Introducing the Notion of 'Contrast' Features for Language Technology

Marina Santini[1(✉)], Benjamin Danielsson[2], and Arne Jönsson[2,3]

[1] Division ICT-RISE SICS East, RISE Research Institutes of Sweden, Stockholm, Sweden
marina.santini@ri.se
[2] Department of Computer and Information Science, Linköping University, Linköping, Sweden
benda425@student.liu.se, arne.jonsson@liu.se
[3] Division ICT-RISE SICS East, RISE Research Institutes of Sweden, Linköping, Sweden

Abstract. In this paper, we explore whether there exist 'contrast' features that help recognize if a text variety is a genre or a domain. We carry out our experiments on the text varieties that are included in the Swedish national corpus, called Stockholm-Umeå Corpus or SUC, and build several text classification models based on *text complexity features*, *grammatical features*, *bag-of-words features* and *word embeddings*. Results show that text complexity features and grammatical features systematically perform better on genres rather than on domains. This indicates that these features can be used as 'contrast' features because, when in doubt about the nature of a text category, they help bring it to light.

Keywords: Genre · Domain · Supervised classification · Features

1 Introduction

Finding a neat divide across text varieties is a difficult exercise. In the experiments presented in this paper, we focus on two text varieties, i.e. genre and domain. We observe that domains have a topical nature (e.g. Medicine and Sport), while genres have a communicative and textual nature (e.g. academic articles, health records, prescriptions or patient leaflets). A domain normally includes documents belonging to several genres; for instance, genres such as patient leaflets, articles and prescriptions are commonly used in the medical domain. Conversely, individual genres may serve several domains (like academic articles) or can be peculiar to a single domain (like health records). Sometimes, genre and domain are conflated together in domain adaptation, a field associated with parsing, machine translation, text classification and other NLP tasks. However, researchers have recently pointed out that mixing up text varieties has a detrimental effect on the final performance. For instance, some researchers (e.g.

G. Anderst-Kotsis et al. (Eds.): DEXA 2019 Workshops, CCIS 1062, pp. 189–198, 2019.
https://doi.org/10.1007/978-3-030-27684-3_24

[14] and [15]) analyse how topic-based and genre-based categories affect statistical machine translation, and observe that translation quality improves when using separate genre-optimized systems rather than a one-size-fits-all genre-agnostic system.

Since it is not straightforward to sort out genre from domain and vice versa, we investigate whether it is possible to decide automatically if a text variety is a genre or a domain. For instance, is "hobby" a genre or a domain? What about "editorial" or "interview"? In this paper, we explore whether there exist 'contrast' features that help recognize if a text category is a **genre** or a **domain**. By 'contrast' features we refer to those features that consistently perform well (or badly) only on one text variety. We explore the text categories that are included in the Swedish national corpus, called Stockholm-Umeå Corpus or SUC. We build several supervised text classification models based on several feature sets, namely *text complexity features*, *grammatical features*, *bag-of-words (BoW) features* and *word embeddings*.

2 Working Definitions

Before setting out any computational explorations across the text varieties of the SUC, we would like to provide working definitions that help understand the core difference between genre and domain.

Domain is a subject field. Generally speaking, 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 (i.e. open class features like nouns and verbs) and specialized terms (such as "anaemia" in Medicine, or "foul" in Sport).

The concept of **genre** is more abstract than domain. 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"; similarly, a "letter" complies to communicative conventions that are different from the conventions of an "interview". "Academic papers", "tweets", "letters" and "interviews" are examples of genres. Genre conventions usually affect the organization of the document (its rhetorical structure and composition), the length of the text, vocabulary richness, as well as syntax and morphology (e.g. passive forms vs. active forms). In text classification, genres are often represented by features such as Parts-of-Speech (POS) tags, character n-grams, POS n-grams, syntactic tags and function words.

In this complex scenario, how can we assess computationally whether a text category is a genre or domain? We propose using 'contrast' features, as explained in the next sections.

