On the Relevance of Syntactic and Discourse Features for Author Profiling and Identification

Juan Soler-Company
UPF
Carrer de Roc Boronat 138
Barcelona, 08018, Spain
juan.soler@upf.edu

Leo Wanner
UPF and ICREA
Carrer de Roc Boronat 138
Barcelona, 08018, Spain
leo.wanner@upf.edu

Abstract

The majority of approaches to author profiling and author identification focus mainly on lexical features, i.e., on the content of a text. We argue that syntactic dependency and discourse features play a significantly more prominent role than they were given in the past. We show that they achieve state-of-the-art performance in author and gender identification on a literary corpus while keeping the feature set small: the used feature set is composed of only 188 features and still outperforms the winner of the PAN 2014 shared task on author verification in the literary genre.

1 Introduction

Author profiling and author identification are two tasks in the context of the automatic derivation of author-related information from textual material. In the case of author profiling, demographic author information such as gender or age is to be derived; in the case of author identification, the goal is to predict the author of a text, selected from a pool of potential candidates. The basic assumption underlying author profiling is that, as a result of being exposed to similar influences, authors who share demographic traits also share linguistic patterns in their writings. The assumption underlying author identification is that the writing style of an author is unique enough to be characterized accurately and to be distinguishable from the style of other authors. State-of-the-art approaches commonly use large amounts of lexical features to address both tasks. We show that with a small number of features, most of them syntactic or discourse-based, we outperform the best models in the PAN 2014 author verification shared task (Stamatatos et al., 2014) on a literary genre dataset and achieve state-of-the-art performance in author and gender identification on a different literary corpus.

In the next section, we briefly review the related work. In Section 3, we describe the experimental setup and the features that are used in the experiments. Section 4 presents the experiments and their discussion. Finally, in Section 5, we draw some conclusions and sketch the future line of our research in this area.

2 Related Work

Author identification in the context of the literary genre attracted attention beyond the NLP research circles, e.g., due to the work by Aljumily (2015), who addressed the allegations that Shakespeare did not write some of his best plays using clustering techniques with function word frequency, word n-grams and character n-grams. Another example of this type of work is (Gamon, 2004), where the author classifies the writings of the Brontë sisters using as features the sentence length, number of nominal/adjectival/adverbial phrases, function word frequencies, part-of-speech (PoS) trigrams, constituency patterns, semantic information and n-gram frequencies. In the field of author profiling, several works addressed specifically gender identification. Schler et al. (2006), Koppel et al. (2002) extract function words, PoS and the 1000 words that have more information gain. Sarawgi et al. (2011) use long-distance syntactic patterns based on probabilistic context-free grammars, token-level language models and character-level language models.

In what follows, we focus on the identification of the author profiling trait 'gender' and on author identification as such. For both, feature engineering is crucial and for both the tendency is to use word/character n-grams and/or function

Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 681–687,
Valencia, Spain, April 3-7, 2017. ©2017 Association for Computational Linguistics

and stop word frequencies (Mosteller and Wallace, 1963; Aljumily, 2015; Gamon, 2004; Argamon et al., 2009), PoS tags (Koppel et al., 2002; Mukherjee and Liu, 2010), or patterns captured by context-free-grammar-derived linguistic patterns; see e.g. (Raghavan et al., 2010; Sarawgi et al., 2011; Gamon, 2004). When syntactic features are mentioned, often function words and punctuation marks are meant; see e.g. (Amuchi et al., 2012; Abbasi and Chen, 2005; Cheng et al., 2009). However, it is well-known from linguistics and philology that deeper syntactic features, such as sentence structure, the frequency of specific phrasal, and syntactic dependency patterns, and discourse structure are relevant characteristics of the writing style of an author (Crystal and Davy, 1969; DiMarco and Hirst, 1993; Burstein et al., 2003).

3 Experimental Setup

State-of-the-art techniques for author profiling / identification usually draw upon large quantities of features; e.g., Burger et al. (2011) use more than 15 million features and Argamon et al. (2009) and Mukherjee and Liu (2010) more than 1,000. This limits their application in practice. Our goal is to demonstrate that the use of syntactic dependency and discourse features allows us to minimize the total number of features to less than 200 and still achieve competitive performance with a standard classification technique. For this purpose, we use Support Vector Machines (SVMs) with a linear kernel in four different experiments. Let us introduce now these features and the data on which the trained models have been tested.

3.1 Feature Set

We extracted 188 surface-oriented, syntactic dependency, and discourse structure features for our experiments. The surface-oriented features are few since syntactic and discourse structure features are assumed to reflect better than surface-oriented features the unconscious stylistic choices of the authors.

For feature extraction, Python and its natural language toolkit, a dependency parser (Bohnet, 2010), and a discourse parser (Surdeanu et al., 2015) are used.

The feature set is composed of six subgroups of features:

Character-based features are composed of the ratios between upper case characters, periods, commas, parentheses, exclamations, colons, number digits, semicolons, hyphens and quotation marks and the total number of characters in a text.

Word-based features are composed of the mean number of characters per word, vocabulary richness, acronyms, stopwords, first person pronouns, usage of words composed by two or three characters, standard deviation of word length and the difference between the longest and shortest words.

