Towards a Reference Corpus of Web Genres for the Evaluation of Genre Identification Systems

Georg Rehm[1], Marina Santini[2], Alexander Mehler[3]
Pavel Braslavski[4], Rüdiger Gleim[3], Andrea Stubbe[5],
Svetlana Symonenko[6], Mirko Tavosanis[7], Vedrana Vidulin[8]

Tübingen University, Germany[1]
SFB 441: Linguistic Data Structures

Stockholm University, Sweden[2]
Department of Linguistics

Bielefeld University, Germany[3]
Computational Linguistics Department

Institute of Engineering Science, RAS[4]
Ekaterinburg, Russia

Centrum für Informations- und Sprach-[5]
verarbeitung, University of Munich, Germany

Nitol, LLC[6]
Moscow, Russia

Università di Pisa, Italy[7]
Dipartimento di Studi italianistici

Jožef Stefan Institute[8]
Ljubljana, Slovenia

1 Introduction

The field of automatic web genre identification is still in its infancy as an established research area.[1] Current approaches are primarily characterised by a certain heterogeneity. They usually work on a collection of web documents compiled by the researchers themselves. A category set is constructed and applied, so that all documents are tagged with one or more genres contained in the category set. Finally, genre categorisation experiments are carried out.

Due to the success of widely used collections such as Reuters-21578 or the Enron mail corpus it is obvious that there are severe problems inherent to this approach and that a reference corpus and a shared category set are needed. Currently there is no such genre benchmark corpus against which to measure results. Only a common dataset can enable researchers to compare and to evaluate their systems and to discuss interoperability issues. Moreover, a reference corpus could prevent people to invest large amounts of time and money to come up with (isolated) solutions to the complex tasks of building a corpus and a suitable category set.

[1]This joint paper is the result of a discussion the authors had at the workshop "Towards Genre-Enabled Search Engines: The Impact of NLP", held, in conjunction with RANLP 2007, on September 30, 2007 (Rehm and Santini, 2007).

Section 2 gives an overview of web genre identification. Section 3 looks at the most important prerequisites for a reference corpus. These are collections of web documents, shared sets of categories, and tools for the annotation of the collection with specific categories so that a gold standard benchmark can be built.

2 Automatic Web Genre Identification: A Short Overview

While keywords express the *topic* of a text, *genre* expresses its type. Keywords can be ambiguous, even misleading – this is why keyword-based searches frequently return irrelevant results. The concept of genre helps to distinguish different types of texts, e. g., *academic paper*, *manual*, *editorial*, and *blog*. These genres show characteristics that are – mostly – topic-independent. In an IR system, genre and topic should be, ideally, used together to increase its accuracy, so that queries such as "*academic papers* about *global warming*" could filter out texts of other genres.

Preliminary results in genre-enabled IR were reported by Karlgren et al. (1998). Xu et al. (2007), Yeung et al. (2007) and nearly all other approaches since the seminal papers by Karlgren and Cutting (1994) and Kessler et al. (1997) suffer from the same shortcomings: genre category sets are built according to subjective criteria for corpus composition, genre annotation, and genre granularity. The field is characterised by small-scale, self-contained, and corpus-dependent experiments.

3 Towards a Reference Corpus of Web Genres

There are three prerequisites for a reference corpus: we need a shared category set that the majority of researchers in this field agree upon (section 3.1), a document collection (section 3.2), and an annotation and processing tool (section 3.3).

3.1 Prerequisite I: A Shared Category Set

Before we can construct a reference corpus of web genres, the majority of the researchers have to agree upon one or more shared category sets, because different category sets make comparisons and evaluations impossible (see table 1). Most category sets do not contain proper genres or web genres but categories that are topical or functional in nature (again, see table 1); in other words, most category sets are not based on the established terms, concepts, and distinctions used in textlinguistics and genre theory, but they contain categories that have been created in an ad hoc fashion (e. g., *discussion*, *simple tables/lists*, *person*, *resources*, *nothing*, *pornographic*, *childrens'*, *content delivery*, *informative*). Another problem is that most approaches assume that web genres can be categorised on the "page" or "document" level. In addition to the assignment of genre categories to single HTML documents, genres also work on an intra-document level because a single document can contain instances of multiple genres (e. g., *contact information*, *list of publications*, *C. V.*, see Rehm, 2002, 2007, Mehler et al., 2007). Furthermore, we need a second category set for the web genre modules that occur on