3 Previous Work

The diversified nature of the text varieties of the SUC has proved to be problematic when the corpus has been used in automatic genre classification experiments.

For instance, it has been observed that the SUC's text categories have a dissimilar nature since some of the SUC's 'genres' are in fact subject-oriented [13]. Additionally, researchers have recommended to work with a more complete and uniform set of SUC genres [13]. Previous SUC genre classification models [13] were based on four types of features, i.e. "words, parts-of-speech, parts-of-speech plus subcategories (differentiating between, e.g., personal and interrogative pronouns), and complete Parole word classifications (which include gender, tense, mood, degree, etc.)." Previous results show that grammatical features tend to perform slightly better than word frequencies when taking all the 9 SUC categories into account. The overall performance was rather modest though, showing wide variations across the 9 SUC's text categories.

Interestingly, a similar low performance was reported also when building genre classification models [10] based on discriminant analysis, 20 easy-to-extract grammatical features and 10 or 15 text categories of the Brown corpus [5]. The Brown corpus is the predecessor of the SUC and contains as well a mixtures of text varieties. More encouraging results were achieved when grouping the Brown corpus' text categories into two subgroups and four subgroups [10].

In situations like those described above [10,13], one can surmize that there is probably not enough labeled training data in the datasets to get higher performance. It could also be that the learning algorithm is not well suited to the data, or that a lower error rate is simply not achievable because the predictor attributes available in the data are insufficient to predict the classes more accurately. However, we argue that this is not the case here, since more recent experiments where SUC categories have been sorted out into different varieties give a different picture. For instance, SUC text varieties have been interpreted as mixture of domains, genres, and miscellaneous categories [4]. The reordering of SUC text varieties was proposed as follows: 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 mixed category, i.e. Miscellaneous (H). The findings show that genre classification based on readability features and the whole SUC (9 classes) has a modest performance, while using readability features only for the subset of six 'proper genres' gives much better results.

Text complexity features (another way to refer to readability features) have also shown good performance on SUC's proper genres [9]. More precisely, PCA-based components extracted from text complexity features perform remarkably well on five proper genres, namely Press Reportage (A), Press Editorial (B), Press Review (C), Learning/Scientific Writing (J) and Imaginative Prose (K) [9].

In the experiments described below we offer a more systematic picture of how to use features to decide about the nature of text varieties.

4 The SUC: Corpus and Datasets

The SUC is a collection of Swedish texts (amounting to about one million words) and represents the Swedish language of the 1990's [6]. The SUC follows the general layout of the Brown corpus [5] and the LOB corpus [8], with 500 sample texts

Table 1. Set 1. Experiments with text complexity features and grammatical features.

Set 1	SUC text categories	Features	SMO	DI4jMlp
Exp1	9 SUC varieties (a_reportage_genre, b_editorial_genre, c_review_genre, e_hobby_domain, f_popular_lore_domain, g_bio_essay_genre, h_miscellaneous_mixed, j_scientific_writing_genre, k_imaginative_prose_genre); 1400 instances	115 complexity features	**0,596**	0,582
		65 components	0,567	0,572
		27 POS tags	0,507	0,526
		62 dependency tags	0,541	0,531
Exp2	5 SUC genres (a_reportage_genre, b_editorial_genre, c_review_genre, j_scientific_writing_genre, k_imaginative_prose_genre); 682 instances	115 complexity features	**0,831**	0,813
		64 components	0,829	0,811
		27 POS tags	0,786	0,773
		62 dependency tags	0,782	0,771
Exp3	4 SUC varieties (2 domains and 2 genres: e_hobby_domain, f_popular_lore_domain, j_scientific_writing_genre, k_imaginative_prose_genre); 402 instances	115 complexity features	**0,785**	0,766
		58 components	0,722	0,704
		27 POS tags	0,743	0,740
		62 dependency tags	0,715	0,711
Exp4	2 SUC genres (j_scientific_writing_genre, k_imaginative_prose_genre); 216 instances	115 complexity features	0,981	0,981
		51 components	0,972	0,949
		27 POS tags	**0,986**	0,981
		62 dependency tags	0,981	0,968
Exp5	2 SUC domains (e_hobby_domain, f_popular_lore_domain,); 186 instances	115 complexity features	0,720	**0,749**
		55 components	0,692	0,674
		27 POS tags	0,674	0,706
		Dependency tags	0,707	0,722