Sentence-based features are composed of the mean number of words per sentence, standard deviation of words per sentence and the difference between the maximum and minimum number of words per sentence in a text.

Dictionary-based features consist of the ratios of discourse markers, interjections, abbreviations, curse words, and polar words (positive and negative words in the polarity dictionaries described in (Hu and Liu, 2004)) with respect to the total number of words in a text.

Syntactic features Three types of syntactic features are distinguished:

1. *Part-of-Speech features* are given by the relative frequency of each PoS tag[1] in a text, the relative frequency of comparative/superlative adjectives and adverbs and the relative frequency of the present and past tenses. In addition to the fine-grained Penn Treebank tags, we introduce general grammatical categories (such as 'verb', 'noun', etc.) and calculate their frequencies.

2. *Dependency features* reflect the occurrence of syntactic dependency relations in the dependency trees of the text. The dependency tagset used by the parser is described in (Surdeanu et al., 2008). We extract the frequency of each individual dependency relation per sentence, the percentage of modifier relations used per tree, the frequency of adverbial dependencies (they give information on manner, direction, purpose, etc.), the ratio of modal verbs with respect to the total number of verbs, and the percentage of verbs that appear in complex tenses referred to as "verb chains" (VCs).

3. *Tree features* measure the tree width, the tree depth and the ramification factor. Tree depth is defined as the maximum number of nodes between the root and a leaf node; the width is the maximum number of siblings at any of the levels of the tree; and the ramification factor is the mean num-

[1] We use the Penn Treebank tagset `http://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html`

ber of children per level. In other words, the tree features characterize the complexity of the dependency structure of the sentences.

These measures are also applied to subordinate and coordinate clauses.

Discourse features characterize the discourse structure of a text. To obtain the discourse structure, we use Surdeanu et al. (2015)'s discourse parser, which receives as input a raw text, divides it into *Elementary Discourse Units* (EDUs) and links them via discourse relations that follow the Rhetorical Structure Theory (Mann and Thompson, 1988).

We compute the frequency of each discourse relation per EDU (dividing the number of occurrences of each discourse relation by the number of EDUs per text) and additionally take into account the shape of the discourse trees by extracting their depth, width and ramification factor.

3.2 Datasets

We use two datasets. The first dataset is a corpus of chapters (henceforth, referred to as "LiteraryDataset") extracted from novels downloaded from the "Project Gutenberg" website[2]. Novels from 18 different authors were selected. Three novels per author were downloaded and divided by chapter, labeled by the gender and name of the author, as well as by the book they correspond to. All of the authors are British and lived in roughly the same time period. Half of the authors are male and half female[3]. The dataset is composed of 1793 instances.

The second dataset is publicly available[4] and was used in 2014's PAN author verification task (Stamatatos et al., 2014). It contains groups of literary texts that are written by the same author and a text whose author is unknown (henceforth, "PANLiterary").

3.3 Experiments

As already mentioned above, we carried out four experiments; the first three of them on the LiteraryDataset, and the last one on the PANLiterary dataset. The LiteraryDataset experiments targeted gender identification, author identification, and identification to which of the 54 books a given chapter belongs, respectively. The PANLiterary experiment dealt with author verification, analogously to the corresponding PAN 2014 shared task.

Used Features	Accuracy Gen	Accuracy Auth
Complete Set	**90.18%**	**88.34%**
Char (C)	67.65%	37.76%
Word (W)	61.79%	38.54%
Sent (S)	60.35%	17.12%
Dict (Dt)	60.62%	17.90%
Discourse (Dc)	69.99%	42.61%
Syntactic (Sy)	88.94%	82.82%
C+W+S+Dt+Dc	80.76%	69.72%
C+W+S+Dt+Sy	89.96%	87.17%
Sy+Dc	89.35%	83.88%
C+W+S+Dt	73.89%	42.55%
MajClassBaseline	53.54%	9.93%
2GramBaseline	79.25%	75.24%
3GramBaseline	75.53%	62.63%
4GramBaseline	72.39%	39.65%
5GramBaseline	65.81%	26.94%

Table 1: Results of the Gender and Author Identification Experiments

4 Experiment Results and Discussion

4.1 Gender Identification

The gender identification experiment is casted as a supervised binary classification problem. Table 1 shows in the column 'Accuracy Gen' the performance of the SVM with each feature group separately as well as with the full set and with some feature combinations. The performance of the majority class classifier (MajClassBaseline) and of four different baselines, where the 300 most frequent token n-grams (2–5 grams were considered) are used as classification features, are also shown for comparison.

The n-gram baselines outperform the SVM trained on any individual feature group, except the syntactic features, which means that syntactic features are crucial for the characterization of the writing style of both genders. Using only this group of features, the model obtains an accuracy of 88.94%, which is very close to its performance with the complete feature set. When discourse features are added, the accuracy further increases.