Meyer zu Eissen and Stein (2004)	Help; Article; Discussion; Shop; Portrayal (non-private); Portrayal (private); Link Collection; Download
Lim et al. (2005)	Personal homepages; Public homepages; Commercial homepages; Bulletin collections; Link collections; Image collections; Simple tables/lists; Input pages; Journalistic materials; Research reports; Official materials; Informative materials; FAQs; Discussions; Product specifications; Others
Stubbe et al. (2007)	Journalism (Commentary; Review; Portrait; Marginal Note; Interview; News; Feature Story; Reportage); Literature (Poem; Prose; Drama); Information (Science Report; Explanation; Recipe; FAQ; Lexicon; Word List; Bilingual Dictionary; Presentation; Statistics; Code); Documentation (Law; Official Report; Protocol); Directory (Person; Catalog; Resources; Timeline); Communcation (Mail/Talk; Forum; Blog; Form); Nothing
Vidulin et al. (2007)	Pornographic; Blog; Childrens'; Commercial/Promotional; Community; Content Delivery; Entertainment; Error Message; FAQ; Gateway; Index; Informative; Journalistic; Official; Personal; Poetry; Prose Fiction; Scientific; Shopping; User Input
Braslavski (2007)	Official, academic, journalistic, literary, and everyday communication style

Table 1: Several recent category sets

the intra-document level, and even a third category set, because web genres can be instantiated on the level of whole websites Mehler and Gleim (2006). Ideally, conventionalised connections between these levels should also be represented within the category sets (for example, that *conference website* contains a *call for papers*).

3.2 Prerequisite II: A Reference Collection of Documents

We plan to build the corpus in two stages: first, we will apply the category sets to existing collections as a proof of concept, then we will use a web crawler to gather a more recent set of documents. Among the collections that we plan to process initially are English (Santini, 2007), German (Mehler et al., 2007), Russian (Braslavski, 2007), and Italian corpora (Tavosanis, 2007). For the final version we plan the inclusion of functions for a monitor corpus so that we can observe and take into account how documents change over time.

3.3 Prerequisite III: Tools

A reference set of genre categories requires tools that operate on various levels of web documents. HyGraph is a system for the construction, storage, management, and retrieval of large corpora of web documents (Gleim et al., 2007). It includes a crawler, a website segmenter (for mapping web pages onto XML-representations of their DOM and link structure), and a viewer as a graphical interface for rapid website skimming (Mehler and Gleim, 2006). HyGraph also contains a tool which allows the annotation of DOM subtrees as well as pages and whole websites as units of analysis. Using this dynamic categorisation tool, annotators can build and maintain their tagsets during the process of annotation – a very important feature

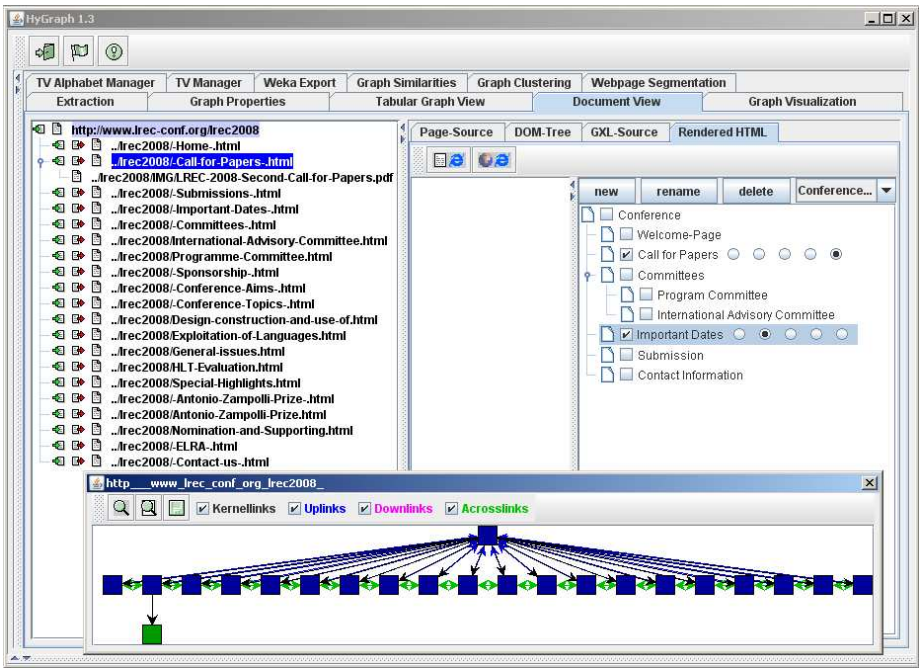

Figure 1: HyGraph's categorisation function

for our overall goal (see figure 1). Further, HyGraph integrates machine learning methods (including structure learning based on hyperlinks) and also maps to standard ML formats (including Weka, SVMLight, LibSVM). We plan to build the corpus using HyGraph.

