An Exploratory Study on Genre Classification using Readability Features

Johan Falkenjack, Marina Santini, Arne Jönsson

Department of Computer and Information Science, Linköping University, Linköping, Sweden
SICS East Swedish ICT AB, Linköping, Sweden
johan.falkenjack@liu.se, marinasantini.ms@gmail.com, arne.jonsson@liu.se

Abstract

We present a preliminary study that explores whether text features used for readability assessment are reliable genre-revealing features. We empirically explore the difference between genre and domain. We carry out two sets of experiments with both supervised and unsupervised methods. Findings on the Swedish national corpus (the SUC) show that readability cues are good indicators of genre variation.

1. Introduction

Texts in national corpora are typically classified into a number of text categories that ensure that a corpus is balanced and representative of a certain language at a certain point in time. This is the case of the Swedish national corpus, called Stockholm-Umeå Corpus or the SUC. The SUC is a collection of Swedish texts from the 1990's, consisting of one million words (Gustafson-Capková and Hartmann, 2006). The SUC follows the general layout (with some exceptions) of the Brown corpus and LOB corpus, which were compiled in the 60s and 70s respectively.

Following a well-established tradition, we use the word "genre" to refer to the text categories in these corpora. However, we are aware that some of these genres are topical, i.e. related to a domain (e.g. Religion or Popular lore), other genres are based on formal/structural/textual similarities (e.g. Reviews or Reportages), and, finally, one genre is mixed (Miscellaneous). Since we expect that these differences affect the behaviour of automatic genre identification models, in this study we argue that, in order to be effective, the automatic classification of domain and genre requires different types of features. This is not a novelty, and some researchers stick to this distinction consistently (e.g. see Sharoff (2007)). As a matter of fact, the distinction between genre and domain is unclear to many, and often these notions are merged together. It starts being acknowledged, however, that bundling domain and genre together may cause erratic errors in NLP applications. Recent studies point out that it is beneficial to keep the two concepts apart (e.g. see van der Wees et al. (2015) and their experience in Statistical Machine Translation).

In this preliminary study, we start exploring possible approaches to pin down an empirical distinction between genre and domain.We present two sets of experiments, one based on unsupervised classification and the second based on supervised classification. More specifically, in the first set we explore how well agglomerative hierarchical clustering performs on the genres included in the SUC. To our knowledge, genre clustering has never been applied to the SUC, but it is worth exploring because it would have the obvious advantage, with respect to supervised classification, of skipping manual genre labelling, which is often controversial (e.g. see Santini (2008)), and always expensive in terms of time and resources.

In the second set of experiments, we investigate whether, empirically, genre modelling would benefit from the separation of subject-based text categories, or "domains", from stylistically- and rhetorically-based textual classes, i.e. "proper" genres (see Section 3).

In both these sets of experiments we wish to explore how well readability assessment features represent genres. The rationale behind this choice lies in the observation, initially put forward by Karlgren and Cutting (1994), that typically genres show different level of readability, for instance highbrow v.s. lowbrow lexicon or simple v.s. complex syntax.

To our knowledge, none of these research lines has been previously investigated in the SUC or in other corpora. The insights provided by the findings may be useful for the improvement of data-driven NLP applications, such as machine translation, domain adaptation, parsing, word sense disambiguation or text simplification.

2. Previous work

Although still limited in terms of scalability, automatic genre classification research has a solid tradition within NLP. For recent advances of this field, see Mehler et al. (2010). The main trend relies on supervised methods. Supervised genre models are described in the most important papers of the field, from the seminal papers by Karlgren and Cutting (1994) and Kessler et al. (1997), who use discriminant analysis and logistic regression/neural networks respectively, to the most recent genre classification experiments often based on SVM. Exploratory unsupervised methods, such as factor analysis and cluster analysis, have been applied in Multi-Dimensional Analysis, an approach created by Biber (1988) to analyse linguistic variation within corpora that are pre-classified into genres and/or registers. Unsupervised algorithms have been used sporadically for automatic text genre classification (e.g. see Santini (2005) or Bekkerman et al. (2006)). In this study, we explore comparatively how supervised and unsupervised methods perform on the SUC.

Many sets of features have been proposed in genre classification research, from Parts-Of-Speech (POS) to Bag-of-Words (BoW), from syntactic patterns to character n-grams. To date, we cannot state however that there exists a

universal feature set that can help us reveal genre automatically. The representativeness of genre features is correlated to many factors, such as the size of the genre collection, the quality or the types of genres, and so on. These correlations have not been fully investigated. In this study, we explore the potential of readability assessment features as genre-revealing features. Empirically, features used for readability assessment have been proved to capture variation within and across genres (e.g. see Dell'Orletta et al. (2014)).

