Identifying Predictive Features for Textual Genre Classification: the Key Role of Syntax

Andrea Cimino[◇], **Martijn Wieling**[•],
Felice Dell'Orletta[◇], **Simonetta Montemagni**[◇], **Giulia Venturi**[◇]
[•]University of Groningen - The Netherlands
m.b.wieling@rug.nl
[◇]Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC–CNR)
ItaliaNLP Lab - *www.italianlp.it*
{name.surname}@ilc.cnr.it

Abstract

English. The paper investigates impact and role of different feature types for the specific task of Automatic Genre Classification with the final aim of identifying the most predictive ones. The goal was pursued by carrying out incremental feature selection through Grafting using different sets of linguistic features. Achieved results for discriminating among four traditional textual genres show the key role played by syntactic features, whose impact turned out to vary across genres.

Italiano. *L'articolo intende indagare il ruolo svolto da diversi tipi di caratteristiche linguistiche nella classificazione automatica del genere testuale al fine di identificare le più efficaci e rilevanti. A questo scopo è stata messa a punto una metodologia basata su un processo incrementale di selezione realizzato mediante un algoritmo di Grafting usando diversi tipi di caratteristiche. I risultati raggiunti mostrano il ruolo chiave delle caratteristiche sintattiche, il cui impatto varia in modo significativo tra generi diversi.*

1 Introduction

Automatic classification of textual genres has always received significant attention from both theoretical and application perspectives. On the one hand, it has been considered relevant by linguists and educators to teach students the correct way of writing in specific communicative scenarios (Biber, 1995; Lee, 2001). On the other hand, the classification of textual genres is seen as a way to cope with the well known problem of information overload: the exploitation of information about document genre can help to develop more accurate Information Retrieval tools. Genre identification has been considered a key factor for reducing irrelevant results of search engines, as users would be able to specify the desired textual genre along with the keywords expressing the content they are looking for (Santini, 2004; Lim, 2004; Santini, 2007). In fact, document genre and document content represent orthogonal dimensions of classifications (Finn, 2003).

A variety of different approaches to Automatic Genre Classification (AGC) has been proposed so far differing at the level of the *genre* and the *typology of features* considered. According to the widely acknowledged fact that no established classification of *genres* exists (see e.g. Sharoff (2010) or Biber (2009)), previous studies focused on 'traditional genres' such as journalism, handbooks, academic prose, see among the others (Kessler, 1997; Stamatatos, 2001; Fang, 2010), and on 'web genres', i.e. genres of web pages, see e.g. (Santini, 2004; Lim, 2004; Mehler, 2010).

Despite the great interest in the investigation of which linguistic features qualify a text genre (Biber, 2009; Fang, 2015), so far little effort has been devoted to use sophisticated NLP techniques, such as syntactic parsing, to capture complex linguistic features for the automatic classification of textual genres. Differently from other application scenarios where the *form* (the style) of a document is investigated, such as e.g. Authorship Attribution (Cranenburgh, 2012), Readability Assessment (Collins, 2014) and Native Language Identification (Tetreault, 2013), AGC approaches proposed so far mainly focus on word level linguistic features, in particular the distribution of function words, word frequency, n–gram models of both characters and Parts–Of-Speech (Santini, 2004; Crossley, 2007; Mehler, 2010) or finer-grained Parts–Of-Speech tags including morpho-syntactic features such as verb tense (Fang, 2010). Very

few studies rely on features extracted from syntactically annotated texts, the exception being Stamatatos (2001) who combines lexical features (i.e. word frequency) with features extracted from the output of a chunk boundary detector (e.g. the distribution of noun, verbal, adjectival phrases), the average number of words included in verbal phrases. Similar structural features have been also used by (Lim, 2004) who combined web–specific features (e.g. HTML tags) with lexical information and features aiming at capturing the syntactic structure of a sentence, e.g. the distribution of declarative and imperative sentences, syntactic ambiguities, etc.

