20th International CODATA Conference:
22-25 October 2006, Beijing, China

Automated Genre Classification for Ingest and Appraisal Metadata

Yunhyong Kim and Seamus Ross
Digital Curation Centre
&
Humanities Advanced Technology and Information Institute
University of Glasgow, Glasgow, UK
{y.kim, s.ross}@hatii.arts.gla.ac.uk

Funded by:
JISC
e-Science

End objectives:

- To enable the automatic extraction of descriptive information for digital objects.

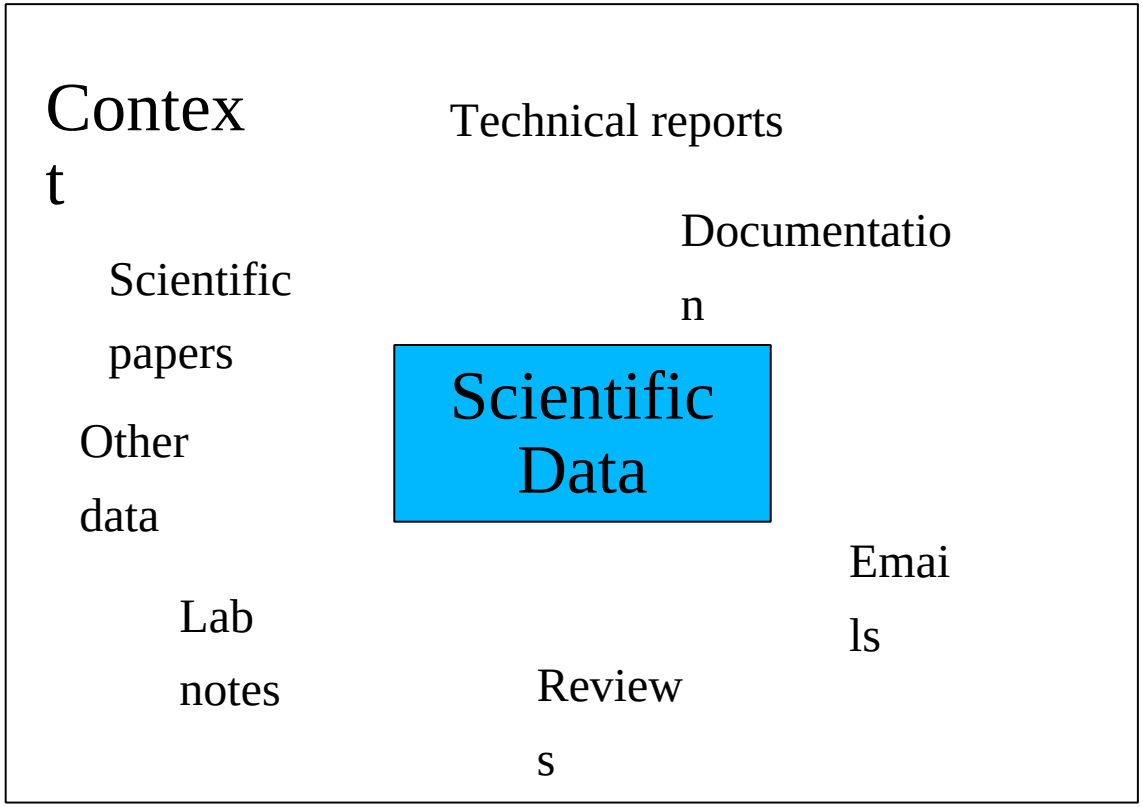
- To enable the automatic identification, selection and management of digital material.

- To create a network of relationships and contexts for information produced independently.

In this presentation we discuss
automatic genre classification

that is
the automatic recognition of document types such as scientific papers, tables, theses etc.