with a length of about 2,000 words each. The SUC text varieties are somehow fuzzy although they are collectively called "genres" by the corpus creators [6]. It is worth stressing that the SUC was created to represent the Swedish language as a whole, and not to represent different text varieties. Certainly, this corpus design affects text categorization models.

Technically speaking, the SUC is divided into 1040 bibliographically distinct text chunks, each assigned to so-called "genres" and "subgenres".

The source dataset containing linguistic features was created using the publicly available toolkit named TECST [3] and the text complexity analysis module called SCREAM [7]. The dataset contains 120 variables [2], thereof 115 linguistic features, three readability indices (LIX, OVIX and NR), and two descriptive variables (file name and the language variety). In these experiments, we only used 115 linguistic features.

The BoW dataset was created from the annotated SUC 2.0 corpus, which can be downloaded from SpråkBanken[1] (Gothenburg University).

[1] See https://spraakbanken.gu.se/eng/resources/corpus.

5 Experiments

The experiments in this study are all based on supervised machine learning, which implies that the classification models are trained and built on labelled data. In this case, the labelled data is the SUC's text categories. The general idea behind supervised classification is that once a model proves effective on a set of labelled data (i.e. the performance on the test data is good), then the model can be safely applied to unlabelled data. The requirement is that the unlabelled data on which a supervised classification model is going to be applied has a similar composition and distribution of the data on which the supervised model has been trained. The supervised paradigm differs from unsupervised classification (clustering) where classification models are trained and built on completely unlabelled data. In that case, a human intervention is needed to name and evaluate the resulting clusters.

In this section, we present three sets of experiments and a detailed report of their performance. The first set is based on text complexity- and grammatical features; the second set relies on BoW features; in the third set we present a comparative experiment based on function words and word embeddings. As mentioned above, we apply supervised machine learning and rely on two stable learning models, namely Support Vector Machines (SVM) and Multilayer Perceptron (MLP). We use off-the-shelf implementations of SVM and MLP to ensure full replicability of the experiments presented here. We run the experiments on the Weka workbench [16]. Weka's SVM implementation is called SMO and includes John Platt's sequential minimal optimization algorithm for training a support vector classifier[2]. Weka provides several implementations of MLP. We use the DI4jMlpClassifier, that is a wrapper for the DeepLearning4j library[3], to train a multi-layer perceptron. Both algorithms were run with standard parameters. Results are shown in Tables 1, 2 and 3. We compared the performance on the Weighted Averaged F-Measure (AvgF)[4] and applied 10-folds crossvalidation.

Set 1. The first set contains five experiments (see Table 1). In Experiment 1, models are created with all the 9 SUC text categories (1400 instances) using four different features sets, namely *115 text complexity features*, *PCA-based components*, *POS tags* and *dependency tags*. The two algorithms (SMO and DI4JMlpClassifier), although they have a very different inductive bias, achieve a very similar performance. We observe however that the DI4JMlpClassifier is slower than SMO, which is extremely fast: SMO took no more than a few seconds on all datasets, while the DI4JMlpClassifier's average processing time was about a couple of minutes. In this set, we observe that the performance on the "domain" varieties (i.e. "hobby" and "popular lore") is quite poor. The majority

[2] See http://weka.sourceforge.net/doc.dev/weka/classifiers/functions/SMO.html.
[3] See https://deeplearning.cms.waikato.ac.nz/.
[4] Weighted Averaged F-Measure is the sum of all the classes F-measures, each weighted according to the number of instances with that particular class label. It is a more reliable metric than the harmonic F-measures (F1).