In this paper, we tackle the AGC task for traditional genres (namely literary, scientific, educational and journalistic texts) by using different types of linguistic features, i.e. lexical, morpho-syntactic and syntactic. In particular, the following research questions are addressed: i) which are the most effective features to classify a textual genre, and ii) whether and to what extent features identified as most effective remain the same across different genres. These questions have been addressed by carrying out incremental feature selection with the final aim of identifying the most predictive ones. So far, studies focused on the *best set* of features to classify textual genres have been carried out mainly on English. In this paper, this issue is investigated for a typologically different language, Italian.

2 Model training and feature ranking

In order to identify and rank the most important features playing a role in genre classification, we used GRAFTING (Perkins, 2003). This approach allows us to simultaneously train a maximum entropy model while also including incremental feature selection. Grafting uses a gradient-based heuristic to select the most promising feature (which is added to the set of selected features S), and subsequently performs a full weight optimization over all features in S. This process is repeated until a certain stopping condition is reached. The stopping condition integrates l_1 regularization in the grafting approach. This means that only those features are included (with a non-zero weight) if the l_1 penalty is outweighed by the reduction of the objective function. Consequently, overfitting is prevented by excluding noisy features, or those that change value infrequently. In our case, the l_1 penalty was selected on the basis of evaluating maximum entropy models (using 10-fold cross validation) using varying l_1 values (range: 1e-11, 1e-10, ..., 0.1, 1).

For selecting the features and estimating their weights, we used TINYEST[1], a grafting-capable maximum entropy parameter estimator for ranking tasks (De Kok, 2011; De Kok, 2013). Even though our task is not a ranking task, it can be used for binary classification by assigning a high score (1) to the correct class and a low score (0) to the incorrect class. A similar approach was followed by Dell'Orletta (2014) for discriminating between easy–to–read vs difficult–to–read sentences. As the focus of the present study is on the classification of texts belonging to different traditional genres, we created four separate binary classifiers which were trained to distinguish Literature texts from non-Literature (i.e. the three remaining genres) texts, Educational texts from non-Educational texts, etc. A text was assigned the class of the classifier which returned the highest score.

3 Typology of Features

Various types of features have been proposed in the literature for the automatic classification of text genres. Following Stamatatos (2001) and Lim (2004), we combine token–based and structural features. Token–based features were extracted from the top list of the most frequent lemmata in the training corpus and represented in terms of the relative frequency of each lemma in each document. Structural features were extracted from the considered corpora morpho–syntactically tagged by the POS tagger described in (Dell'Orletta, 2009) and dependency–parsed by the DeSR parser using Multi–Layer Perceptron (Attardi, 2009). As shown in Table 1, they range across different linguistic description levels (lexical, morpho–syntactic and syntactic) for a total of 90 features that resulted to be informative "fingerprints" of the form of a text, on issues of e.g. genre, style, authorship or readability.

4 Experiments

4.1 Experimental Setup

We used an Italian corpus including documents representative of four different genres: educational material (Dell'Orletta, 2011), newspaper ar-

[1] http://github.com/danieldk/tinyest

ticles (Marinelli, 2003), literary texts (Marinelli, 2003) and scientific papers (Dell'Orletta, 2014). The whole corpus was split up into a training set (136 documents for the Education genre, 579 for the Journalism genre, 365 for the Literature genre and 317 for the Scientific genre), and a held-out test set (60 documents for each genre).

To assess the influence of including structural features over simply using the most frequent words (lemmata), we used two sets of features (each consisting of about 200 features). The first set of features (taken as the baseline) corresponds to the relative frequency of the 200 top-most frequent words (henceforth referred to as the `tw200` set). [2] The second set combines token-based and structural features: i.e. in addition to the relative frequency of the 100 top-most frequent words, it contains the (90) structural features illustrated above and detailed in Table 1 (this set is henceforth referred to as the `lingtw` set). To guarantee comparability of values, for each feature the values were scaled between 0 and 1 on the basis of the data from the training set. If a (non-scaled) feature value in the held-out test set exceeded the maximum non-scaled value of that feature in the training set, it was set to the maximum value (1).