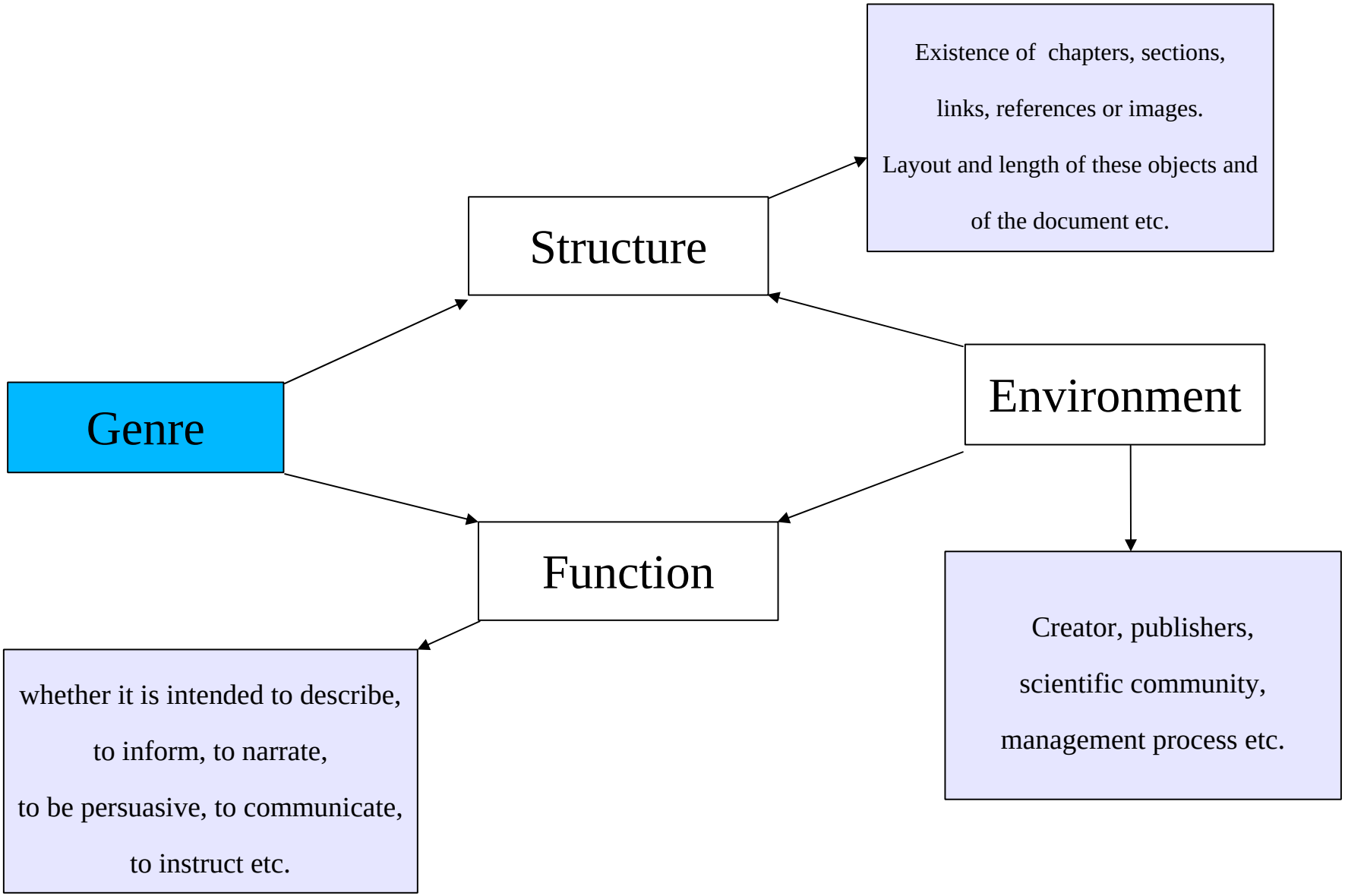
Understanding data: why genre classification?