4 Scope of this Contribution

Should this submission be accepted for LREC 2008, our final paper will include a motivation for the need of a reference corpus of web genres. We will emphasise the aspect of shared category sets and will analyse some sets used in research papers to show their commonalities and differences. Based on these results, previous work (e. g., Rehm, 2007), related approaches (Rosso, 2005) as well as best-practices in textlinguistics (Heinemann and Heinemann, 2002), we will suggest at least three shared category sets, agreed upon by all authors, that can be used for the construction of a reference corpus. Furthermore, we will provide an initial version of a web genre corpus that integrates several collections used for this purpose and that will be annotated using the shared category sets. The annotation will be carried out by as diverse a group of web users as possible so that real users (in contrast to the researchers themselves) construct this part of the resource; inter-coder reliability should be taken into account. We will show how the existing collections are processed by HyGraph to build a multilingual and multi-purpose resource for both web genre identification experiments and corpuslinguistic analyses.

References

Braslavski, Pavel (2007): "Combining Relevance and Genre-Related Rankings". In: Rehm and Santini (2007), pp. 1–4.

Gleim, R.; Mehler, A. and Eikmeyer, H.-J. (2007): "Representing and Maintaining Large Corpora". In: *Proc. of Corpus Ling. 2007*. Birmingham, UK.

Heinemann, Margot and Heinemann, Wolfgang (2002): *Grundlagen der Textlinguistik*. Tübingen: Niemeyer.

Karlgren, Jussi; Bretan, Ivan; Dewe, Johan; Hallberg, Anders and Wolkert, Niklas (1998): "Iterative Information Retrieval Using Fast Clustering and Usage-Specific Genres". In: *Proc. of the 8th DELOS Workshop on User Interfaces in Digital Libraries*. Stockholm, pp. 85–92.

Karlgren, Jussi and Cutting, Douglass (1994): "Recognizing Text Genres with Simple Metrics Using Discriminant Analysis". In: *Proc. of COLING 94*. Kyoto, pp. 1071–1075.

Kessler, Brett; Nunberg, Geoffrey and Schütze, Hinrich (1997): "Automatic Detection of Text Genre". In: *Proc. of the 35th Annual Meeting of the ACL*. San Francisco: Morgan Kaufmann, pp. 32–38.

Lim, Chul Su; Lee, Kong Joo and Kim, Gil Chang (2005): "Multiple Sets of Features for Automatic Genre Classification of Web Documents". *Information Processing and Management* 41 (5): pp. 1263–1276.

Mehler, A. and Gleim, R. (2006): "The Net for the Graphs – Towards Webgenre Representation for Corpus Linguistic Studies". In: *WaCky! Working Papers on the Web as Corpus*, edited by Baroni, M. and Bernardini, S., Bologna: Gedit, pp. 191–224.

Mehler, Alexander; Gleim, Rüdiger and Wegner, Armin (2007): "Structural Uncertainty of Hypertext Types". In: Rehm and Santini (2007), pp. 13–20.

Meyer zu Eissen, Sven and Stein, Benno (2004): "Genre Classification of Web Pages". In: *Proc. of the 27th German Conf. on Artificial Intelligence (KI-2004)*. Ulm.

Rehm, Georg (2002): "Towards Automatic Web Genre Identification". In: *Proc. of the 35th Hawaii Int. Conf. on System Sciences (HICSS-35)*. Big Island, Hawaii.

Rehm, Georg (2007): *Hypertextsorten: Definition – Struktur – Klassifikation*. Norderstedt: Books on Demand. (PhD thesis in Applied and Comp. Ling., Giessen University, 2005).

Rehm, Georg and Santini, Marina (editors) (2007): *Proc. of the Int. Workshop Towards Genre-Enabled Search Engines: The Impact of NLP*, Borovets, Bulgaria.

Rosso, Mark A. (2005): *Using Genre to Improve Web Search*. Ph. D. thesis, School of Information and Library Science, University of North Carolina at Chapel Hill.

Santini, Marina (2007): *Automatic Identification of Genre in Web Pages*. Ph.D. thesis, University of Brighton, United Kingdom.

Stubbe, Andrea; Ringlstetter, Christoph and Goebel, Randy (2007): "Elements of a Learning Interface for Genre Qualified Search". In: Rehm and Santini (2007), pp. 21–28.

Tavosanis, Mirko (2007): "Juvenile Netspeak and Subgenre Classification Issues in Italian Blogs". In: Rehm and Santini (2007), pp. 37–43.

Vidulin, Vedrana; Luštrek, Mitja and Gams, Matjaž (2007): "Using Genres to Improve Search Engines". In: Rehm and Santini (2007), pp. 45–51.

Xu, J.; Cao, Y.; Li, H.; Craswell, N. and Huang, Y. (2007): "Searching Documents Based on Relevance and Type". In: *ECIR 2007*. Rome.

Yeung, P.; Büttcher, S.; Clarke, C. and Kolla, M. (2007): "A Bayesian Approach for Learning Document Type Relevance". In: *ECIR 2007*. Rome.