The feature ranking for each genre was obtained using grafting on the full training data set. The performance (i.e. the percentage of correctly classified documents) of the algorithm was evaluated for an increasing number of features (starting from including only the first (best) feature for each genre to including all features for each genre) against both a 10-fold cross-validation test set and a held-out test set.

The 10-fold cross-validation procedure was performed on the basis of the training set (i.e. the feature weights were determined on the basis of 90% of the training data, whereas the performance was evaluated on the remaining 10% of the training data; this procedure was repeated 10 times). As stated before, the genre of the document in the test set was assigned to the genre whose binary classification model (in this case with the same number of features) resulted in the highest score.

The classification accuracy was assessed with respect to the held-out test set for different numbers of features: *i)* the number of features associated with the best performance on the cross validation set, and *ii)* the lowest number of features such that the performance dropped when a new feature was added (i.e. performance kept increasing for each additional feature up to the selected number of features).

[2] In our preliminary analyses, we also assessed the effect of including the most frequent bigrams as features. However, as the performance was similar to only using unigrams, we did not include bigrams as features.

Typology	Feature
Raw Text	Sentence and token length
Lexical	Rate of words in the Basic Italian Vocabulary, Type/Token ratio
Morpho-syntactic	Part-Of-Speech unigrams, Lexical density, Verbal mood
Syntactic	Dependency type unigrams, Parse tree depth features, Arity of verbal predicates, Distribution of subordinate vs main clauses, Length of dependency links

Table 1: Typology of features automatically extracted from linguistically annotated texts.

4.2 Replication

Results reported below can be replicated by downloading the docker image `italianlp-wieling/dockergenreclassification` which contains all data and scripts necessary for the feature extraction and the grafting procedure, and also contains all results. The Docker file including all commands to setup the virtual machine can be found at https://github.com/italianlp-wieling/dockergenreclassification.

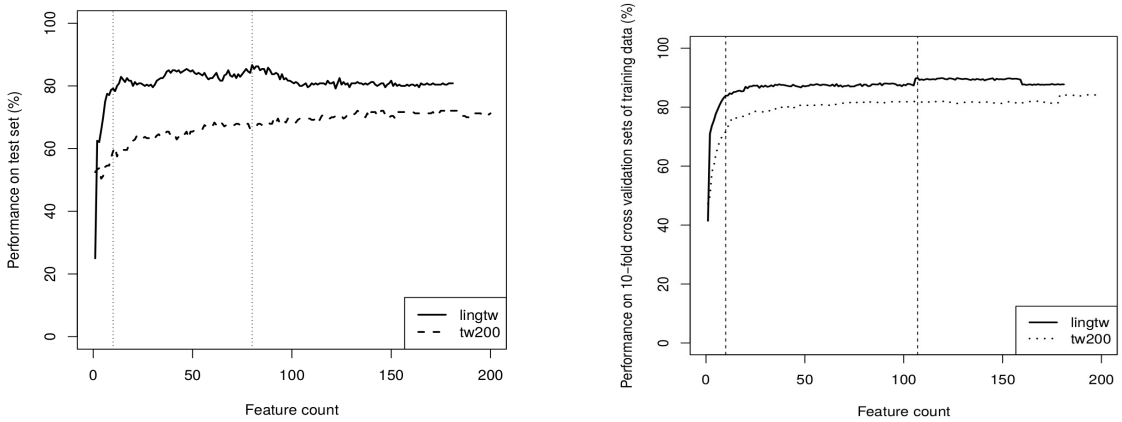
5 Results

5.1 Genre Classification Results

Figures 1(a) and 1(b) report the classification results using the `lingtw` vs `tw200` features sets: it can be clearly observed that inclusion of structural features is highly beneficial. With only 10 features, the 10-fold cross validation performance is 83.89% for the `lingtw` set, whereas it is only 71.51% for the `tw200` set. The optimal performance for the `lingtw` set is reached with 106 features (89.72%), whereas a significantly lower performance is reached using the `tw200` set (84.17%) despite the much higher number of features used (179). The performance on the held-out test set turned out to be slightly lower: 79.16% for 10 features using the `lingtw` set, against 59.16% with the `tw200` set. The optimal performance for the `lingtw` set is reached with 80 features (86.66%), whereas for the `tw200` set a lower per-

