Cross-Testing a Genre Classification Model

Marina Santini
MarinaSantini.MS@gmail.com

Abstract

A genre classification model is presented and cross-tested with a number of genre collections. This model provides some insights into open issues in Automatic Genre Identification.

1 Introduction

This poster paper focuses on the automatic identification of genre. Broadly speaking, genres indicate *how* information is *packaged* in a certain pragmatic and communicative context. For instance, the EDITORIAL genre indicates that a document contains an argumentation representing a newspaper or a magazine as a whole, and this argumentation has the power to shape public opinion. EDITORIALS, REPORTAGES, ACADEMIC PAPERS, BLOGS or HOME PAGES are some of the many genres characterizing today's communication, in the paper world or in digital environments.

Genres are based on more or less tight conventions. The recognition of these conventions allows people to reconstruct or infer the context in which texts have been produced, together with their purposes and functions. For this reason, genre information would benefit all those fields where language variation is important, and especially research areas where language technology can be enhanced or refined by a more fine-grained document typology, e.g. corpus linguistics, Natural Language Processing (NLP), information retrieval, automatic summarization, or machine translation. Automatic Genre Identification (AGI) would then be a great advantage for many researchers, since the manual annotation of documents by genre is very expensive, time-consuming and often controversial.

To date, although plentiful, AGI findings are still fragmented. One of the reasons underlying this fragmentation is the lack of genre benchmarks that would permit more systematic comparisons, and help understand how stable AGI models are, the maximum number of genres that can be identified automatically in a certain environment (e.g. intranets, digital libraries, or the web), which features are the most profitable, how robust genre models are to noise and how they can be updated with emerging genres or purged from vanishing ones.

In this poster paper, a genre classification model is presented and cross-tested with a number of genre collections[1]. This model provides some insights into AGI open issues.

2 A Genre Model in a Difficult Scenario

The model discussed in this section is not based on supervised learning, but on a simplified form of the *subjective Bayesian method* and on inferential *if-then* rules. Since it relies on "inference" rather than "learning", we refer to it as the *genre inferential model*. This model has already been presented with a partial evaluation (e.g. see Santini, 2007). Here, a more comprehensive evaluation is carried out.

The genre inferential model relies upon linguistically rich features and follows a multi-label scheme. The model has been devised for an open digital environment, like the web, where the level of noise is high and the population is difficult to approximate. The task of the model is to apply either *no genre* – when a document is highly individualized – or *one genre* – when the document belongs to a single genre – or *multiple genres* – when a document contains several genres or is hybrid. The genre palette (i.e. the genres that the model can automatically identify) is static and includes two different types of genres, namely four rhetorical genres and seven social genres. The underlying assumption of this double-layered genre model is that rhetorical genres, which represent universal communicative purposes, help harness the instability of the web or other noisy digital environments, because they are more stable than social genres. Social genres are historical entities, so they come and go, and are very linked to technology (like BLOGS or HOME PAGES).

This version of the genre inferential model relies on a web corpus and on the following steps:

- Automatic extraction of functionally-motivated features from the web corpus.
- Inference of four rhetorical genres (DESCRIPTIVE–NARRATIVE, EXPLICATORY–INFORMATIONAL, ARGUMENTATIVE–PERSUASIVE and INSTRUCTIONAL) using the subjective Bayesian method. Inferred rhetorical genres are associated to a probability value.
- Probabilities are interpreted in terms of "gradations", and ranked in descending order.

After the ranking, two hypotheses are tested:

[1] These collections are available from the WEBGENREWIKI <http://purl.org/net/webgenres>.

- The first hypothesis says that the combination of a number of rhetorical genres is sufficient to derive four BBC web genres (EDITORIALS, DIY MINI-GUIDES, SHORT BIOGRAPHIES and FEATURE ARTICLES), more traditional in their textuality.
- The second hypothesis says that the combination of two predominant rhetorical genres, i.e. the top-ranked ones, plus a combination of additional traits is sufficient to derive seven web genres (BLOGS, ESHOPS, FAQS, FRONT PAGES, LISTINGS, PERSONAL HOME PAGES, and SEARCH PAGES), more influenced by the functionalities allowed by the web.

This model achieves an accuracy of about 86% on an initial corpus of 2480 web pages (see Table 1) on a single label.

