EMPIRICAL TEXT MINING FOR GENRE DETECTION

Vasiliki Simaki[1], Sofia Stamou[1,2] and Nikos Kirtsis[2]
[1]*Computer Engineering and Informatics Department, Patras University, Patras, Greece*
[2]*Department of Archives and Library Science, Ionian University, Corfu, Greece*

Keywords: Genre Detection, Annotation, Human Study.

Abstract: In this paper, we report on a preliminary study we carried out for identifying patterns that characterize the genre type of Greek texts. In the course of our study, we address four distinct genre types, we record their observable stylistic elements and we indicate their exploitation for automatic genre-based document classification. The findings of our study demonstrate that texts contain lexical features with discriminative power as far as genre is concerned, however modeling those features so that they can be explored by computer-based applications is still in early stages.

1 INTRODUCTION

The genre is the totality of characteristics we observe in a text that gives a unique print. It is an heterogeneous categorical principle in that it provides clues for classifying texts into specific styles. Genre clues are closer to the field of semantics and unlike text type that conveys information about the texts' structure, they give out information about the style of a text. Genre clues are employed for characterizing a text as subjective/objective, positive/negative about the subject it elaborates, opinionated, factual, etc. They may also be employed for unravelling stylistic features reflected within a text's content, with these ranging from literary content to procedural, descriptive and so forth.

In this paper, we report on an observatory study we carried out in which we try to identify structural and contextual clues within text in order to be able to characterize its underlying genre accurately. Our observations verify that there are discrete informational points within texts that constitute their genre clues. Based on this finding, we take a step further and try to mine lexical and syntactic patterns from text, which may be treated as genre indicators for automatically classifying texts according to genre classes.

2 RELATED WORK

Automatic genre and style detection of texts is an active field of research over the past years essentially because there are many challenges and thus yet-unresolved issues associated with it. In this respect, several researchers have proposed various solutions to the problem of identifying the characteristics that communicate the genre and the stylistic print of texts. In particular, Kessler et al. (1997) proposed a theory of genres that compares various surface textual features and they attempted the automatic genre detection via the exploitation of genre clues in a text's body. The clues they determined are: *structural*, *lexical*, *character-level* and *derivative*; and, despite their surface nature, they are as successful as deep structural cues. Finn et al. (2002) proposed the application of machine learning techniques for the automatic classification of text genres. They investigated three approaches to automatically classify text documents by genre, i.e. the bag-of-words method, part-of-speech statistics and handcrafted shallow linguistic features. The application of their approach to a collection of news articles revealed that part-of-speech tagging and statistical analysis of a text's content may effectively contribute towards genre-based document classification. Later, Finn and Kushmerick (2003) elaborated on the details of textual genres and showed how genre can be distinguished across different texts. Lee and Myaeng (2004) implemented the so-called '*deviation-based statistical feature selection*' method and proved its efficacy in automatically extracting genre-related features from texts. Santini et al. (2006) proposed a new definition for genre based on the traits

Simaki V., Stamou S. and Kirtsis N..
EMPIRICAL TEXT MINING FOR GENRE DETECTION.
DOI: 10.5220/0003956207330737
In *Proceedings of the 8th International Conference on Web Information Systems and Technologies* (WEBIST-2012), pages 733-737
ISBN: 978-989-8565-08-2
Copyright © 2012 SCITEPRESS (Science and Technology Publications, Lda.)

of *hybridism*, *individualization* and *evolution*. This theoretical characterization constitutes the base for an inferential model that automatically identifies genres within web pages. Later, Santini (2007) presented a model based on a scheme performing zero-to-multi genre assignments. Sharoff (2007) performed experiments to find the best set of features that delimit a typology useful for classifying web pages according to their domain and genre. With respect to Greek, Stamatatos et al. (2000) described an approach that relies on classification schemes and natural language processing modules for categorizing Greek texts according to their genre and author. The application of their technique to various Greek corpora proved the feasibility of their proposed method.

