sets.

For the review classification task (Table 8) this approach performs better than POS and TS, but not a good as BOW. In the domain transfer case (Table 10), this approach performs best on average.

This approach to classification exploits the fact that the three different feature-sets do not all make mistakes on the same documents. So a mistake made by the model based on one feature-set can be corrected by the models based on the other feature-sets. This works best in situations where all three feature-sets achieve good performance, such as the subjectivity classification task. When each feature-set performs well, they are more likely to correct each others mistakes.

In cases where some of the feature-sets perform poorly (such as the review classification task), this approach will achieve performance that is proportional to the relative performance of the individual feature-sets.

It seems likely that for genre classification tasks where it is not clear which feature-set is most suitable for the task, this approach will increase the likelihood of the classifier performing well.

6.2 Multi-view selective sampling

We wish to minimize the human effort involved in building a new genre classifier. To this end we wish to actively select documents for labeling such that we achieve better performance with less training data.

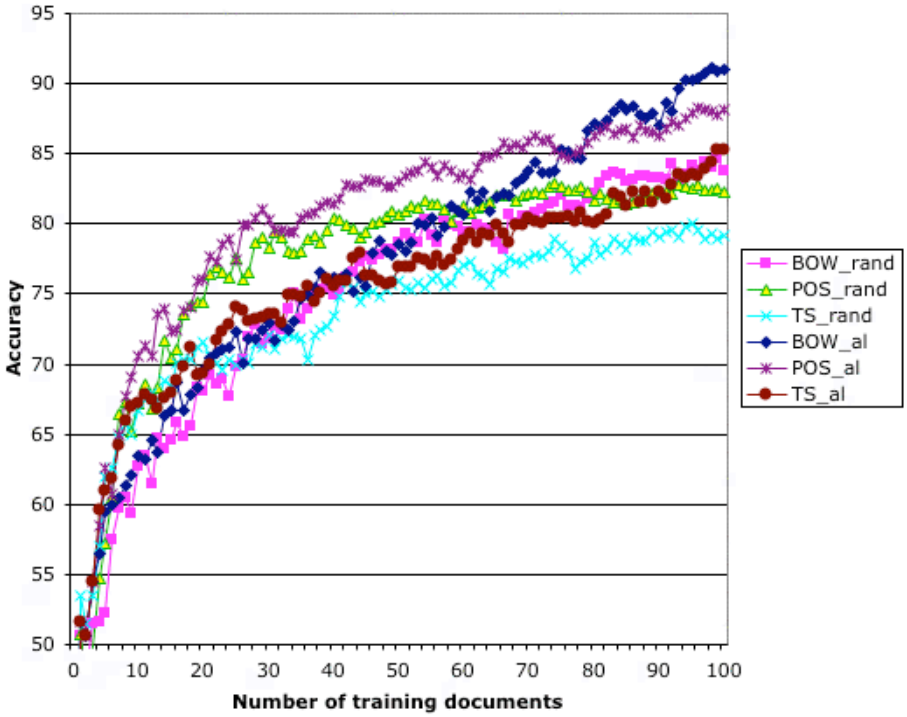

Figure 3: Multi-view selective sampling on the subjectivity classification task

Figure 3 shows the first 100 points of the learning curves for the subjectivity classification task averaged over the three topic domains. The vertical axis shows the average accuracy taken over ten trials. The horizontal axis shows the number of training documents.

The naive way of choosing documents to add to the training set is to choose a document at random. This approach is shown for each of the three feature-sets (BOW_rand, POS_rand and TS_rand). The POS learning rate is better on this task than the learning rate for the other two feature-sets.

We seek to improve the learning rate of the genre classifiers. One method of improving the learning rate is to use active learning to selectively sample documents to add to the training set. The aim is to select training documents that will most improve the learned model, thus achieving maximum performance for minimal training data.

The approach we investigate to improve the learning rate uses the level of agreement between the learned models based on different feature-sets. The first document to be labeled is selected at random. The next document to be labeled is the one where the models based on the three different feature-sets disagree most about the classification.

Applying this approach to our subjectivity classification task gives an improvement in learning rate for all three feature-sets (BOW_al, POS_al, TS_al). For each feature-set, there is little difference between the random and active learning approaches initially. However as the classification accuracy improves, the active learning approach begins to exhibit a better learning rate that the random approach. This indicates that the active learning approach consistently chooses documents that improve the performance of the classifier.

7 Conclusion

In theory, genre and topic are orthogonal. However, our experiments indicate that in practice they partially overlap. It may be possible to automatically identify genre in a topic independent way, but the results of our domain transfer experiments show that the feature-sets we investigate result in models that are partially topic dependent.

From a single topic point of view, our approach was very successful. If we used only the usual methods of evaluation, we would conclude that genre classification is not a difficult task and can easily be achieved using standard machine learning techniques. On the subjectivity classification task, all our feature-sets achieved high accuracy, while on the review classification task a standard bag-of-words approach achieved good accuracy.

We have argued that standard methods of evaluation are not sufficient when evaluating genre classifiers and that in addition the genre classifier's ability to transfer to new topic domains must also be evaluated. When we evaluate this additional aspect of the genre classifiers, we find that it is difficult to build classifiers that transfer well to new domains.

For the subjectivity classification task we have shown that it is possible to build a genre classifier that can automatically recognize a document as being either subjective or objective. High accuracy in a single topic domain can be achieved using any of the three feature-sets we investigated (BOW, POS or TS) but when domain transfer is measured for this task, the POS feature-set performs best. Overall, the

POS feature-set is best for this genre classification task as it performs well both in a single topic domain and when transferred to new topic domains.