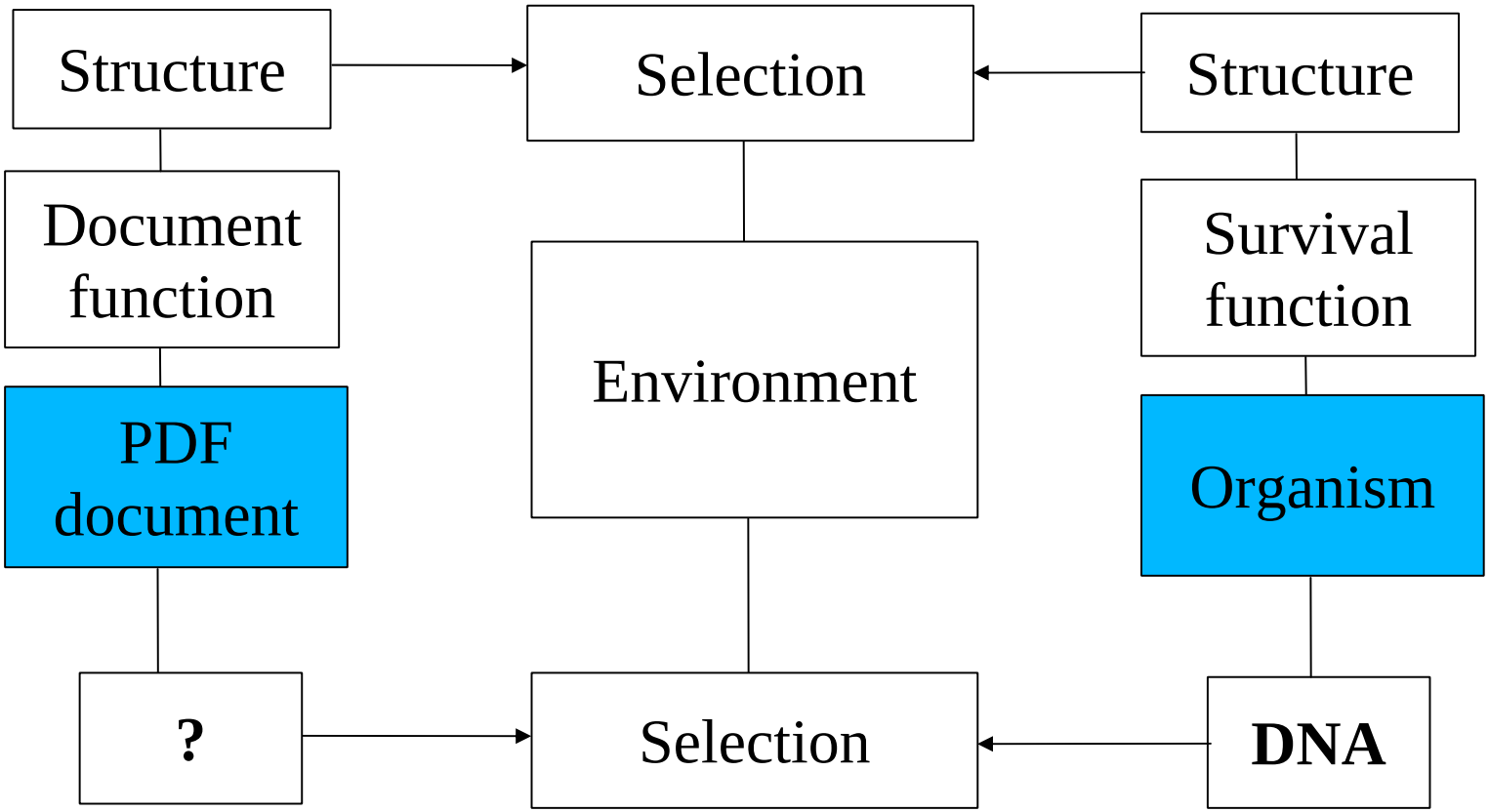
What is genre?

Sample research in genre:

- Biber, D.: Dimensions of Register Variation:a Cross-Linguistic Comparison. Cambridge University Press (1995).
- Karlgren, J. and Cutting, D.: Recognizing Text Genres with Simple Metric using Discriminant Analysis. Proc. 15th conf. Comp. Ling. {\bfseries Vol 2} (1994) 1071--1075.
- Kessler, B., Nunberg, G., Schuetze, H.: Automatic Detection of Text Genre. Proc. 35th Ann. Meeting ACL (1997) 32—38.
- Rauber, A. and Müller-Kögler, A.: Integrating Automatic Genre Analysis into Digital Libraries In: Fox, E.A., and Borgman, C.L. (eds.), Proceedings of the ACM/IEEE Joint Conference on Digital Libraries 2001 (JCDL01), June 24 - 28 2001, Roanoke, VA, pp.1-10, ACM, 2001.
- Bagdanov, A. D., Worring, M.: Fine-Grained Document Genre Classification Using First Order Random Graphs. Proceedings of International Conference on Document Analysis and Recognition 2001 (2001) 79.
- Boese, E. S.: Stereotyping the web: genre classification of web documents. Master's thesis, Colorado State University (2005).
- Finn, A. and Kushmerick, N.: Learning to Classify Documents According to Genre. Journal of American Society for Information Science and Technology, 57 (11), 1506-1518, 2006


Existence of chapters, sections,
links, references or images.
Layout and length of these objects and
of the document etc.
Structure
Genre
Environment
Function
Creator, publishers,
scientific community,
management process etc.
whether it is intended to describe,
to inform, to narrate,
to be persuasive, to communicate,
to instruct etc.