3 METHODOLOGY

Our text genre detection approach operates upon pre-determined classes of text types in which it looks for features that signify the genre of the respective texts. Before delving into the details of our method, we should point out that we distinguish text genre from text type in that the former represents the way in which information is communicated via texts, whereas the latter represents the type of information texts convey. As an example consider the current paper, whose style might be characterized as technical or formal and whose type might be characterized as scientific publication or research paper.

Based on the discrimination between style and type, we propose a method that tries to identify the genre (i.e. style) of texts that belong to particular types. To obviate the need for pre-classifying texts according to their type, we rely on an existing textual type classification scheme[1] that defines the following categories of text types: (i) ***narrative***: an account of events, (ii) ***expository***: text that explains something, (iii) ***procedural***: text that gives instructions on how to do something and (iv) ***descriptive***: text that lists the characteristics of something. Each of the above category types contains sub-types, which share common elements in structure and content with their parental type and with each other, and at the same time they deliver clues that are unique and representative of their respective sub-type category. As sub-type example categories consider the following: (i) narrative texts may be sub-categorized into *novels*, *poetry* or *short stories*, (ii) expository texts may be sub-categorized into *news articles*, *travel books* and *periodicals*, (iii) procedural texts may be sub-categorized into various kinds of *guidebooks*, e.g. cooking books, manuals, installation guides and (iv) descriptive texts may be sub-categorized into *encyclopedias*, *grammars* and *dictionaries*. Given the type and/or sub-type of a text, our method tries to identify structural and contextual elements that signify the text's style or else genre. To ensure the accurate identification of genre, we relied on a dataset of texts already classified into one of the above types and after manually inspecting them we extracted their genre-indicative elements. Manual inspection of texts was performed by human expert annotators, i.e. linguistics, to whom we distributed a number of texts along with their respective type categories and we asked them to indicate for every text type the features (both contextual and structural if feasible) that yield its style. Note that we did not ask our study participants to directly indicate the genre of every examined text type, but rather to indicate the elements that shape the stylistic identity of every examined text. This is because the quest in our study is to determine the characteristics that distinguish the genre of text types from one another rather than annotate texts with stylistic information. Based on the above determination, one can employ supervised clustering methods for automatically grouping texts in terms of their common underlying stylistic elements. Turning back to the description of our method, we should point out that due to space and time constraints, we did not examine all the sub-categories of every text type but rather we relied on texts categorized under a specific (randomly selected) type sub-category. Thus, for the narrative category, we relied on texts belonging to the *short stories* sub-category, for the expository category we relied on texts belonging to the *news articles* sub-category, for the procedural category we relied on texts belonging to the *cookbooks* sub-category and for the descriptive category we relied on texts belonging to the *dictionaries* sub-category.

The dataset used for each of the above categories, concerned online texts that we harvested from suitable websites, i.e. sites containing (i) short stories, (ii) online newspapers, (iii) sites about recipes and cooking and (iv) lexicographic sites. In the next section, we report on the statistics of our collected dataset. Now, we present the way in which human annotators assessed our collected data sources in order to identify their structural and contextual elements that signify their genre. To carry out our study, we recruited 5 experienced linguists (3 female, 2 male), who volunteered to read the texts collected for every type category and indicate the

[1] http://www.sil.org/linguistics/glossaryoflinguisticterms/WhatIsAText.htm

elements that they judged as indicative of the texts' style. The only instruction given to our study participants was that their role would not be to indicate the texts' genre but rather to identify elements that demonstrate the style. To acquaint our users with their task, we run a supervised laboratory experiment in which we asked them to experiment with a number of texts and via the think-aloud protocol to indicate which elements they deemed as style-descriptive for each of the test texts. Note also that our volunteers were aware of the type sub-category of every text they examined. Finally, we advised our participants to indicate only elements for which they felt absolutely confident that they are indicative of the texts' genre and we did not allow them to communicate with each other during their participation in the survey. The time duration of the study was five consecutive days, with two 3-hour sessions per day. At the end of the study, we collected the notes (in the form of free-style comments) our participants delivered, we grouped them by text type and we examined for every text type the genre-indicative elements that were selected by the majority of our study participants. Obtained results for each of the text types considered are presented and discussed in the following section.