Genre	First 10 features					First 50 features					First 80 features				
	S	**P**	**L**	**R**	**W**	**S**	**P**	**L**	**R**	**W**	**S**	**P**	**L**	**R**	**W**
Journalism	30	40	30	0	0	40	38	10	0	12	38.75	28.75	6.25	1.25	25
Literature	40	30	20	0	10	34	28	8	2	28	33.75	26.25	6.25	1.25	32.50
Education	50	20	10	20	0	44	32	6	4	14	42.5	30	5	2.5	20
Science	60	10	20	10	0	50	32	10	4	4	42.5	30	6.25	2.5	18.75

Table 2: Percentage distribution of different typologies of ranked syntactic (**S**), morpho-syntactic (**P**), lexical (**L**), raw-text (**R**) and token-based (**W**) features selected via GRAFTING on the held-out test set.



(a) Held-out test set.

(b) 10-fold cross-validation test set.

Figure 1: Genre classification results using a held-out test set (a), and a 10-fold cross-validation procedure (b).

formance (72.08%) is obtained using 133 features.

5.2 Feature Ranking Results

In order to investigate the typology of linguistic features most significantly contributing to AGC we focused on the `lingtw` set. In particular, we carried out an in-depth analysis of the grafting-based feature ranking resulting from the classification of the held-out test set. Ranked features were categorized into five classes: syntactic, morpho-syntactic, lexical, raw and token-based features. Figure 2 provides a genre-independent view reporting the percentage average distribution (across genres) of different feature types within the first 10, 50 and 80 ranked feature sets. As shown, *syntactic* features play the most relevant role. They cover the 45% and 42% of the first 10 and 50 features respectively, and remain the most predictive ones also when 80 features are considered (representing 39.38% of the set). On the other hand, the distribution of token-based features increases as far as a wider amount of ranked features is considered (they cover 2.5%, 14.5% and 24.06% in the 10, 50 and 80 feature sets respectively).

Consider now the distribution of different types of features across genres reported in Table 2: notable differences can be observed. In particular, *Literature* and *Scientific prose* represent two opposite poles. Token-based features (*W*) are more predictive for literary texts with respect to other genres (i.e. they represent 10%, 28% and 32.50% in the top 10, 50 and 80 features respectively). On the contrary, syntactic features (*S*) play for *Scientific prose* a more important role than for the other genres (covering respectively 60%, 50% and 42.50% of the top 10, 50 and 80 features).

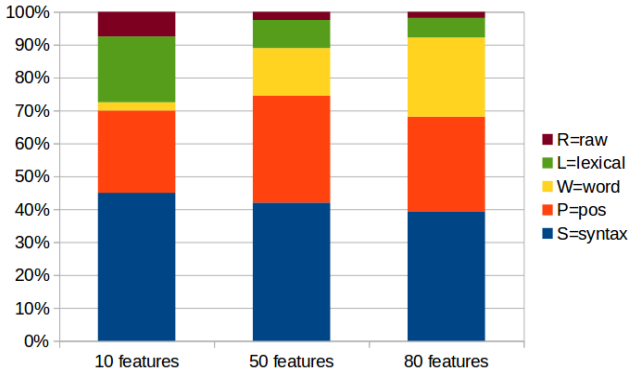


Figure 2: Genre-independent average distribution of different feature types in the top 10, 50 and 80 ranked sets.