Documents as a dynamic entities



Properties that characterise documents

- **Image (white space analysis)**
- **Style (length, average length of words, word frequency analysis, number of font changes, difference between largest and smallest font size)**
- **Language model (N-gram model)**
- **Semantics (proportion of objective nouns, argumentation structure etc.)**
- **Context (who created it for whom and where is it from)**

Experiments

- Clustering documents
- Binary predictions in a pool of nineteen genres
  - Retrieving Periodicals
  - Retrieving Thesis
  - Retrieving Scientific Articles
  - Retrieving Business Reports
  - Retrieving Forms
- Classification of five genres

Results I (Cluster)

Group	Genre	Visual	Stylistic	Combinec
Book	Academic Book	100	60	100
	Fiction	92.8	83.3	75
	Other Book	70.6	82.4	70.6
Article	Scientific Article	76	92	64
	Other Research	94.7	73.7	84.2
	Magazine article	61.5	84.6	61.5
Serial	Periodicals (Newspaper, M	100	62.5	87.5
	Newsletter	54.2	83.3	58.3
Treatise	Thesis	100	90	90
	Business Report	81.8	90.9	81.8
	Technical Report	88.9	72.2	83.3
Information structure	List	71.4	85.7	71.4
	Form	61.5	69.2	53.8
Evidential document	Minutes	100	76.9	100
Other	Guideline	95	78.6	90
	Job Description	73.3	50	73.3
	Product Description	62.5	66.7	56.3
	Fact Sheet	72.4	85.7	64.3
	Slides	61.5	91.7	61.5

Results II (Periodicals)

image classifier (acc. 88.6%) →

Genre	Precision (%)	Recall (%)
Period	29.8	87.5
Other	99.2	88.7

style (acc. 88.52%) →

Genre	Precision (%)	Recall (%)
Period	14.8	25
Other	92	93.8

language (acc. 94.79%) →

Genre	Precision (%)	Recall (%)
Period	0	0
Other	96.9	100

Results III (Scientific Article, Thesis)

Scientific Article

Classifier	Precision (%)	Recall (%)
image	21	80
Style	50	76
Language	100	15

Thesis

Classifier	Precision (%)	Recall (%)
image	13.6	80
Style	7	60
Language	40	17.4

Results IV (Business Report, Forms)

Business Report

Classifier	Precision (%)	Recall (%)
image	5.6	63.6
Style	9.1	72.7

Forms

Classifier	Precision (%)	Recall (%)
image	7.2	38.5
Style	10.6	76.9

Classification: five genres (language model)

Genre	Precision (%)	Recall (%)
Academic Book	42.9	60
Business Report	50	81.8
Fictional Book	100	100
Minutes	86.7	92.9
Thesis	75	60

Conclusions

- different genres have different feature strengths.
- retrieval of selected genres dependent on strong feature types may perform better than global analysis of all features to classify a large number of genres.
- binary decisions divide document space into groups less likely and more likely to contain a given genre type .

Future Work

- Improvement of the classifiers
  - Extended image classifier
  - Extended Language model classifier
  - Augmented stylistic classifier
- More classifiers
  - Semantic classifier
  - Contextual classifier
- Human Labelling experiments
  - Document retrieval exercise
  - Re-labelling exercise

Errors for Periodicals, Thesis, Scientific Article: Confusion Matrix

Group	Genre	Period	Sci. Article	Thesis
Book	Academic Book	0	2	3
	Fiction	1	4	9
	Other Book	2	8	7
Article	Scientific Article	0	24	1
	Other Research	0	14	5
	Magazine article	3	9	1
Serial	Periodicals (News, Magazine)	15	1	0
	Newsletter	7	17	0
Treatise	Thesis	0	1	9
	Business Report	1	5	5
	Technical Report	1	10	7
Information structure	List	1	10	3
	Form	1	12	0
Evidential document	Minutes	0	11	2
Other	Guideline	1	17	3
	Job Description	2	13	0
	Product Description	6	9	1
	Fact Sheet	3	11	0
	Slides	4	6	3