Table 2. Set 2. Experiments with bag-of-words features

Set 2	SUC text categories	BOW features	SMO	DI4jMlp
Exp1	9 SUC varieties (a_reportage_genre, b_editorial_genre, c_review_genre, e_hobby_domain, f_popular_lore_domain, g_bio_essay_genre, h_miscellaneous_mixed, j_scientific_writing_genre, k_imaginative_prose_genre); 1400 instances	Including stopwords	**0,767**	0,640
		Without stopwords	0,741	0,614
Exp2	5 SUC genres (a_reportage_genre, b_editorial_genre, c_review_genre, j_scientific_writing_genre, k_imaginative_prose_genre); 682 instances	Including stopwords	**0,903**	0,854
		Without stopwords	0,863	0,824
Exp3	4 SUC varieties (2 domains and 2 genres: e_hobby_domain, f_popular_lore_domain, j_scientific_writing_genre, k_imaginative_prose_genre); 402 instances	Including stopwords	**0,905**	0,828
		Without stopwords	0,880	0,792
Exp4	2 SUC genres (j_scientific_writing_genre, k_imaginative_prose_genre); 216 instances	Including stopwords	**0,991**	**0,991**
		Without stopwords	**0,991**	**0,991**
Exp5	2 SUC domains (e_hobby_domain, f_popular_lore_domain,); 186 instances	Including stopwords	**0,925**	0,858
		Without stopwords	0,892	0,842

of the texts labelled "hobby" were classified as "reportage" (65 instances). The same thing happened with the majority of the "popular lore" texts. One could surmize that this happens because "reportage" is the most populated class in the dataset. Interestingly, however, a very small class like "bio_essay", that contains two genres, is not attracted towards "reportage". Unsurprisingly, also the "miscellaneous" class has a high number of misclassified texts, notably 42 out of 145 miscellaneous texts have been classified as "reportage", and only 70 texst have been classified with the correct label.

In Experiment 2, we created models with only five single "proper genres". We ditched out the "bio-essay" class because it included two genres that are not necessarily close to each other and because the class was very little populated. All the instances labelled with text varieties other than the five genres were removed, and we ended up with a model built on 682 instances. We observe that the performance increases dramatically (up to AvgF = 0.831 when using 115 complexity features) if compared to the models in Experiment 1.

In Experiment 3, a balanced dataset was created with two domains and two genres. The best performance is achieved by SMO based on 115 complexity

features. We observe that the performance on the two domains (and especially on "popular lore") is definitively lower than the performance on the two genres. This decline could be interpreted as a sign that text complexity features are not representative of topic-based varieties.

In Experiment 4, we created models with two genres only, and we observe that the performance soars up dramatically (up to AvgF = 0.986) when using only 27 POS tags and SMO.

In Experiment 5, we created models with two domains only. The performance is definitely lower, reaching its peak with 115 complexity features in combination with the DI4jMlpClassifier (AvgF = 0.749). Again, the "popular lore" domain suffers from many misclassifications. We interpret the moderate performance on these two subject-based classes as the effect of the inadequacy of the features to represent the nature of domain-based text varieties. From this set of experiments, it appears that text complexity- and grammatical features are genre-revealing features, and their performance on topic-based categories is lower.

Set 2. Also the second set contains five experiments, but this time models are built with BoW features. Two BoW datasets were used: one including stopwords, and the other one without stopwords. Stopwords are normally removed when classifying topic-based categories, while they are helpful for genre classification [14]. A breakdown of the results is shown in Table 2. In Exp1, the best performance is achieved by SMO in combination with BoW+stopwords (AvgF = 0.767). This overall performance is much higher than in Exp1, Set 1. BoW features perform much better on the domains and on the miscellaneous class. The inclusion of stopwords helps genre classification. This is good news for the classification task in itself, but less so for the distinction between genre and domain, that is important in some other NLP tasks, as mentioned earlier. In Exp2, Exp3, Exp4 and Exp5, the best performance is achieved by SMO in combination with BoW features including stopwords. AvgF is very high in all cases (always greater than 0.90), reaching the peak (AvgF = 0.991) on two proper genres.