	Journalism	Literature	Education	Science
Sentence length	–	82	6	32
Word length	56	50	3	3
Type/Token Ratio (forms)	3	95	94	4
Parse tree depth	11	3	37	94
Maximum length of dependency links	42	24	48	56
Post-verbal subject	16	22	90	76
Pre-verbal object	31	21	47	42
Passive subject	7	53	17	5

Table 3: Different ranking positions of a selection of features across genres. Features which were not selected during ranking have no specified rank in the table.

Let's focus now on the role played by individual features across genres. Table 3 reports the different rank positions associated with a selection of features in the classification of the four genres. Raw text features (i.e. sentence and word length) resulted to play a key role in the classification of educational materials (*Education*) with respect to the other genres (e.g. *Literature*). A feature capturing the lexical richness of texts such as Type/Token Ratio (TTR), which refers to the ratio between the number of lexical types and the number of tokens (considered as single forms) within a text, is similarly ranked for *Journalism* and *Science* while it plays a less relevant role in the classification of educational material and literary texts. Moving to syntax, it should be noted that two features characterizing the overall sentence structure, i.e. the depth of the whole parse tree (calculated in terms of the longest path from the root of the dependency tree to some leaf) and the maximum length of dependency links (calculated in terms of the words occurring between the syntactic head and the dependent), play a key role in the classification of the *Literature* and *Journalism* genres. For the latter, it is interesting to contrast the high rank associated with the parse tree depth feature and the irrelevant role played by sentence length (typically taken as a proxy of the underlying grammatical structure): this clearly shows that syntactic features are more effective in discriminating genres. Other features which turned out to play a relevant role in ACG are concerned with the relative ordering of subject and object with respect to the verbal head: their non-canonical orders, i.e. post-verbal subject and pre-verbal object, play a key role in the classification of *Literature* and *Journalism* genres. On the contrary, the use of passive voice (inferred from the presence of passive subjects) is less relevant for the classification of *Literature*, whereas it is highly ranked in the characterization of scientific writing and newspaper articles.

6 Conclusion

In this paper we investigated impact and role of different feature types for Automatic Genre Classification. The goal was pursued by carrying out incremental feature selection through Grafting augmented with TinyEst. Two sets of features were taken into account, token-based and structure-based. Achieved results show the key role played by syntactic features, a result which is new with respect to the AGC literature. Another original contribution is concerned with the role of different feature types which turned out to vary across textual genres, suggesting the specialization of features in binary genre classification tasks (e.g. Literature vs. other genres). The features contributing to AGC for Italian are possibly influenced by the language dealt with. Although it is widely acknowledged that linguistic variation across genres is a language universal, the question is whether similar linguistic features are expected to play a similar role across languages. If this might be the case of features such as e.g. TTR, use of passive voice, tenses or pronouns, on the other hand features concerned with the ordering of sentence constituents or the overall sentence structure (e.g. parse tree depth or dependency length) may be distinctive to a specific language or language family. Further directions of research thus include comparison of results in a multilingual perspective as well as across a wider variety of genres.

Acknowledgements

Reported research started in the framework of a Short Term Mobility program of international exchanges funded by CNR, and continued within the project "Smart News, Social sensing for breakingnews", funded by the Tuscany Region under the FAR-FAS 2014 program.

References

Giuseppe Attardi, Felice Dell'Orletta, Maria Simi and Joseph Turian 2009. Accurate Dependency Parsing with a Stacked Multilayer Perceptron. *Proceedings of Evalita 2009*.

Douglas Biber. 1995. Dimensions of register variation: A cross-linguistic comparison. *Cambridge University Press* Press, Cambridge, UK.

Douglas Biber and Susan Conrad 2009. Genre, Register, Style. *Cambridge University Press*

Andreas van Cranenburgh 2012. Literary authorship attribution with phrase-structure fragments. *Proceedings of the ACL Workshop on Computational Linguistics for Literature*, 59–63

Kevin Collins-Thompson 2014. Computational assessment of text readability: a survey of current and future research. *Recent Advances in Automatic Readability Assessment and Text Simplification. Special issue of International Journal of Applied Linguistics*, (165–2), 97–135

Scott A. Crossley and Max Louwerse. 2007. Multidimensional register classification using bigrams. *International Journal of Corpus Linguistics)*, (12–4) 453–478

Daniël de Kok 2011. Discriminative features in reversible stochastic attribute-value grammars. *Proceedings of the EMNLP Workshop on Language Generation and Evaluation*, 54–63

Daniël de Kok 2013. Reversible Stochastic Attribute-Value Grammars. Rijksuniversiteit Groningen.