Previous genre classification experiments carried out on the Brown corpus (Karlgren and Cutting, 1994) and on the SUC (Wastholm et al., 2005) show that classifiers tend to perform better when the number, quality and type of the genre classes are taken into consideration. For instance, Wastholm et al. (2005) acknowledge that the SUC's text categories have a diversified nature and unmistakably the authors state: "some of the SUC's "genres" are in fact subject-oriented.". Additionally, they warn future researchers: "To anyone wishing to further refine our genre recognizer, we suggest that some resources be put into defining a more complete and uniform set of genres". Well-aware of these differentiations, we incorporate their suggestion in our experiments, and try to empirically explore the difference between genres and domains in the SUC.

3. Text Categories: Genre and Domain

There is a long-standing but still ongoing discussion on the definition of the concept of genre. Genre is a multi-faceted notion whose characterization overlaps with neighbouring textual dimensions, such as domain, register or style. Some of these discussions have been summarized in Lee (2001).

In this study, we wish to disentangle the notions of genre and domain empirically. This discussion is relevant to automatic text classification in general because it has a bearing on feature selection and on the performance of automatic classifiers. We propose the following theoretical distinctions between the concepts of genre and domain:

- Domain is a subject field. Domain refers to the shared general topic of a group of texts. For instance, "Fashion", "Leisure", "Business", "Sport", "Medicine" or "Education" are examples of broad domains. In text classification, domains are normally represented by topical features, such as content words.
- Genre is a more abstract concept. It characterizes text varieties on the basis of conventionalized textual patterns. For instance, an *academic paper* obeys to textual conventions that differ from the textual conventions of a *tweet*; or a *letter* complies to conventions that are different from the conventions of an *interview*. *Academic papers*, *tweets*, *letters*, *interviews* are examples of genres. Genre conventions usually affect the organization of the documents (its rhetorical structure and composition), the length of the text, the syntax and the morphology (e.g. passive forms v.s. active forms), vocabulary richness, etc. In text classification, genres are often represented by features such as POS tags, character n-grams, or POS n-grams.

If we apply this distinction to the nine top genres included in the SUC, we end up with six "proper" genres (i.e. *Press Reportage* (***A***), *Press Editorial* (***B***), *Press Review* (***C***), *Biographies/Essays* (***G***), *Learning/Scientific Writing* (***J***) and *Imaginative Prose* (***K***); two subject-based categories or domains (Skills/Trades/Hobbies (**E**) and Popular Lore (**F**); and a mix category (Miscellaneous (**H**)). The total number of subgenres is 48. See the Appendix for a breakdown.

It is worth noting that while the six proper genres can virtually contain any topics, because the concept of genre is not necessarily binding in terms of content, the two domains are subject-specific and can virtually contain any genres. Domain and genre are not perfectly orthogonal in real life, because there might be some correlations created by use (e.g. the *recipe* genre contains food-related topics), but they tend to highlight different textual properties.

4. Experimental settings

In our experiments, we explore two research hypotheses.

Hypothesis 1. Since previous research points out that readability assessment features are potentially good genre-revealing features, we put forward the hypothesis that these features show some degree of robustness in the identification of SUC genres even when used with an unsupervised method, such as hierarchical clustering (see Section 4.1).

Hypothesis 2. We assume that domain and genre are two different notions that are not represented by the same type of features. We put forward the hypothesis that since readability assessment features are genre-revealing features, they work better on proper genres and less efficiently on domains or mixed text categories. We test this hypothesis in a supervised classification paradigm (see Section 4.2).

Corpus: The SUC contains 500 samples of texts with a length of about 2,000 words each. Technically speaking, the SUC is divided into 1040 bibliographically distinct text chunks, each assigned to a genre and subgenre. The SUC contains nine top genres and 48 subgenres.

Features: In this study we use the full set of 118 features proposed for assessment of readability for the Swedish language (Falkenjack et al., 2013). This feature set contains a mixture of 21 lexical, morphological and syntactic features, such as average sentence length, highly frequent lemmas, and average dependency distance. It also comprises 13 combined measures, such as LIX and OVIX.

4.1 Agglomerative Hierarchical Clustering

The goal of this set of experiments is to explore to what extent the combination of agglomerative hierarchical clustering and readability assessment features make sense of SUC's genres. Since clustering does not rely on labelled examples, it needs robust features capable of revealing sensible patterns in data. We used the AGNES function of the cluster package in R, and the Ward linkage, which is usually more robust in the presence of noise.