Set 3. Set 3 contains only a single experiment. In this set of experiments, we compare the contrastive power of two very different feature sets – function words and word embeddings – on the 9 SUC text categories. Function words (a.k.a. stopwords) are grammatical closed classes. They can be used in the form of POS tags (like here) or as word frequencies. They are light-weight features that can be successfully used for genre detection [14], although they are less powerful than other features [12]. Here we used 15 POS tags that represent function words. It took a few seconds to create the models.

Word embeddings are one of the most popular representation of document vocabulary to date. They can be used for many tasks (e.g. sentiment analysis, text classification, etc.). Word embeddings are capable of capturing the context of a word in a document, as well as semantic and syntactic similarity. Here we use Word2Vec word embeddings [11], extracted with the unsupervised Weka filter called DI4jStringToWordVec in combination with the MI4jMLPClassifier (SMO

cannot handle word embeddings). Training a neural networks-based classifier is time consuming. After laborious parameters' tuning, we used a configuration based on three convolutional layers, a global pooling layer, a dense layer and an output layer. It took 5 days to run this configuration with 10-folds crossvalidation. Results (shown in Table 3) are definitely modest on the SUC dataset.

Table 3. Function words (15 POS tags) vs Word2Vec word embeddings

Set 3	SUC text categories	Features	SMO	DI4jMlp
Exp	9 SUC varieties (a_reportage_genre, b_editorial_genre, c_review_genre, e_hobby_domain, f_popular_lore_domain, g_bio_essay_genre, h_miscellaneous_mixed, j_scientific_writing_genre, k_imaginative_prose_genre); 1400 instances	Function Words	0,371	**0,448**
		Word Embeddings	n/a	0.340

Discussion. In the first set of experiments, we explored whether text complexity features (taken individually or aggregated in PCA-based components) and grammatical features have enough contrastive power to disentangle genres and domains. It turns out that these features are more representative of genres than domains and mixed classes since they perform consistently better on genre classes, as neatly shown in all the five experiments in Table 1. It appears that they can be safely used as 'contrast' features. In most cases, the best performance is achieved with 115 text complexity features in combination with SMO. In the second set of experiments, BoW features perform equally well on genres and on domains. The best performance is achieved with BoW features including stopwords. All in all, the classification performance of BoW features on the five experiments is higher than the performance based on text complexity- and grammatical features. However, it is unclear whether the BoW models are generalizable. We speculate that these models somehow overfit the corpus (although we applied 10-folds crossvalidation). For the purpose of our investigation, we observe that BoW features have no or little contrastive power, since their behaviour is rather indistinct across genres and domains. In the third set of experiments, we compared the performance of two very different feature, namely function words and word embeddings. Both feature sets are quite weak. All in all, function words perform better in combination with SMO. We observe that although the overall performance is quite modest, function words are very effective with the "Reportage" genre and the "Imaginative prose" genre, where the number of misclassifications is very limited. Word embeddings are a thorny kind of feature. It has been pointed out that "[d]espite the popularity of word embeddings, the precise way by which they acquire semantic relations between words remain unclear" [1]. Their performance on the SUC is definitely modest.