Genres	# web pages	Proportions
NOISE (randomly selected web pages from the SPIRIT collection)	1000	40.32%
Blogs	200	8.07%
Eshops	200	8.07%
FAQs	200	8.07%
Front pages	200	8.07%
Listings	200	8.07%
Personal Home Pages	200	8.07%
Search Pages	200	8.07%
BBC Editorials	20	0.81%
BBC DIY mini-guides	20	0.81%
BBC Short Biographies	20	0.81%
BBC Features	20	0.81%
Total	2480	100%

Table 1. Santini's initial web corpus

In order to test its robustness to scalability, an increase of corpus size was simulated by adding another genre collection to the initial corpus, namely the KI-04 corpus (Meyer zu Eissen and Stein, 2004), containing 1205 web pages. On this enlarged corpus (i.e. 3685 web pages), the accuracy achieved on a single genre is about 81%. We deemed these results encouraging, since a size increase of 35% causes only 5% decrease in accuracy. This evaluation was considered to be partial because the multi-label classification part of the model could not be tested, since there were no multi-genre labelled collections available up to very recently, and the SPIRIT collection included in the initial corpus was not annotated by any genre at that time. It is worth pointing out that the SPIRIT collection included in the initial corpus represented the *noise* that can be found on the web. In this model, noise can be paraphrased as "DON'T KNOW". Simply put, the SPIRIT collection is a random slice of the web whose content is unknown. Therefore, it contains not only genres that are different from those included in the model's genre palette, but also genres that might be in the palette. Since we do not know the number and the distribution of genres on the web, this DON'T KNOW class is an attempt to bypass the constraint underlying machine-learning-based models, where the documents must be necessarily pre-assigned to known and well-defined classes.

In this new and final part of the evaluation, the automatic classification of the SPIRIT collection, recently annotated by the author of the inferential model, is assessed. Additionally, two new genre collections have been employed for comparative analyses, i.e. a hierarchical genre collection (Stubbe and Ringlstetter, 2007), and a multi-labelled genre collection (Vidulin et al. 2007).

All in all, the model has been cross-tested with 6404 genre-annotated web pages. We conjure that this final composite corpus of 6404 web pages well represents a noisy environment like the web, where documents come from disparate communities enacting different genre conventions and classification schemes.

In this difficult scenario, the genre model shows some robustness and stability, but results are still far from optimal. This conclusion conforms to the assessment of other genre-enabled applications, i.e. WEGA genre add-on (Stein et al., 2008; Santini and Rosso, 2008), and X-Site (Freund, 2008).

More specifically, the major problem that negatively affects the genre inferential model is the conceptual elusiveness and lack of mutual understanding of genre classes. This leads to conceptually different classes bearing similar names. One example is the class referring to "online stores". In Santini's web corpus, ESHOPS are intended as interactive documents, with their own purpose (i.e. selling products), rhetorical function (persuade potential buyers), textual conventions (e.g. use of exhortations), and expectations (e.g. prices, special offers, etc.); in KI-04, SHOPS are "all kinds of pages whose main purpose is product information or sale"; more broadly, Vidulin et al.'s SHOPPİNG class includes online stores, classified ads, price comparators and pricelists. The inferential model performs disappointingly on Vidulin et al.'s SHOPPİNG class because it was not designed to cover classified ads and price comparators.

In conclusion, it seems that the diverse definitions of the concept of genre have a strong bearing on the characterization of genre classes, thus affecting the generability of AGI models.

References

L. Freund. 2008. Exploiting task-document relations in support of information retrieval in the workplace. PhD thesis. University of Toronto.

S. Meyer zu Eissen S. and B. Stein. 2004. Genre Classification of Web Pages: User Study and Feasibility Analysis. In Biundo S., Fruhwirth T. and Palm G. (eds.). KI 2004: Advances in Artificial Intelligence. Springer.

M. Santini. 2007. Automatic Genre Identification: Towards a Flexible Classification Scheme. FDIA 2007.

M. Santini and M. Rosso. 2008. Testing a Genre-Enabled Application: A Preliminary Assessment. FDIA 2008.

B. Stein, S. Meyer zu Eissen, and N. Lipka (in preparation). Web genre analysis: Use cases, retrieval models, and implementation issues.

A. Stubbe and C. Ringlstetter. 2007. Recognizing Genres. Colloqium "Towards a Reference Corpus of Web Genres", Birmingham.

V. Vidulin, M. Luštrek, and M. Gams. 2007. Using genres to improve search engines. Workshop "Towards Genre-enable Search Engines: The Impact of NLP", Borovets.