Automatic Urdu Text Genre Identification

Presenter: Farah Adeeba

25th June, 2014 | Center for Language Engineering (CLE)

Outline

- What is Genre
- How to Define Genre
- Urdu Text Genre Identification

Genre Classification

- Automated genre classification is concerned with predicting the genre of an unknown text correctly, independent of its topic, style or any other characteristic.

Applications

- Accuracy Improvement for
  - Parsing
  - POS Tagging
  - Word-sense Disambiguation
  - Information Retrieval

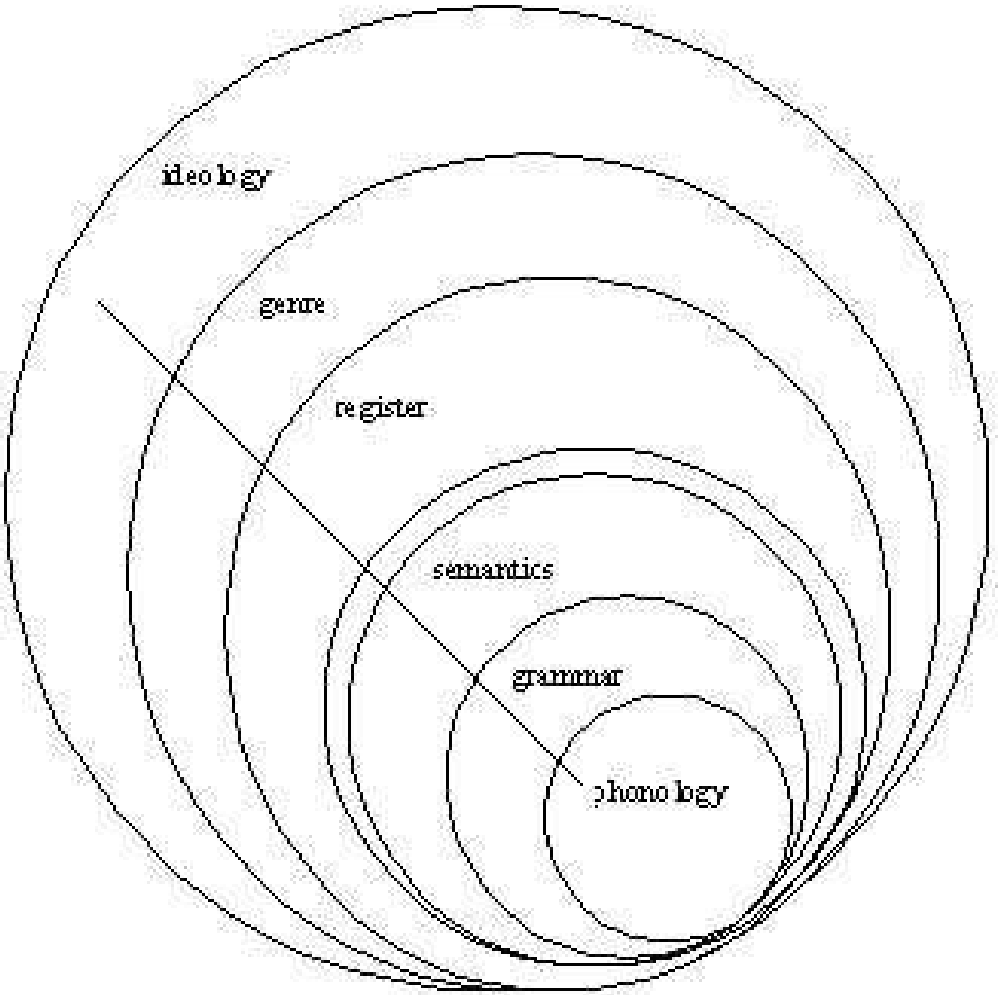
Concepts

- Genre
- Register
- Text Type
- Domain
- Style

Genre, Register

- Crystal (1991, p. 295) defines register as "a variety of language defined according to its use in social situations, e.g. a register of scientific, religious, formal English.

Genre, Register (Martin,1993)



Genre & Text Type(Biber ,1988)

- Genre – External, non-linguistic, "traditional" Criteria
  - intended audience
  - purpose
  - activity type
  - Assigned on the basis of use rather than on the basis of form
- Text type- internal, linguistic characteristics of texts themselves

Genre & Style

- Style is defined by authorship
- This is not necessarily restricted to single persons, but can be extended to newspapers, companies or other institutions with binding style guides
- Geographic regions and periods in time

Genre & Topic

- The topic of a text is what it is about.
- Examples: countries, sports, financial mattes or crocodiles, regardless of whether it is a song or an FAQ section of a website.