4 IDENTIFYING GENRE CLUES

Table 1 summarizes our experimental dataset.

Table 1: Statistical distribution of the examined dataset across text types.

Documents	**853**
Short stories (narrative)	157
News articles (expository)	359
Recipes (procedural)	334
Dictionaries (descriptive)	3

Having collected the experimental texts for our human survey, we grouped them by type and we asked our study participants to read the texts in every type and note down the clues they considered indicative of the respective texts' genre. To facilitate the work of our volunteers, we asked them to write down in the form of (self-selected) keywords and/or short phrases brief descriptions of the clues they identified. All texts across all four types were examined by all the participants, thus we received feedback for the entire dataset. Based on the user-selected genre clues for each of the text types considered, we built an indexing module where we stored the user-selected genre clues for further processing. Genre clues processing concerned manual assessment of the collected feedback to identify recurrent genre identifiers. In our study, we deem a genre clue as recurrent in the user feedback if at least three of the participants indicated the same clue as genre-indicative even if the terms used to verbalize the clue were not identical among their comments. Table 2 gives the statistics of the user-defined genre clues for each of the types examined.

Table 2: User-selected genre clues for text types.

Text type: short stories	
Genre clues selected by more that 3 users	5
Genre clues selected by 3 or less users	9
Text type: news articles	
Genre clues selected by more that 3 users	4
Genre clues selected by 3 or less users	9
Text type: recipes	
Genre clues selected by more that 3 users	4
Genre clues selected by 3 or less users	6
Text type: dictionaries	
Genre clues selected by more that 3 users	3
Genre clues selected by 3 or less users	5

According to the reported data, there are high levels of user agreement with respect to the elements that signify the genre of different text types. This might be due to the fact that all our study participants were experienced linguistics with a solid knowledge of what text genre is and how it can be encapsulated in the structural and lexical elements of texts.

4.1 Analysis of Genre Clue Indicators

Having collected and processed user feedback with respect to what constitute the indicative characteristics of genre within texts, we proceed with the presentation and analysis of those indicators in an attempt to shed light on the text properties that signify style. Table 3 reports the stylistic characteristics of texts belonging to the each of the categories examined as given by at least 3 of our participants.

As the table shows, there are relatively high levels of repetition among the genre clues that characterize text types. To identify how genre clues are pronounced into the structural and contextual properties of their corresponding texts, we relied on the combined analysis of the texts employed in our study and the comments our participants delivered for each of the examined texts and we interpret our findings as follows.

For the ***short stories*** category, the limited text size and the quick narration are the most frequently indicated clues of genre. The text size feature accounts to the number of words narrative texts contain, which according to our data ranges between

Table 3: Genre clues for the ***short stories, new articles, recipes and dictionary*** categories respectively.

Text type: short stories		**Text type: news articles**	
Genre clue	**Fraction**	**Genre clue**	**Fraction**
Limited text size	61.14%	Headings	100.00%
Quick narration	57.89%	Images, tables	100.00%
Presence of dialogue	47.77%	Syntactic coherence	22.31%
Syntactic coherence	44.45%	Named entities	13.74%
Recursion	40.25%		
Text type: recipes		**Text type: dictionaries**	
Genre clue	**Fraction**	**Genre clue**	**Fraction**
Structure replication	100.00%	Structure replication	100.00%
Sequential short NPs	25.77%	Short, elliptical phrases	98.00%
Precise verbal structures	11.88%	Use of abbreviations/symbols	23.00%
Use of command words	10.51%		

1,000 and 3,500 terms. The quick narration feature accounts to the use of descriptive language, the succession of events (usually in a chronological order), the use of pronouns and the presence of factual statements within text. Modeling the characteristic of quick narration is a quite complex task that requires large volumes of human-annotated data (i.e. narrative corpora) as well as the availability of sophisticated language processing modules. The syntactic coherence feature accounts to the presence of full syntactic phrases within texts and the recursion feature accounts to embedding phrases of the same type within texts. Both of them together with the dialogue feature are genre descriptive elements for narrative texts and can be captured after applying syntactic parsing to the contents of the respective texts.

