Marina Santini, Georg Rehm, Serge Sharoff, Alexander Mehler

Editorial

In recent years, a multitude of most interesting research has been carried out in linguistics, psycholinguistics, computational linguistics and information retrieval with regard to the topic of web genres. Despite the increasing interest in this novel and innovative field within different communities, there is still a significant lack in literature, especially concerning edited collections and journal issues that provide an overview of recent research. The aim of this special issue of the *Journal for Language Technology and Computational Linguistics* is to contribute to filling this gap. More specifically, this issue is dedicated to automatic genre identification.[1]

Genres are categories that subsume texts which have multiple features in common, most importantly, a shared communicative purpose. Genres form and evolve within specific discourse communities, are instantiated as well as enforced and most often also given a name by members of their respective discourse communities. An important characteristic is that their users are able to recognise certain genres, such as, for example, an invoice, a business letter, a shopping list, or a menu very quickly based on genre-specific properties (such as, for example, a conventionalized text structure, certain key words, a specific layout and several other properties). Recognizing that a text belongs to a certain genre helps in the assessment of its importance and significance as well as the communicative goals its original author associated with the text. In other words, genres are a very important means for effective communication.

Karlgren and Cutting (1994) and Kessler et al. (1997) were among the first to suggest that genres might be useful to enhance information retrieval systems. The automatic identification of text genres could be used to generate additional metadata about text collections that, in turn, could enable users to search for texts based on their genre properties (with hypothetical queries such as "find instances of the genres *invoice* or *business letter* that contain *company car*"). Later on, when it became evident that computational linguistics had a new, exciting and most challenging target in the form of the World Wide Web, this idea was extended to web documents (see, for example, Crowston and Williams, 1997; Haas and Grams, 2000; Roussinov et al., 2001; Rehm, 2002). If we could identify automatically the genres of web documents we could extend the procedure and functionality sketched above to the whole web, giving users new and innovative means of locating relevant content online.

There is a plethora of open questions with regard to genre and web genre research. People who approach the field from a linguistic and text linguistic perspective try to find ways of describing and representing genres and web genres with the help of knowledge representation formalisms such as ontologies. Another interesting question concerns the evolution of genres and web genres, how they are formed in dynamic social communication system and discourse communities and how certain optional properties

[1] It is interesting to note that most articles in this issue are authored or co-authored by students who are still engaged in ongoing research or who have recently completed their research projects.

or features of genre instances slowly change their status into obligatory components. Researchers who work in psycholinguistics are, for example, concerned with the problem of genre recognition: how do we identify instances of certain genres, what kind of cues or inherent properties of a document do we employ in order to categorize a specific text into a certain genre? Is it primarily words that we use for this process or layout features or probably a significant document structure? Do ordinary web users even think in terms of "genres", do they categorize web documents into different types? Computational linguists try to fit all of these currently still fragmented pieces and insights together in order to build systems that are able to identify genres and especially web genres automatically. A few of these open questions are of utmost importance: how do web genres work in the hypertext environment of the World Wide Web with regard to the document, sub-document and super-document level? What kind of features could or should be used and extracted from proper documents in order to compute their respective web genres? Can scalable systems be built that are able to identify hundreds of different genres – both traditional genres and genuine web genres? How can a reference corpus of web genres be built so that the underlying theoretical assumptions inherently encoded into the corpus are met by the probably differing theoretical opinions of other researchers?

This special issue and other activities its editors and contributors are involved in has its origin in 2007. Intense networking among members of the genre community lead to a number of events. More precisely, in July 2007, Marina Santini and Serge Sharoff organised the colloquium "Towards a Reference Corpus of Web Genres" (Santini and Sharoff, 2007) which was held at Corpus Linguistics 2007 in Birmingham, UK.[2] Shortly afterwards, Marina Santini and Georg Rehm organised a follow-up workshop, "Towards Genre-Enabled Search Engines: The Impact of Natural Language Processing" (Rehm and Santini, 2007) which was held in conjunction with the conference Recent Advances in Natural Language Processing (RANLP) in Borovets, Bulgaria, in September 2007.[3] One result of this workshop was a common paper on the construction of a reference corpus of web genres (Rehm et al., 2008). Yet another event organised in 2007 was the panel[4] arranged by Mark Rosso, "Towards the Use of Genre to Improve Search in Digital Libraries: Where Do We Go from Here?". This panel was sponsored by SIG-CR and SIG-HCI and held in conjunction with ASIST 2007 in Milwaukee, Wisconsin (USA). It lead to the publication of the special issue "Bringing genre into focus" (Freund and Ringlstetter, 2008). At roughly the same time, the *Scandinavian Journal of Information Systems* published a special issue titled "Genre Lens on Information Systems" (Päivärinta et al., 2008) which brings into focus an additional perspective on genre usefulness.[5]