The results of these experiments are shown in Table 1, second and third columns. AHCW stands for Agglomerative Hierarchical Clustering with Ward's linkage. We report the BCubed F1 as well as the weighted average regular F1-scores, where each cluster has been classified according to majority membership. Regular F1 scores (e.g. the harmonic mean of precision and recall) is indicative, but "favors coarser clusterings, and random clusterings do not

(Exp n.) SUC genres	AHCW		SVM	NB
	Avg F1	F(BCubed)	Wgh F1	Wgh F1
(Exp1) 48 subgenres	0.257	0.226	0.358	0.329
(Exp2) 9 genres (**A, B, C, E, F, G, H, J, K**)	0.386	0.341	0.628	0.541
(Exp3) 8 genres (without **H**)	0.451	0.459	0.689	0.609
(Exp4) 6 selected genres(**A, B, C, G, J, K**)	**0.555**	0.460	**0.824**	**0.714**
(Exp5) 2 domains + Miscellanous (**E, F, H**)	0.534	0.530	0.644	0.625
(Exp6) 2 domains (**E, F**)	0.675	0.649	0.786	0.737

Table 1: Supervised and unsupervised methods: F-scores

receive zero values" as observed by Amigó et al. (2009). Since these authors suggest that BCubed is a better measure, we calculate it following their formulas.

In this set of experiments, texts are clustered into the number of genres indicated in the experiment number. For example, in Exp1 (first row), the full set of 1040 texts is clustered into 48 clusters, i.e. the same as the number of subgenres, and each cluster is classified according to its majority subgenre label.

As expected, the higher the number of clusters, the sparser they are. In some cases, genres do not have a matching cluster. For example in Exp1, many subgenres do not have a matching cluster (31 out of 48). However, we can observe a consistent regularity across the clustering experiments: readability assessment features consistently identify **K**, **A** and **J**. In particular, these features appear to be robust features for the **K** (Imaginative prose), which is a genre that usually differentiates itself from factual genres because of the richness of stylistic and rhetoric devices. This characterization is well captured in the results that show the best performance on this genre. For instance, in Exp2 **K**'s F1 is 0.62 (followed by **J**'s F1=0.53, and **A**'s F1=0.50) and in Exp4 **K**'s F1 is 0.69 (followed by **A**'s F1=0.66 and **J**'s F1=0.54).

4.2 Supervised Classification: SVM and Naïve Bayes

In this set of experiment, we try to identify the computational difference between genres and domains. As already explained in Section 3, while genres are defined on the basis of organizational, rhetorical, syntactic and morphological devices, domains relates to the subject field or the content of documents. Since the readability assessment feature set contains lots of grammatical information, in this set of experiments we put forward the hypothesis that these features are expected to perform better on proper genres (i.e. **A, B, C, G, J, K**) than on domains and mix classes (i.e. **E, F** and **H**). More specifically, we compare the robustness of readability assessment features in different grouping using SVM and Naïve Bayes (See Table 1, forth and fifth column). We rely on the Weka workbench (SMO and Naïve Bayes with standard parameters, 10-fold-crossvalidation) and report weighted F-scores.

Usually the performance of a supervised classifier increases when the number of classes decreases, so it is remarkable that the performance of SVM on the six proper genres (i.e. **A, B, C, G, J, K**) is undeniably higher than the performance on two domains and a mix class (i.e. **E, F** and **H**), namely 0.824 v.s. 0.644. The performance of the Naïve Bayes (when compared with SVM) suggests that this algorithm is not the best suited for readability features. It must be noticed however that the trend stays the same. In general, we can observe that the six proper genres (Exp4) are strongly discriminated if compared with 2 domains and 1 mix class (Exp5. Interestingly, this tendency emerges also with AHCW Avg F1 (but it is not so evident with AHCW BCubed). This divide is less noticeable when Exp4 is compared with Exp6. Arguably, these results confirm Hypothesis 2 and show empirically the existence of a theoretical divide between genres and domains.

In Table 2, we compare our results with Wastholm et al. (2005) based on accuracy values (since this is the evaluation measure used by the authors). Wastholm et al. (2005) report comparative experiments where different feature sets are tried out on the SUC's genres using a Naïve Bayes classifier. They get the best accuracy with POS trigrams (namely, 60%).

SUC genres	NB (WKM2005)	NB	SVM
9 genres (**A, B, C, E, F, G, H, J, K**)	60%	53.3%	**64.1**%

Table 2: Supervised classification: accuracies

The combination of SVM and readability assessment features outperforms the baseline reported in Wastholm et al. (2005). Namely SVM has an accuracy of 64.1% on the nine top genres, while Wastholm et al. (2005) report an accuracy of 60% using POS trigrams. This seems to confirm the goodness of the readability assessment feature set in combination with SVM. SVM is a powerful algorithm and makes the best of readability features on SUC's nine noisy classes (i.e. a mix of domains and genres). We observe that the combination of readability features and Naïve Bayes classifier seems not to be ideal since accuracy is rather low (i.e. 53.3%). As already emerged in previous research (e.g. Sharoff (2007)), POS trigrams are robust features for genre classification because they incorporate shallow but effective syntactic information that well differentiate across genres. Possibly, a combination of POS trigrams, readability features and SVM would outperform this set for results. We point out however that since SUC's genres, when taken as a whole, are noisy classes, any classification algorithms can predictably underachieve.

