Urdu Text Genre Identification

Farah Adeeba, Sarmad Hussain, Qurat-ul-Ain Akram[1]
Center for Language Engineering
[1]ainie.arkram@kics.edu.pk

Abstract

Automatic genre identification of document is becoming ever-increasing important since the availability of more and more text in digital form. This study describes automatic Urdu text genre identification by evaluating state of the art classifiers on different sets of features. The features are extracted by structural and lexical analysis of Urdu text. In addition, term frequency and inverse document frequency of features are computed. Different types of experiments are performed on two types of Urdu text corpora to evaluate the features set and classifiers for automatic Urdu text genre identification system. SVM classifier outperforms irrespective of the features set . The experimental analysis reveals that lexical features are more effective than structural features, and significantly improve the genre identification accuracy.

1. Introduction

Automated genre identification deals with prediction of genre of an unknown text, independent of its topic and style. With the tremendous increase in digital data, automated genre identification is becoming important for information retrieval to classify huge text into different categories. In addition, automatic genre classification of a document also helps to retrieve relevant documents according to the user requirements. It also plays an important role to improve the performance of Natural Language Processing (NLP) applications including part of speech (POS) tagging, parsing and word sense disambiguation system [1].

Urdu belongs to Arabic script and is national language of Pakistan, spoken by more than 100 million people[4]. It has rich morphology i.e. word has many surface forms, requires five agreement (case, gender, respect, number, person**)**. In this paper, Urdu text genre classification technique is presented which automatically classifies text into culture, science, religion, press, health, sports, letters and interviews genres. This article investigates the impact of structural and lexical information for the genre identification. In addition, different machine learning techniques including Support Vector Machines(SVM), Naive Bayes and Decision trees are also evaluated. The proposed systems are objectively evaluated by using two different corpora to explore the effect of corpora size in genre identification. A series of experiments show that lexical level approach is quite effective and significantly improves the identification accuracy.

The rest of this paper is organized as follows. Section 2 gives an overview of state of the art techniques for automatic genre identification. Section 3 presents the methodology of the genre identification of Urdu documents, which involves two main phases;(1) extraction of structural and lexical features, and (2) classification of Urdu documents based on extracted features. In addition, text pre-processing is also discussed in Section 3. Urdu benchmark corpora used in this study are described in Section 4. System experiments along with results are presented in Section 5. Finally, Section 6 concludes the research findings of this study.

2. Related Work

Efforts have been carried out for automatic text genre identification using rule based and statistical techniques. The documents have different features which are used to segregate the genre of these documents. These features are used to classify the genres using machine learning algorithms. Classifiers suggest the genre based on the computed features. In literature structural, lexical, sub lexical and hybrid features are being used for genre identification.

Lexical features are usually computed in terms of word frequencies in context of a document, and other documents (TF-IDF). Stamatatos et al. [2] developed a method of text genre identification by using common word frequencies. Wall street journal of the year 1989 is used as dataset consisting of 2560K words, which is further divided into training and testing files. 640k text

[4] http://www.ethnologue.com/language/urd

files are used in training phase and 16k text files are used for testing. For genre classification, 30 most frequent words of British National Corpus (BNC) are used. The Discriminant analysis is performed to extract most frequent words from training corpus. The experimental results show that 30 most frequent words of BNC corpus play important role by giving 2.5% error rate, as compared to the words extracted from training corpora.

Lee and Myaeng[3] proposed genre classification by using word statistics from different class sets, genre and subject class. Goodness value of a term for a genre is computed in two ways; (1) term's document frequency ratios for genre and subject class, and (2) term frequency (TF) ratios. These term's document frequency ratios and term frequency are used to compute the probability of a document belonging to a genre. Dataset of English and Korean languages are used for this study. A total of 7,615 of English and 7,828 documents of Korean are collected from web consisting of reportage, editorial, technical paper, critical review, personal home page, Q&A, and product specification. Naive Bayesian and similarity approaches are used for genre identification. Results show that term frequency (TF) ratios performs well as compared to the document frequency ratios.

Lexical features are also used for Urdu text classification [4] to categorize the text document into six genres. They compute different words based statistical features from corpus. These features are classified using Naive Bayes and SVM. The experimental results show SVM performs well with reasonable genre identification accuracy. Zia et al.[5] also used lexical level information for Urdu text classification using different classifiers including k-NNs, SVM and decision trees. The system has 96% f-measure to label genre of document among one of four genres.

Structural features are usually extracted from POS, phrase related and chunk related information. Lim et al.[6] used structural features for genre identification of web documents. A total of 1,224 documents are used to extract POS, phrase and chunk level features. K-Nearest-Neighbor algorithm is used to classify the genre by using these structural features. The accuracy of the system is 36.9%, 38.6% and 37% for POS, phrase and chunk level features respectively. Such Structural features cannot be applied for resource scarce languages which have very limited annotated language resource.

Classification accuracy has direct relationship with dataset size. The change in sample size results into direct change in classification accuracy. Sordo and Zeng[7] investigated the dependency between sample size and classification accuracy. It is observed that classification accuracy increases with the increase in dataset size. Irrespective of size, dataset itself plays an important role for genre classification.

Accurate genre classification needs minimal overlapping between genres' lexical and contextual information. Kanaris and Stamatatos[9] used two different corpora for genre identification. Both corpora have different accuracies.

In addition, number of genres has significant impact on genre classification. Limited number of genres results into higher accuracy. Moreover, significant training data also improve the genre identification accuracy.

Ali and Ijaz [4] used lexical features for Urdu text classification using SVM and Naive Bayes methods. Maximum accuracy i.e. 93.34% is achieved by using SVM. Presented system classify document in one of six categorize. Zia et. al. [5] used different feature selection approaches for Urdu text classification. Lexical features are used for classification of document into four genres. The available techniques for Urdu are mainly focused on the evaluation of machine learning classification approaches by using lexical features. In this paper, the effect of lexical and structural features is observed by applying different machine learning classification approaches. In addition, two different text corpora are used to investigate the impact of training dataset size variability on genre classification.

3. Methodology

In this paper, impact of lexical and structural features is analyzed for the development of Urdu text genre identification system. Urdu words unigram and bigrams are extracted as lexical features whereas words POS and words sense information is used as structural information. These features are classified using state of the art machine learning classification approaches. The details of each phase of the presented technique are discussed in sub sequent sections.

3.1 Text pre-processing

In this paper, two different datasets are used which are discussed in detail in Section 4. Dataset-1 is cleaned, POS tagged, sense tagged corpus which is manually distributed into respective genres by expert linguists. Dataset-2 is large corpus in size which is only manually distributed into genres by linguists.

Therefore, preprocessing is applied to extract features from these corpora. The details are given below.

3.4.1. Corpus cleaning.

As mentioned above, Dataset-1 is cleaned therefore Urdu words unigram and bigram extraction is straight forward. Words are extracted by tokenizing the corpus on space. But, in Dataset-2 the extracted words list has some issues. This list is manually analyzed and an automatic corpus cleaning is developed by applying different heuristics to clean the corpus. Some examples of Urdu space insertion issues are discussed below. Heuristics to resolve these issues are also discussed.

1. عرصہ۳۰ سال سے پی ٹی سی ٹیچر
2. دنیا بھر میں موجود 65کے لگ بھگ باقاعدہ اوپن یونیورسٹیوں کے ساتھ
3. برطانیہ میںUKOU کے قیام کے محض تین سال بعد
4. HKEY_LOCAL_MACHINE\SOFTWARE\Microsost\Windows\CurrentVersion\Explorer\BitBucketاب

Due to missing space between Urdu digits and text, the two words e.g. ۳۰ سال are treated as single word. Same type of issue is observed between sequence of Latin digits and Urdu text e.g. 65کے.The space is inserted at start and end of sequence of Urdu/Latin digits to resolve this issue. There are also examples of joined Latin character sequence with Urdu text e.g. میںUKOU in Example 3. These Latin strings can be URLs as shown in example 4.

Hence to resolve this issue, space is automatically inserted at start and end of Latin character sequence. In addition, space is inserted between normal text and punctuation marks so each punctuation mark is considered as individual word.

After resolving space insertion issues, Dataset-1 and Dataset-2 is further processed to add tags which are useful to process lexical features. The complete URL is replaced with special word tag. To give tag to the web URL, the corpus is processed and web URLs are extracted by using regular expression. The complete web URL are replaced with special tag i.e. "httpaddr". In the same manner, email address is extracted using regular expression and replaced with "emailaddr" tag. Moreover, Latin cardinal number strings are extracted and replaced with a tag as "CD". These special word tags help to manipulate the lexical information for Urdu genre identification.

3.4.2. Stemming.

Urdu is morphological rich language i.e. a word may have more than one surface forms resulting into need of large amount of data containing all surface forms of the words so that machine learning classifier can better learn all forms. The requirement of this huge amount of dataset due to multiple surface form is resolved by applying Urdu stemmer. Before extracting words unigrams and bigrams from Dataset-1 and Dataset-2, Urdu stemming algorithm [10] is applied on both datasets. For better results of Urdu stemmer [10], all closed class words are extracted from both datasets using Urdu closed class list[5].

3.4.3. POS tagging.

Dataset-1 is manually annotated with POS tags [11] (details are given in Dataset Section). Dataset-2 is large corpus as compared to Dataset-1 and is not annotated with POS tagged. Manual annotation of this corpus with POS tags is troublesome task. Hence, an automatic Urdu POS tagger is used [11] to automatically tag the Dataset-2 so that word POS features can be extracted.

3.2. Features extraction

In any machine learning based NLP application, features play an important role to improve the accuracy of the application. In the same way, for the development of automatic genre identification, text document is processed to extract useful feature set which better distinguishes the genre of the respective document. For the development of Urdu genre identification system, two different types of features are extracted; (1) lexical features, and (2) structural features. To extract lexical features, words unigram and bigrams are computed along with their Term Frequency (TF) and Inverse Document Frequency(TF-IDF). These lexical features are separately extracted from both datasets.

To compute structural features, POS and sense information of word is used. Word along its POS feature set is computed from both datasets where as word with its sense information is only extracted from Dataset-1 as only this dataset is manually annotated with sense tags. Structural level information is used for experimentation of Dataset-1 to investigate the impact of word POS and sense on genre identification accuracy.

For dimensionality reduction low frequent terms are discarded. To see the impact of features set on genre identification, separate systems are developed for each feature, presented in Table 1. Details of number of features in each feature set are given in Section 5.

3.3. Classifiers

[5] http://cle.org.pk/software/ling_resources/UrduClosedClassWordsList.htm

The features are used to train the machine learning classification algorithms. The features are extracted from training data. These features along with label of the genre class are fed to the classifier in the training phase.

Table 1: Systems for the Urdu genre identification

System	Features
System 1	Word Unigram
System 2	Word Bigrams
System 3	Word/POS
System 4	Word/Sense

In the recognition, the features of the input document are computed and then based on the learning model, classifier predicts genre. In this study, state of the art classification algorithms including Support Vector Machines, Naive Bayes and C4.5 are used for Urdu genre identification system. Each classifier is separately trained on each of the features-based system (Table 1). The results are discussed in Section 5.

4. Dataset

Two different datasets are used for Urdu text genre identification. These are (1) CLE Urdu Digest 100K [8]named as Dataset-1, and (2) CLE Urdu Digest 1 Million named as Dataset-2.

Dataset-1 is a balanced corpus having 100K Urdu words which is collected from multiple genres of Urdu Digest corpus. This corpus is manually cleaned by linguist to resolve space insertion and space deletion issues. In addition, same corpus is manually annotated with POS tags by linguist [11]. The complete Urdu POS tagset along with guidelines for corpus annotation is defined. A total of 35 POS tags are defined in Urdu POS tagset.

Dataset-2 is 1 million Urdu words corpus, distributed into multiple domains. This corpus is automatically cleaned using the heuristic discussed in Corpus Cleaning Section. In addition, Dataset-2 is automatically annotated with Urdu POS tagset using POS tagger [11] which has 96.8% accuracy. The motivation behind using two different version of CLE Urdu digest is to investigate the effect of dataset size on the accuracy.

Eight genres i.e. culture, science, religion, press, health, sports, letters and interviews of both datasets are used for classification experiment. Details of training and testing documents of Dataset-1 and Dataset-2 against each genre are presented in Table 2.

5. Experiments and results

The lexical and structural features from Dataset-1 and Dataset-2 are extracted. Each feature set is labeled with different system and is evaluated separately to analyze its impact on genre identification accuracy. The number of features extracted from training data of Dataset-1 and Dataset-2 for each system are provided in Table 3. For Dataset-1, 228 training and 56 testing documents are used. While, Datset-2 includes 686 and 160 documents for training and testing, respectively (Table 2).

The features extracted from training data are used to train each classifier for Dataset-1 and Dataset-2 separately. Testing data is used to test the performance of each system trained by respective classifier.

Table 2: Datasets

Genre	Data Set 1		Data Set 2	
	Training Document	Testing Document	Training Document	Testing Document
Culture	34	8	120	30
Science	45	10	98	21
Religion	23	6	95	20
Press	23	6	94	24
Health	23	6	129	31
Sports	23	6	25	6
Letters	28	7	90	21
Interviews	30	7	35	7
Total	**229**	**56**	**686**	**160**

Table 3. Number of features in Dataset-1 and Dataset-2

System	Features	No. of features for Dataset-1	No. of features for Dataset-2
System 1	Word Unigram	156	1,665
System 2	Word Bigrams	316	4,798
System 3	Word/POS	1,037	6,548
System 4	Word/Sense	1,570	.....

The accuracy measures including Precision(P), Recall(R) and F-measure(F) are computed to evaluate accuracy results. Recall(R) is the number of correctly classified documents divided by the number of total documents. Precision(P) is the number of correct classifications divided by the number of classification

Table 4, Table 5, and

Table **6**, respectively.