6 Conclusion and Future Work

In this paper, we argued that text varieties have a diversified nature. We limited our empirical investigation to two text varieties, namely genre and domain. The core of our investigation was the quest of 'contrast' features that could automatically distinguish between genre classes and domain classes. We explored the contrastive power of several feature sets, and reached the conclusion that text complexity features and grammatical features are more suitable as 'contrast' features than BoW features. In particular, text complexity features perform consistently better on genres than on domains. This means that these features can help out when in doubt about the nature of text varieties. Function words and word embeddings seem less suitable as 'contrast' features. A valuable by-product of the empirical study presented here is a comprehensive overview of the performance of different feature sets on the text varieties included in the SUC.

Future work includes the exploration of additional 'contrast' features as well as the application of this approach to other corpora containing mixed text varieties (e.g. the Brown corpus and the British National Corpus).

Acknowledgements. This research was supported by E-care@home, a "SIDUS – Strong Distributed Research Environment" project, funded by the Swedish Knowledge Foundation [kk-stiftelsen, Diarienr: 20140217]. Project website: http://ecareathome.se/

References

1. Altszyler, E., Sigman, M., Slezak, D.F.: Corpus specificity in LSA and word2vec: the role of out-of-domain documents. arXiv preprint arXiv:1712.10054 (2017)
2. Falkenjack, J., Heimann Mühlenbock, K., Jönsson, A.: Features indicating readability in Swedish text. In: Proceedings of the 19th Nordic Conference of Computational Linguistics (NoDaLiDa-2013), No. 085 in NEALT Proceedings Series 16, Oslo, Norway, pp. 27–40. Linköping University Electronic Press (2013)
3. Falkenjack, J., Rennes, E., Fahlborg, D., Johansson, V., Jönsson, A.: Services for text simplification and analysis. In: Proceedings of the 21st Nordic Conference on Computational Linguistics, pp. 309–313 (2017)
4. Falkenjack, J., Santini, M., Jönsson, A.: An exploratory study on genre classification using readability features. In: Proceedings of the Sixth Swedish Language Technology Conference (SLTC 2016), Umeå, Sweden (2016)
5. Francis, W.N., Kucera, H.: Brown Corpus Manual: Manual of Information to Accompany a Standard Corpus of Present-Day Edited American English for Use with Digital Computers. Brown University, Providence (1979)
6. Gustafson-Capková, S., Hartmann, B.: Manual of the Stockholm Umeå Corpus version 2.0. Stockholm University (2006)
7. Heimann Mühlenbock, K.: I see what you mean. Assessing readability for specific target groups. Dissertation, Språkbanken, Department of Swedish, University of Gothenburg (2013). http://hdl.handle.net/2077/32472
8. Johansson, S., Leech, G.N., Goodluck, H.: Manual of information to accompany the Lancaster-Oslo/Bergen Corpus of British English, for use with digital computer. University of Oslo, Department of English (1978)

9. Jönsson, S., Rennes, E., Falkenjack, J., Jönsson, A.: A component based approach to measuring text complexity. In: Proceedings of the Seventh Swedish Language Technology Conference 2018 (SLTC-2018) (2018)
10. Karlgren, J., Cutting, D.: Recognizing text genres with simple metrics using discriminant analysis. In: Proceedings of the 15th Conference on Computational Linguistics, vol. 2, pp. 1071–1075. Association for Computational Linguistics (1994)
11. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
12. Santini, M.: Automatic Identification of Genre in Web Pages: A New Perspective. LAP Lambert Academic Publishing, Saarbrücken (2011)
13. Wastholm, P., Kusma, A., Megyesi, B.: Using linguistic data for genre classification. In: Proceedings of the Swedish Artificial Intelligence and Learning Systems Event, SAIS-SSLS (2005)
14. Van der Wees, M., Bisazza, A., Monz, C.: Evaluation of machine translation performance across multiple genres and languages. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018) (2018)
15. Van der Wees, M., Bisazza, A., Weerkamp, W., Monz, C.: What's in a domain? Analyzing genre and topic differences in statistical machine translation. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), vol. 2, pp. 560–566 (2015)
16. Witten, I.H., Frank, E., Hall, M.A., Pal, C.J.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, Cambridge (2016)