5. Conclusions

In this paper, we presented two sets of experiments and showed that readability assessment features can be profitably used as genre-revealing features with both supervised

and unsupervised methods. Results show that our readability feature set performs satisfactorily.

A crucial point that requires in-depth reflection is the theoretical and empirical distinction between textual dimensions, such as genre and domain. Although textual distinctions have already been analyzed and applied in several settings, in practical terms it is hard to provide clear-cut guidelines. In this respect, we suggest that the automatic discrimination of genres and domains is potentially useful.

Acknowledgements

This research was financed by VINNOVA, Sweden's innovation agency, and SICS East Swedish ICT. We thank SLTC reviewers for useful comments.

References

E. Amigó, J. Gonzalo, J. Artiles, and F. Verdejo. 2009. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information retrieval, 12(4), 461-486*.

R. Bekkerman, K. Eguchi, and J Allan. 2006. *Unsupervised non-topical classification of documents*. Ph.D. thesis, Massachusetts University Amherst.

D. Biber. 1988. *Variation across speech and writing*. Cambridge University Press.

F. Dell'Orletta, S. Montemagni, and G. Venturi. 2014. Assessing document and sentence readability in less resourced languages and across textual genres. *ITL-International Journal of Applied Linguistics,*, 165(2):163–193.

J. Falkenjack, K. Heimann Mühlenbock, and A. Jönsson. 2013. Features indicating readability in Swedish text. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NoDaLiDa-2013), Oslo, Norway*, NEALT Proceedings Series 16.

S. Gustafson-Capková and B. Hartmann, 2006. *Manual of the Stockholm Umeå corpus version 2.0*. Stockholm University.

J. Karlgren and D Cutting. 1994. Recognizing text genres with simple metrics using discriminant analysis. In *Proceedings of the 15th conference on Computational Linguistics-Volume 2*, pages 1071–1075.

B. Kessler, G. Numberg, and H. Schütze. 1997. Automatic detection of text genre. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 32–38.

D.Y. Lee. 2001. Genres, registers, text types, domains and styles: Clarifying the concepts and nevigating a path through the bnc jungle. *Language Learning and Technology 5(3):37–72*.

A. Mehler, S. Sharoff, and M. Santini, editors. 2010. *Genres on the Web: Computational Models and Empirical Studies*, volume 42. Springer.

M. Santini. 2005. Clustering web pages to identify emerging textual patterns. *RECITAL*.

M. Santini. 2008. Zero, single, or multi? genre of web pages through the users' perspective. *Information Processing & Management*, 44(2):702–737.

S. Sharoff. 2007. Classifying web corpora into domain and genre using automatic feature identification. In *Proceedings of the 3rd Web as Corpus Workshop*, pages 83–94.

M. van der Wees, A. Bisazza, W. Weerkamp, and C. Monz. 2015. What's in a domain? analyzing genre and topic differences in statistical machine translation. In *Proceedings of the Joint Conference of the 53rd ACL and the 7th IJCNLP*.

P. Wastholm, A. Kusma, and B. Megyesi. 2005. Using linguistic data for genre classification. In *Proceedings of the Swedish Artificial Intelligence and Learning Systems Event, SAIS-SSLS*.

Appendix. SUC's Genres, Subgenres and Domains

	Top genres	**Subgenres**	**Genre/Domain**
A	**Press,Reportage**		**Genre**
		AA. Political AB. Community AC. Financial AD. Cultural AE. Sports AF. Spot News	
B	**Press, Editorials**		**Genre**
		BA. Institutional BB. Debate articles	
C	**Press, Reviews**		**Genre**
		CA, Books CB. Films CC. Art CD. Theater CE. Music CF. Artists, shows CG. Radio, TV	
E	Skills, trades and hobbies		Domain
		EA. Hobbies, amusements EB. Society press EC. Occupational and trade union press ED. Religion	
F	Popular lore		Domain
		FA. Humanities FB. Behavioural sciences FC. Social sciences FD. Religion FE. Complementary life styles FG. Health and medicine FH. Natural science, technology FJ. Politics FK. Culture	
G	**Biographies, essays**		**Genre**
		GA. Biographies, memoirs GB. Essays	
H	Miscellaneous		Mixed
		HA. Federal publications HB. Municipal publications HC. Financial reports, business HD. Financial reports, non-profit organisations HE. Internal publications, companies HF. University publications	
J	**Learned and scientific writing**		**Genre**
		JA. Humanities JB. Behavioural science JC. Social sciences JD. Religion JE. Technology JF. Mathematics JG. Medicine JH. Natural sciences	
K	**Imaginative prose**		**Genre**
		KK. General fiction KL. Mysteries and science fiction KN. Light reading KR. Humour	