Felice Dell'Orletta 2009. Ensemble system for Part-of-Speech tagging. *Proceedings of Evalita'09, Evaluation of NLP and Speech Tools for Italian*.

Felice Dell'Orletta, Simonetta Montemagni, Eva Maria Vecchi and Giulia Venturi. 2011. Tecnologie linguistico-computazionali per il monitoraggio della competenza linguistica italiana degli alunni stranieri nella scuola primaria e secondaria. *Percorsi migranti: uomini, diritto, lavoro, linguaggi*, 319–366

Felice Dell'Orletta, Simonetta Montemagni and Giulia Venturi 2014. Assessing document and sentence readability in less resourced languages and across textual genres. *International Journal of Applied Linguistics*, 165:2, 319–366

Felice Dell'Orletta, Martijn Wieling, Andrea Cimino, Giulia Venturi and Simonetta Montemagni 2014. Assessing the Readability of Sentences: Which Corpora and Features? *Proceedings of 9th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2014)*, 163–173

Alex Chengyu Fang, and Jing Cao. 2010. Enhanced Genre Classification through Linguistically Fine-Grained POS Tag. *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation (PACLIC)*, 223–232

Chengyu Alex Fang, and Jing Cao 2015. Text Genres and Registers: The Computation of Linguistic Features. Springer.

Aidan Finn and Nicholas Kushmerick 2003. Learning to classify documents according to genre. *Proceedings of the IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, 1–26

Brett Kessler, Geoffrey Nunberg, and Hinrich Schütze 1997. Automatic detection of text genre. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Chapter of ACL (ACL/EACL'97)*, 223–232

David Lee. 2001. Genres, registers, text types, domains, and styles: clarifyng the concepts and navigating a path through the BNC jungle. *Proceedings of ECIR 2004 (26th European Conference on IR Research)*, University of Sunderland (UK), (3), 37–72

Chul Lim, Kong Lee, and Gil Kim. 2004. Multiple sets of features for automatic genre classification of web documents. *Information processing and management* (41) 1263–1276

R. Marinelli, L. Biagini, R. Bindi, S. Goggi, M. Monachini, P. Orsolini, E. Picchi, S. Rossi, N. Calzolari and A. Zampolli. 2003. The italian parole corpus: an overview. *Computational Linguistics in Pisa, Special Issue, XVI-XVII*, 401-421

Alexander Mehler, Serge Sharoff and Marina Santini (Eds.) 2010. Genres on the Web. *Springer Series - Text, Speech and Language Technology*

Simon Perkins, Kevin Lacker and James Theiler 2003. Grafting: Fast, incremental feature selection by gradient descent in function space. *The Journal of Machine Learning Researchs*, (3), 1333–1356

Marina Santini. 2004. Identification of Genres on the Web: a Multi–Faceted Approach. *Language Learning and Technology*, 1–8

Maria Santini 2007. Enhanced Genre Classification through Linguistically Fine-Grained POS Tag. *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation (PACLIC)*, 223–232

Serge Sharoff 2010. In the garden and in the jungle: Comparing genres in the BNC and internet. in Mehler (2010), 149–166

Efstathios Stamatatos, Nikos Fakotakis and George Kokkinakis 2001. Automatic text categorization in terms of genre and author. *Computational Linguistics*, (26) 471–495

Joel Tetreault, Daniel Blanchard and Aoife Cahill 2013. A Report on the First Native Language Identification Shared Task *Proceedings of the ACL Workshop on Innovative Use of NLP for Building Educational Applications*, 48–57