Genre , Style, Topic Example

- An example would be a letter about a trip to Inverness written by Robert Burns. The trip would be considered the topic of the document. The letter is the genre and the text is written in the personal style of Robert Burns.

Genre Label in Corpus

- problem with genre labels is that they can have so many different levels of generality
- A second problem is that different "genres" can be based on so many different criteria (domain, topic, participants, setting, etc.)

SUPERORDINATE	Mammal	Literature ["SUPER-GENRE"?]	Advertising ["SUPER-GENRE"]
BASIC-LEVEL	Dog/Cat	Novel, Poem, Drama [GENRE]	Advertisement [GENRE]
SUBORDINATE [PROTOTYPE]	Cocker spaniel / Siamese	Western, Romance, Adventure [SUB-GENRE]	Print ad, Radio ad, TV ad, T-shirt ad [SUB-GENRE]

Genre Classification

- Automated genre classification is concerned with predicting the genre of an unknown text correctly, independent of its topic, style or any other characteristic.
- Main Tasks
  - Feature Set
  - Classification Algorithm

Feature Set

- Sub lexical level Cues
- Lexical Cues
- Structural Cues
- Hybrid Cues

Classification Algorithms

- K-nearest neighbour
- Decision trees
- Naive Bayes
- multinomial Naive Bayes
- Support vector machines
- Self Organized Maps
- Neural Network

Urdu Text Genre Identification

- Experiments
  - Lexical level Experiments
  - Deeper Information
  - Sub lexical level Experiments
  - Hybrid
- Data Set
  - CLE Urdu Digest 100K
    - POS Tagged
    - Sense Tagged
  - CLE Urdu Digest 1M

Data Set

Genre	Data Set 1 (100K CLE Urdu Digest)		Data Set 2 (1M CLE Urdu Digest)	
	Training Document	Testing Document	Training Data	Testing Data
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**

Automatic Urdu Text Genre Identification

Feature Set Details

System #	Features	Data Set 1 Count	Data Set 2 Count	Cut off
System 1	Word Unigram	156	1665	50
System 2	Word Bigrams	316	4798	20
System 3	POS Unigram	36	36	1
System 4	POS Bigram	175	510	50
System 5	POS Trigram	317	2326	50
System 6	Sense	800	.....	1
System 7	Ligature Bigram	2375	13488	10
System 8	Ligature Trigram	2281	12531	10,50
System 9	Ligature 4-grams	642	10500	20
System 10	Word/POS	1037	6548	10
System 11	Word/Sense	1570	.....	1
System 12	Ligature Trigram + POS Bigram	527	13040	50,50
System 13	Ligature Trigram + POS Trigram	669	14872	50,50
System 14	Ligature 4-gram + POS Unigrams	676	8868	20,1

Classifier

- SVM
- Naive Bayes
- C4.5

Text Pre-processing

- Corpus Cleaning
  - Space inserted between Latin numbers/Latin Characters and text

1. عرصہ ۳۰ سال سے پی ٹی سی ٹیچر
2. دنیا بھر میں موجود 65 کے لگ بھگ باقاعدہ اوپن یونیورسٹیوں کے ساتھ
3. برطانیہ میں UKOU کے قیام کے محض تین سال بعد
4. HKEY_LOCAL_MACHINE\SOFTWARE\Microsost\Windows\CurrentVersi on\Explorer\BitBucket کا
5. کیاآپ ہمیشہ ایک Good Dreamer رہے ہیں۔

  - URL is termed as Web URL and replaced with special tag of "httpaddr
  - Email address are also replaced with tag of "emailaddr".
  - Latin cardinal number strings are extracted and replaced with special symbol of "CD"

Text Pre-processing (2)

- Pos Tagging
  - For POS n-grams and Word/POS feature Data Set 2 is POS tagged using automatic Urdu POS tagger
- Ligature Splitting

Experiments

- Systems Experiments with Two Genres
- System Results with Four Genres
- System Results with Eight Genres