The genre clues identified for the ***news articles*** category have strong discrimination power as these are present in the striking majority of journalistic texts. In particular, we observe that what characterizes the genre of a text as expository and specifically as journalistic is the presence of headings, images and the utilization of named entities in the text contents. Modeling the above features in order to come up with automated genre detection methods is relatively easy. This is because most of the features that characterize expository texts and *news articles* can be identified within the texts' structural properties where the application of shallow parsing would be sufficient for their detection. For the remaining features that serve as indicators of expository texts, we would need to apply deep syntactic analysis to the texts' contents to be able to automatically extract them.

The genre clues identified for procedural texts are pretty different from both narrative and expository texts with their main differences accounting for the texts' structural properties. In particular, texts belonging to the ***recipes*** type replicate the same structural and syntactic patterns, i.e. they begin with a title/heading, which is followed by a list of sequential short noun phrases (NPs), which are topically related to each other. Noun phrases are followed by short text nuggets, which contain simple and repetitive terminology, command words as well as precise verbal structures, i.e. verbs in indicative plural. Our data indicates that we can capture and therefore model the stylistic features of procedural texts by relying mainly on the texts' structure and less on their content. This facilitates the genre characteristics modeling task as it does not require the application of deep lexical analysis to the examined texts. Of course we are aware of the fact that different types of procedural texts may vary in their structure, but still one could apply our user-selected features as a starting base for detecting texts that describe a process of doing something.

Finally, ***dictionaries*** are specialized instances of text with unique stylistic features. As such, we can safely determine their genre by relying on the analysis of their structural elements, i.e. by looking for the pattern: *lemma entry + sequences of short elliptical phrases* that may contain abbreviations and/or symbols. Although instances of the above patterns may vary across different categories of descriptive texts, still the structure of their contextual elements, i.e., lemma entries followed by short descriptions, is what constitutes their genre print.

5 CONCLUDING REMARKS

We have presented an exploratory study for identifying the features that characterize the genre of different text types. Our findings show that there exist specific structural and contextual elements within texts that can be modeled as genre clues in order to be explored by automatic genre classification modules. In the future, we plan to investigate ways of semi-automatically assigning texts of particular

types to pre-specified genres.

REFERENCES

Finn, A. and Kushmerick, N. 2003. Learning to classify documents according to genre. In *Proceedings of the Computational Approaches to Style Analysis and Synthesis Workshop*.

Finn, A., Kushmerick, N. and Smyth, B. 2002. Genre classification and domain transfer for information filtering. In *Proceedings of the European Colloquium on Information Retrieval Research*, pp. 353-362, Glasgow.

Karlgren, J. 1999. Stylistic experiments in information retrieval. *Natural Language Information Retrieval*, Kluwer.

Lee, Y. B. and Myaeng, S. H. 2004. Automatic identification of text genres and their roles in subject-based categorization. *In the 37th Hawaiian Conference on System Sciences.*

Santini, M., Power, R. and Evans, R. 2006. Implementing a characterization of genre for automatic genre identification of web pages. *ACL Computational Linguistics Conference*.

Santini, M. 2007. Automatic genre identification: towards a flexible classification scheme. In *the BCS IRSG Symposium: Future Directions in Information Access*, Glasgow, Scotland.

Sharoff, S. 2007. Classifying web corpora into domain and genre using automatic feature identification. In *the Web as Corpus Workshop*, Louvain-la-Neuve.

Stamatatos E., Fakotakis N. and Kokkinakis G. 2000. Automatic text categorization in terms of genre and author. *Computational Linguistics*, vol.26, no.4, pp. 461-485, MIT Press

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS