6TH INTERNATIONAL CONFERENCE ON LANGUAGE & TECHNOLOGY
17TH - 18TH NOVEMBER, 2016

Urdu Text Genre Identification

Farah Adeeba, Sarmad Hussain, Qurat-ul-Ain Akram

مرکز تحقیقاتِ لسانیات
Center for Language Engineering
www.cle.org.pk

Outline

- Introduction
- Urdu Text Genre Identification
  - Corpus
  - Features
  - Classifiers
- Results
- Conclusion

Introduction

- Automated genre identification deals with prediction of genre of an unknown text, independent of its topic and style.
- Improve Performance of:
    - Parsing
    - word sense disambiguation
    - Information Retrieval

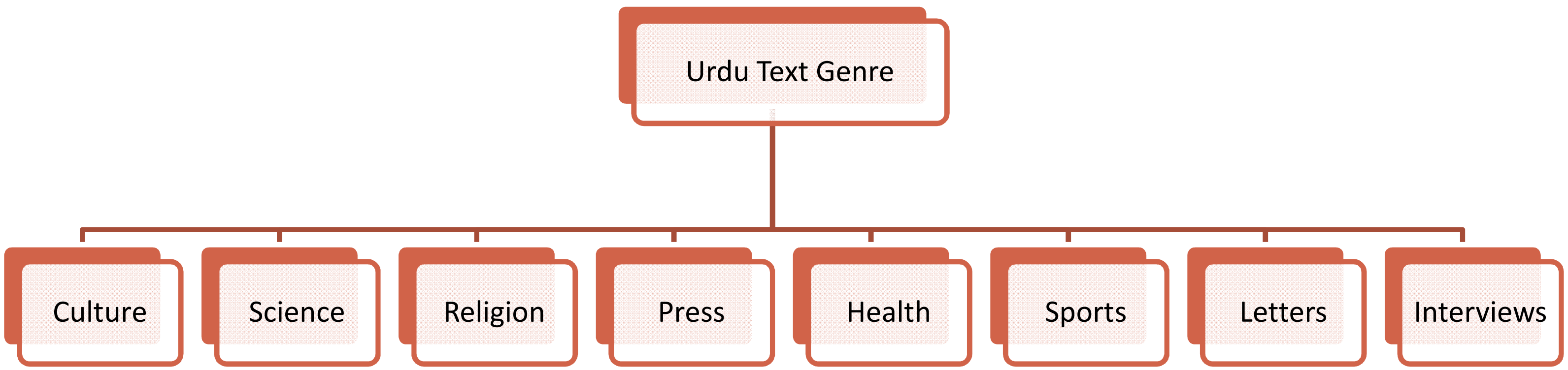
What do we need for Automatic Genre Identification

1. A genre taxonomy
2. A corpus of different genres
3. Measurable attributes (features) that can be extracted automatically
4. An automatic classifier, i.e. a computational model that does the classification for us

Urdu Text Genre Identification

Genre Taxanomy

- Eight genres



What do we need for Automatic Genre Identification

1. A genre taxonomy
2. A corpus of different genres
3. Measurable attributes (features) that can be extracted automatically
4. An automatic classifier, i.e. a computational model that does the classification for us

Corpus

- CLE Urdu Digest 100K
  - Manually Cleaned
  - Manually POS Tagged
  - Manually Sense Tagged
- CLE Urdu Digest 1M
  - Not cleaned
  - Un-Annotated

1. Informational (80%)	
a) Informal (20%)	Letters
	Interviews
b) Formal	
	Press
	Religion
	Sports
	Culture (travel, history)
	Entertainment
	Health
	Science (education, technology)
2. Imaginative (20%)	
	Short Stories
	Translation of foreign literature
	Novels
	Book reviews

Corpus

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**

Text Pre-processing (1)

1. Corpus cleaning
   - Space insertion deletion issues
   - Latin digits and Urdu text
   - The complete web URL are replaced with special tag "httpaddr“
   - Email address is extracted using regular expression and replaced with "emailaddr"
   - Latin cardinal number strings are extracted and replaced with a tag as "CD"

عرصہ۳۰ سال سے پی ٹی سی ٹیچر

دنیا بھر میں موجود 65کے لگ بھگ باقاعدہ اوپن یونیورسٹیوں کے ساتھ
کے قیام کے محض تین سال UKOUبرطانیہ میں بعد

HKEY_LOCAL_MACHINE\SOFTWARE\Microsost\Windows\CurrentVersion\Explorer\BitBucketاب

Text Pre-processing (2)

2. Stemming
   - Datasets are stemmed using Urdu Stemmer [1]
3. POS Tagging
   - Dataset 2 is automatically POS tagged using Urdu POS Tagger[2]

1: http://www.cle.org.pk/software/langproc/UrduStemmer.htm
2: http://cle.org.pk/clestore/postagger.htm

What do we need for Automatic Genre Identification

1. A genre taxonomy
2. A corpus of different genres
3. Measurable attributes (features) that can be extracted automatically
4. An automatic classifier, i.e. a computational model that does the classification for us

Features

- Impact of different features
  - Lexical
  - Structural
- Features are computed along with their Term Frequency (TF) and Inverse Document Frequency(TF-IDF)
- Each feature set is labeled with different system

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

Features

- For dimensionality reduction low frequent terms are discarded

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	.....

What do we need for Automatic Genre Identification

1. A genre taxonomy
2. A corpus of different genres
3. Measurable attributes (features) that can be extracted automatically
4. An automatic classifier, i.e. a computational model that does the classification for us

Classifiers

- Features are computed and then based on the learning model, classifier predicts genre of a document
  - Support Vector Machines(SVM)
  - Naive Bayes
  - C4.5

System Evaluation

- Accuracy is measured for
  - Each feature set
  - Classifier
    - SVM
    - Naïve Bayes
    - Decision Tree
- 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 made
- F-measure(F) is computed by using the following equation

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

System 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	...	...	...

System Results using Naïve 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	...	...	...

System 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	...	...	...

Conclusion

- Lexical features provide higher accuracy as compared to the structural features
- SVM outperforms the other classifiers irrespective of feature type
- Dataset-2 having more training examples gives better results as compared to the Dataset-1 for each system and each classifier

Acknowledgments

مرکز تحقیقات لسانیات

Center for Language Engineering

www.cle.org.pk

Thank you for your time and attention