Systems Experiments with Two Genres

- Health and Religion

System	Data Set 1			Data Set 2		
	Precision	Recall	F1	Precision	Recall	F1
System 1	0.929	0.917	0.916	0.905	0.902	0.9
System 2	0.929	0.917	0.916	0.941	0.941	0.941
System 3	0.875	0.833	0.829	0.902	0.902	0.901
System 4	0.875	0.833	0.829	0.963	0.961	0.96
System 5	**1**	**1**	**1**	**1**	**1**	**1**
System 6	0.875	0.833	0.829	---	---	---
System 7	0.929	0.917	0.916	**1**	**1**	**1**
System 8	**1**	**1**	**1**	0.981	0.98	0.98
System 9	0.929	0.917	0.916	0.946	0.941	0.94
System 10	0.833	0.833	0.833	0.981	0.98	0.98
System 11	0.78	0.846	0.811	---	---	---
System 12	0.929	0.917	0.916	0.963	0.961	0.96
System 13	0.929	0.917	0.916	**1**	**1**	**1**
System 14	0.929	0.917	0.916	0.981	0.98	0.98

System Results with Four Genres

- Health, Culture, Religion, Letters

System	Data Set 1			Data Set 2		
	Precision	Recall	F1	Precision	Recall	F1
System 1	0.646	0.556	0.53	0.755	0.696	0.672
System 2	0.599	0.593	0.594	0.79	0.755	0.752
System 3	0.313	0.333	0.32	0.376	0.529	0.433
System 4	0.673	0.667	0.661	0.74	0.735	0.737
System 5	0.487	0.481	0.478	0.747	0.735	0.734
System 6	0.543	0.556	0.542	---	---	---
System 7	0.77	0.778	0.772	0.818	0.804	0.793
System 8	**0.891**	**0.899**	**0.888**	**0.826**	**0.814**	**0.803**
System 9	0.358	0.481	0.393	0.49	0.598	0.515
System 10	0.911	0.63	0.742	0.79	0.784	0.775
System 11	0.519	0.5	0.519	---	---	---
System 12	0.7	0.667	0.657	0.384	0.529	0.438
System 13	0.6	0.593	0.59	0.833	0.824	0.818
System 14	0.774	0.778	0.768	0.827	0.804	0.796

Automatic Urdu Text Genre Identification

System Results with Eight Genres

System	Data Set 1			Data Set 2		
	Precision	Recall	F1	Precision	Recall	F1
System 1	0.377	0.357	0.312	0.559	0.544	0.52
System 2	0.374	0.375	0.327	0.735	0.663	0.665
System 3	0.287	0.286	0.282	0.521	0.513	0.51
System 4	0.434	0.393	0.402	0.571	0.563	0.563
System 5	0.359	0.339	0.331	0.638	0.613	0.614
System 6	0.569	0.5	0.492	---	---	---
System 7	0.758	0.554	0.564	0.763	0.731	0.729
System 8	0.608	0.571	0.565	**0.774**	**0.763**	**0.762**
System 9	0.486	0.429	0.422	0.736	0.725	0.725
System 10	**0.692**	**0.589**	**0.607**	0.721	0.688	0.686
System 11	0.599	0.536	0.535	---	---	---
System 12	0.561	0.518	0.52	0.549	0.556	0.547
System 13	0.515	0.429	0.427	0.754	0.731	0.73
System 14	0.673	0.554	0.536	0.746	0.725	0.723

System Results with Different Classification Algorithm

Classification Algorithm	System	Data Set 1			Data Set 2		
		Precision	Recall	F1	Precision	Recall	F1
Naive Bayes	System8	0.509	0.464	0.459	0.648	0.606	0.604
Decision Tree	System 8	0.397	0.393	0.381	0.417	0.413	0.408
SVM	**System 8**	**0.608**	**0.571**	**0.565**	**0.774**	**0.763**	**0.762**

CONCLUSION

- For accurate genre identification impact of structural, lexical, sub lexical and hybrid features are being used.
- Moreover the impact of number of genres and data set size is also investigated.
- Impact of classification algorithm is also explored.
- The features extracted from sub lexical level information are most appropriate for identifying the genres of documents.

References(1)

- KESSLER B., NUMBERG G., SHÜTZE H. (1997),"Automatic Detection of Text Genre", Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics.
- STAMATATOS E., FAKOTAKIS N., KOKKINAKIS G. (2000),"Text Genre Detection Using Common Word Frequencies", Proceedings of the 18th International Conference on Computational Linguistics (COLING2000).
- Vasiliki Simaki, Sofia Stamou, Nikos Kirtsis: Empirical Text Mining for Genre Detection. WEBIST 2012:

References(2)

- Peter Wastholm , Annette Kusma , Beáta B. Megyesi (2005) ,Using linguistic data for genre classification In Proceedings of the Swedish Artificial Intelligence and Learning Systems Event SAIS-SSLS, Mälardalen University, Schweden
- Sung Hyon Myaeng (2002). "Text genre classification with genre-revealing and subject-revealing features" in the Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval P 145-150