Dataset-2 having more training examples gives better results as compared to the Dataset-1 for each system and each classifier. Moreover, results reveal that lexical features provide higher accuracy as

made whereas F-measure(F) is computed by using the following equation

F = 2 * (Precision*Recall) / (Precision + Recall).

Each classifier is trained separately on a system of each dataset, excluding System 4 which is trained and tested only for Dataset-1. The accuracy results of SVM, Naive Baye and C45 classifiers are presented in compared to the structural features. The System 2 i.e. word bigrams yields higher precision, recall and f-measure as compared to the other systems. It has been observed from the results that SVM outperforms the other classifiers irrespective of feature type.

Table 4. Systems results using SVM

System	Dataset-1			Dataset-2		
	P	R	F	P	R	F
System 1	0.50	0.50	0.48	0.68	0.68	0.67
System 2	0.38	0.33	0.35	**0.74**	**0.70**	**0.70**
System 3	0.63	0.62	0.62	0.72	0.68	0.68
System 4	0.53	0.35	0.38	...	...	...

Table 5. Systems results using Naive Bayes

System	Data Set 1			Data Set 2		
	P	R	F	P	R	F
System 1	0.45	0.37	0.37	0.68	0.67	0.66
System 2	0.37	0.39	0.37	**0.70**	**0.7 0**	**0.69**
System 3	0.59	0.58	0.58	0.67	0.65	0.63
System 4	0.34	0.35	0.32	...	...	...

Table 6. Systems results using C4.5

System	Dataset-1			Dataset-2		
	P	R	F	P	R	F
System 1	0.34	0.32	0.32	0.45	0.45	0.45
System 2	0.44	0.41	0.42	**0.47**	**0.45**	**0.46**
System 3	0.46	0.44	0.43	0.44	0.44	0.43
System 4	0.171	0.179	0.161	...	...	...

In addition, after doing detailed analysis, it has been observed that some texts in the genres are overlapping e.g. science and health which results in misclassification. The genre which is not overlapping e.g. sports has 100% genre identification accuracy.

6. Conclusion

In this paper, automatic Urdu text genre identification system has been presented by evaluating the impact of lexical and structural features along with different state of the art classifiers. From the results, it is observed that lexical features are most appropriate for identifying the genres of Urdu documents. In addition, results indicate that SVM has an advantage over other classifiers irrespective of feature type. The size of the training data also affects the accuracy. It is reinforced that significant amount of training data improves the document classification accuracy.

References

[1] B. Kessler, G. Bumberg and H. Schütze, "Automatic detection of text genre," in *in proc. European Chapter*

of the Association for Computational Linguistics (EACL), Madrid, Spain, 1997.

[2] E. Stamatatos, N. Fakotakis and G. Kokkinakis, "Text genre detection using common word frequencies," in *proc. Conference on Computational linguistics*, Stroudsburg, PA, USA, 2000.

[3] L. Y. and M. S. H., "Text genre classification with genre-revealing and subject-revealing features," in *proc. international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, 2002.

[4] A. R. Ali and M. Ijaz, "Urdu text classification," in *pro. International conference on Frontiers of Information Technology*, 2009.

[5] T. Zia, Q. Abbas and M. P. Akhtar, "Evaluation of Feature Selection Approaches for Urdu Text Categorization," *International Journal Intelligent Systems and Applications,* pp. 33-40, May 2015.

[6] C. S. Lim, K. J. Lee, G. C. Kim, K. Su, J. Tsujii, J. Lee and O. Kwong, "Automatic genre detection of web documents," in *Natural Language Processing-IJCNLP* , Berlin, Heidelberg, Springer Berlin Heidelberg, 2004, pp. 310-319.

[7] S. M. and Z. Q., "On sample size and classification accuracy: a performance comparison," in *proc. International conference on Biological and Medical Data Analysis* , Berlin, Heidelberg, 2005.

[8] S. Urooj, S. Hussain, F. Adeeba, F. Jabeen and R. Parveen, "CLE Urdu Digest Corpus," in *proc. Conference on Language and Technology*, Lahore, Pakistan, 2012.

[9] K. I. and S. E., "Webpage genre identification using variable-length character n-grams," in *proc. IEEE International Conference on Tools with Artificial Intelligence*, Washington, DC, USA, 2007.

[10] Q. Akram, A. Naseer and S. Hussain, "Assas-band, an Affix-Exception-List Based Urdu Stemmer," in *in the Proc. of the 7th Workshop on Asian Language Resources*, Suntec City, Singapore, 2009.

[11] T. Ahmed, S. Urooj, S. Hussain, A. Mustafa, R. Parveen, F. Adeeba, A. Hautli and M. Butt, "The CLE Urdu POS Tagset," in *poster presentation in Language Resources and Evaluation Conference(LERC 14)*, Reykjavik, Iceland., 2014.