The interest in genre research has not declined over the years, the year 2009 is also rich of genre-related activities. First, the workshop "Automated Document Genre Classification Workshop: Supporting Digital Curation, Information Retrieval, and

[2] http://corpus.leeds.ac.uk/serge/webgenres/colloquium/

[3] http://www.sics.se/use/genre-ws/

[4] http://www.asis.org/Conferences/AM07/panels/18.html

[5] http://iris.cs.aau.dk/index.php/volume-20-40200841-no-1.html

Knowledge Extraction"[6] continues the general discussion and tries to establish a common roadmap for future projects. Second, the forthcoming book *Genres on the Web: Computational Models and Empirical studies* (Mehler et al., Forthcoming)[7] presents a wide range of conceptualisations of the notion of web genre together with an overview of computational approaches to web genre classification and structure modelling. Finally, this special issue of the *Journal for Language Technology and Computational Linguistics*, which has been conceived as the book's companion volume, is now available online with a collection of recent genre research. Taken together, the book and this JLCL special issue show the most up to date and comprehensive range of theoretical, computational and empirical research on web genres.

Structure of this Special Issue

Research on the automatic identification of web genres usually falls in one of three different categories, i. e., theoretical approaches, implementations, and the creation of resources for analysis and evaluation.

Theoretical approaches describe the extremely hard problem that web genres are able both to span multiple documents (i. e., a hypertext) as well as to instantiate only a small part of a single web page. There are, however, evolutionary processes at work that are responsible for the existence of multiple conventions and preferred instances of genres and web genres (see, for example, Rehm, 2007; Mehler, 2009). Due to these evolutionary processes, web genres come into being and, afterwards, slowly and continuously change. In their article "The Evolution of Genre in Wikipedia", Malcolm Clark, Ian Ruthven and Patrik O'Brian Holt examine how genres develop over a period of six years in the online encyclopedia Wikipedia. The authors concentrate on the biographical article and generate a number of follow-up research questions as well as plans for experimental work. A related area of research is addressed by Philip M. McCarthy, John C. Myers, Stephen W. Briner, Arthur C. Graesser and Danielle S. McNamara who present "A Psychological and Computational Study of Sub-Sentential Genre Recognition". In their article the authors describe three experiments on genre recognition based on words alone. Their conclusions not only have implications for research on the automatic identification of web genres but also for a better understanding of genre recognition in general.

Implementations are concerned with the creation of actual systems that accomplish the overall goal of this field of research. When building a genre identification system, the decision of which features to use for this process is very important. In their paper "Cost-Sensitive Feature Extraction and Selection in Genre Classification", Ryan Levering and Michal Cutler describe a complex approach for the automatic selection of features for the task of genre identification and report experimental results on two datasets. Chaker Jebari concentrates on processes that are able to categorize documents into multiple genres. In his paper "A New Centroid-based Approach for Genre Categorization of Web Pages", Jebari uses machine learning algorithms in order to compute multiple

[6] http://www.dcc.ac.uk/events/genre-classification-2009/

[7] http://sirao.kgf.uni-frankfurt.de/webgenrebook/index.html

ranks for each documents. This approach reflects the fact that a single web page often contains instances of multiple genres. It is this core problem of web genre research that is also addressed by the article "Multi-Label Approaches to Web Genre Identification". In their paper, Vedrana Vidulin, Mitja Luštrek and Matjaž Gams extract multiple features from web pages in order to test the performance of several classifiers for the task of assigning multiple genre labels to a document.

The creation of resources is closely related to the two categories mentioned above and deals with web genre corpora and datasets. In their article "Building a Corpus of Italian Web Forums: Standard Encoding Issues and Linguistic Features", Silvia Petri and Mirko Tavosanis examine linguistic properties of postings in web discussion groups and construct a corpus of these documents. They use an encoding and annotation scheme that is based on the TEI guidelines. Undoubtedly, one of the most significant gaps in current web genre research is the lack of a reference corpus of web genres that interested parties could use to evaluate their own systems based on a shared resource that was built specifically for evaluation purposes (see, for example, Rehm et al., 2008). The article "Web Genre Benchmark Under Construction" by Marina Santini and Serge Sharoff discusses this problem in detail and suggests a solution.

Acknowledgments

This special issue would not have been possible without the time and dedication of the following reviewers: Eric Atwell, Malcom Clark, Luanne Freund, Mikael Gunnarsson, Theresa Heyd, Marie-Paule Jacques, Yunhyong Kim, Nedim Lipka, Mark Rosso and Efstathios Stamatatos. Further, we gratefully acknowledge support of the German Federal Ministry of Education (BMBF) through the research project *Linguistic Networks* and of the German Research Foundation (DFG) through the Excellence Cluster 277 *Cognitive Interaction Technology* via the Project *Knowledge Enhanced Embodied Cognitive Interaction Technologies* (KnowCIT), and the Research Group 437 *Text Technological Information Modeling* via the Project A4 *Induction of Document Grammars for Webgenre Representation*.

References

Crowston, Kevin and Williams, Marie (1997): "Reproduced and Emergent Genres of Communication on the World-Wide Web". In: *Proceedings of the 30th Hawaii International Conference on Systems Sciences (HICSS-30)*. IEEE Computer Society, volume 6, pp. 30–39.

Freund, Luanne and Ringlstetter, Christoph (2008): "Special Issue: Bringing Genre into Focus". *Bulletin of the American Society for Information Science and Technology* 34 (5).

Haas, Stephanie W. and Grams, Erika S. (2000): "Readers, Authors, and Page Structure – A Discussion of Four Questions Arising from a Content Analysis of Web Pages". *Journal of the American Society for Information Science* 51 (2): pp. 181–192.

Karlgren, Jussi and Cutting, Douglass (1994): "Recognizing Text Genres with Simple Metrics Using Discriminant Analysis". In: *COLING 94 – The 15th International Conference on*

Computational Linguistics. Association for Computational Linguistics, Kyoto, volume 2, pp. 1071–1075.

Kessler, Brett; Nunberg, Geoffrey and Schütze, Hinrich (1997): "Automatic Detection of Text Genre". In: *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Meeting of the European Chapter of the Association for Computational Linguistics*. San Francisco: Morgan Kaufmann, pp. 32–38.

Mehler, Alexander (2009): "A quantitative graph model of social ontologies by example of Wikipedia". In: *Towards an Information Theory of Complex Networks: Statistical Methods and Applications*, edited by Dehmer, Matthias; Emmert-Streib, Frank and Mehler, Alexander, Boston/Basel: Birkhäuser.

Mehler, Alexander; Sharoff, Serge and Santini, Marina (editors) (Forthcoming): *Genres on the Web: Computational Models and Empirical Studies*. Springer, Berlin/New York.

Päivärinta, T.; Shepherd, M.; Svensson, L. and Rossi, M. (2008): "Special Issue Genre Lens on Information Systems". *Scandinavian Journal of Information Systems* 20 (1).

Rehm, Georg (2002): "Towards Automatic Web Genre Identification – A Corpus-Based Approach in the Domain of Academia by Example of the Academic's Personal Homepage". In: *Proceedings of the 35th Hawaii International Conference on System Sciences (HICSS-35)*. Big Island, Hawaii: IEEE Computer Society.

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

Rehm, Georg and Santini, Marina (editors) (2007): *Proceedings of the International Workshop Towards Genre-Enabled Search Engines: The Impact of Natural Language Processing*, Borovets, Bulgaria. Held in conjunction with RANLP 2007.

Rehm, Georg; Santini, Marina; Mehler, Alexander; Braslavski, Pavel; Gleim, Rüdiger; Stubbe, Andrea; Symonenko, Svetlana; Tavosanis, Mirko and Vidulin, Vedrana (2008): "Towards a Reference Corpus of Web Genres for the Evaluation of Genre Identification Systems". In: *Proceedings of the 6th Language Resources and Evaluation Conference (LREC 2008)*. Marrakech, Morocco.

Roussinov, Dmitri; Crowston, Kevin; Nilan, Mike; Kwasnik, Barbara; Cai, Jin and Liu, Xiaoyong (2001): "Genre Based Navigation on the Web". In: *Proceedings of the 34th Hawaii International Conference on Systems Sciences (HICSS-34)*. IEEE Computer Society.

Santini, Marina and Sharoff, Serge (editors) (2007): *Proceedings of the Colloquium Towards a Reference Corpus of Web Genres*, Birmingham, UK. Held in conjunction with Corpus Linguistics 2007.