The review classification task is more difficult. Good accuracy can be achieved in a single topic domain using the BOW approach. The POS approach is not suitable for this genre classification task. All three feature-sets fail to achieve good domain transfer on this task.

We also investigated methods of combining the predictions of models based on the different feature-sets and show that this improves performance. This approach is perhaps best when approaching a new genre classification problem, where it is not clear which feature-set is most suitable for the task.

We also show that the learning rate of the genre classifier can be improved by actively selecting which document to add to the training set. This selection is based on the level of disagreement of models built using each feature-set.

These two approaches further facilitate the aim of automating as much as possible the process of building genre classifiers. All three feature-sets can be extracted automatically. The ensemble learning approach can give good performance on the genre classification task and the active learning approach can improve performance on small amounts of training data.

Future work

We identified two sample genre classification tasks. These particular genre classes could be usefully applied to improve existing information retrieval systems. Applications that utilize genre classification to provide noticeable benefits to the end user must be developed to determine whether genre classification can be a useful, practical technique for improving document retrieval systems.

In building such systems it will be useful to identify additional genres that can improve a users ability to filter documents and reduce the number of documents that are potentially relevant to them. An expanded genre taxonomy is needed together with appropriate techniques for automatically identifying genres. We found that the techniques that were successful on one genre classification task (subjectivity classification), were less successful on another genre classification task (review classification).

The ability to achieve good domain transfer is important for genre classifiers. The techniques we used did not provide a complete separation of genre and topic. Further investigation is needed to determine methods of identifying genre in a topic independent way. We also need to refine methods of evaluating domain transfer and determine how to meaningfully compare the performance of different genre classifiers.

Ideally once a general genre taxonomy is defined we need techniques for automatically constructing genre classifiers within this taxonomy. One would hope that there are general techniques that could be used to build all classifiers for all genres within a taxonomy and that these genre classifiers will transfer easily to new topic domains. However, our experience has shown that this is difficult and methods for achieving these aims need further investigation.

Other feature-sets could be generally useful for building genre classifiers. The addition of further feature-sets may also improve the performance of the ensemble learner and active learning approaches.

In general future work consists of extending the work we have done on two genre classification tasks to a general genre taxonomy. Classifiers built to identify genre classifiers within this genre taxonomy should be easy to build and domain independent. The other major area for future work is to implement applications that use genre classification to improve the users experience.

Acknowledgments

This research was funded by Science Foundation Ireland and the US Office of Naval Research. Thanks to Barry Smyth for his advice and assistance.

References

[1] Shlomo Argamon, Moshe Koppel, and Galit Avneri. Routing documents according to style. In *First International Workshop on Innovative Information Systems*, 1998.

[2] Eric Brill. Some advances in transformation-based parts of speech tagging. In *AAAI*, 1994.

[3] Nigel Dewdney, Carol VanEss-Dykema, and Richard McMillan. The form is the substance: Classification of genres in text. In *ACL Workshop on Human Language Technology and Knowledge Management*, 2001.

[4] J. Karlgren. Stylistic experiments in information retrieval. In T. Strzalkowski, editor, *Natural Language Information Retrieval*. Kluwer, 1999.

[5] Jussi Karlgren. The wheres and whyfores for studying text genre computationally. In *Style and Meaning in Language, Art, Music and Design*, Washington D.C., 2004. AAAI Symposium series.

[6] Jussi Karlgren, Ivan Bretan, Johan Dewe, Anders Hallberg, and Niklas Wolkert. Iterative information retrieval using fast clustering and usage-specific genres. In *Eight DELOS workshop on User Interfaces in Digital Libraries*, pages 85–92, Stockholm, Sweden, 1998.

[7] Jussi Karlgren and Douglass Cutting. Recognizing text genres with simple metrics using discriminant analysis. In *Proceedings of the 15th. International Conference on Computational Linguistics (*COLING 94*)*, volume II, pages 1071–1075, Kyoto, Japan, 1994.

[8] Brett Kessler, Geoffrey Nunberg, and Hinrich Schutze. Automatic detection of text genre. In *ACL/EACL*, 1997.

[9] Teuvo Kohonen. The self-organising map. *Proceedings of IEEE*, 78(9):1464–1479, 1990.

[10] V. Lavrenko, M. Schmill, D. Lawrie, P. Ogilvie, D. Jensen, and J. Allan. Mining of concurrent text and time-series. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2000.

[11] M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[12] Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufman, 1993.

[13] A. Rauber and A. Muller-Kogler. Integrating automatic genre analysis into digital libraries. In *First ACM-IEEE Joint Conf on Digital Libraries*, 2001.

[14] Dmitri Roussinov, Kevin Crosswell, Mike Nilan, Barbara Kwasnik, Jin Cai, and Xiaoyong Liu. Genre based navigation of the web. In *34th International Conference on System Sciences*, 2001.

[15] E. Stamatatos, N. Fakotakis, and G. Kokkinakis. Text genre detection using common word frequencies. In *18th International Conference on computational Linguistics*, 2000.

[16] John M. Swales. *Genre Analysis*. Cambridge University Press, 1990.

[17] Richard M. Tong. An operational system for detecting and tracking opinions in on-line discussions. In *SIGIR Workshop on Operational Text Classification Systems*, 2001.

[18] Janyce M. Wiebe. Learning subjective adjectives from corpora. In *AAAI*, 2000.