[2] https://www.gutenberg.org/

[3] The 18 selected authors are: Virginia Woolf, Arthur Conan Doyle, Anne Brontë, Charlotte Brontë, Lewis Carroll, Agatha Christie, William Makepeace Thackeray, Oscar Wilde, Maria Edgeworth, Elisabeth Gaskell, Bram Stoker, James Joyce, Jane Austen, Charles Dickens, H.G Wells, Robert Louis Stevenson, Mary Anne Evans (known as George Eliot) and Margaret Oliphant.

[4] http://pan.webis.de/clef14/pan14-web/author-identification.html

4.2 Author Identification

The second experiment classifies the texts from the LiteraryDataset by their authors. It is a 18-class classification problem, which is considerably more challenging. Table 1 (column 'Accuracy Auth') shows the performance of our model with 10-fold cross-validation when using the full set of features and different feature combinations.

The results of the 10-fold author identification experiment show that syntactic dependency features are also the most effective for the characterization of the writing style of the authors. The model with the full set of features obtains 88.34% accuracy, which outperforms the n-gram baselines. The high accuracy of syntactic dependency features compared to other sets of features proves again that dependency syntax is a very powerful profiling tool that has not been used to its full potential in the field.

Analyzing the confusion matrix of the experiment, some interesting conclusions can be drawn; due to the lack of space, let us focus on only a few of them. For instance, the novels by Elisabeth Gaskell are confused with the novels by Mary Anne Evans, Jane Austen and Margaret Oliphant. This is likely because not only do all of these authors share the gender, but Austen is also considered to be one of the main influencers of Gaskell. Even though, Agatha Christie is predicted correctly most of the times, when she is confused with another author, it is with Arthur Conan Doyle. This may not be surprising since Arthur Conan Doyle and, more specifically, the books about Sherlock Holmes, greatly influenced her writing, resulting in many detective novels with Detective Poirot as protagonist (Christie's personification of Sherlock Holmes). Other mispredictions (such as the confusion of Bram Stoker with Elisabeth Gaskell) require a deeper analysis and possibly also highlight the need for more training material.

4.3 Source Book Identification

To further prove the profiling potential of syntactic and discourse features, we carried out an additional experiment. The goal was to identify from which of the 54 books a given chapter is, making use of syntactic and discourse features only. Using the same method and 10-fold cross-validation, 83.01% of accuracy was achieved. The interesting part of this experiment is the error analysis. "Silas Marner", written by Mary Anne Evans (known as George Elliot), is one of the books that created the highest confusion; it is often confused with "Mill on the Floss" written by the same author. "Kidnapped" by Robert Louis Stevenson, which is very different from the other considered books by the same author, is confused with "Treasure Island" also by Stevenson, and "Great Expectations" by Charles Dickens. "Pride and Prejudice" by Jane Austen is confused with "Sense and Sensibility" also by her. The majority of confusions are between books by the same author, which proves our point further: syntactic and discourse structures constitute very powerful, underused profiling features (recall that for this experiment, we used only syntactic and discourse features; none of the features was content- or surface-oriented). When the full set of features was used, the accuracy improved to 91.41%. In that case, the main sources of confusion were between "Agnes Grey" and "The Tenant of Wildfell Hall", both by Anne Brontë and between "Silas Marner" and "Mill on the Floss", both by G. Elliot.

4.4 PAN Author Verification

The literary dataset in the PAN 2014 shared task on author verification contains pairs of text instances where one text is written by a specific author and the goal is to determine whether the other instance is also written by the same author. Note that the task of author verification is different from the task of author identification. To apply our model in this context, we compute the feature values for each pair of known-anonymous instances and substract the feature values of the known instance from the features of the anonymous one; the feature values are normalized. As a result, a feature difference vector for each pair is computed. The vector is labeled so as to indicate whether both instances were written by the same author or not.

The task performance measure is computed by multiplying the area under the ROC curve (AUC) and the "c@1" score, which is a metric that takes into account unpredicted instances. In our case, the classifier outputs a prediction for each test instance, such that the c@1 score is equivalent to accuracy. In Table 2, the performance of our model, compared to the winner and second ranked of the English literary text section of the shared task (cf. (Modaresi and Gross, 2014) and (Zamani et al., 2014) for details), is shown.

Our model outperforms the task baseline as well

Approach	Final Score	AUC	c@1
Our Model	**0.671**	**0.866**	**0.775**
Modaressi & Gross	0.508	0.711	0.715
Zamani et al.	0.476	0.733	0.650
META-CLASSIFIER	0.472	0.732	0.645
BASELINE	0.202	0.453	0.445

Table 2: Performance of our model compared to other participants on the "PANLiterary" dataset

as the best performing approach of the shared task, the META-CLASSIFIER (MC), by a large margin. The task baseline is the best-performing language-independent approach of the PAN-2013 shared task. MC is an ensemble of all systems that participated in the task in that it uses for its decision the averaged probability scores of all of them.

4.5 Feature Analysis

Table 3 displays the 20 features with the highest information gain, ordered top-down (upper being the highest) for each of the presented experiments.[5] Syntactic features prove again to be relevant in all the experiments. The table shows that there are features that work well for the majority of the experiments. This includes, e.g., the usage of verb chains (VC), syntactic objects (OBJ), commas, predicative complements of control verbs (OPRD), or adjective modifiers (AMOD). It is interesting to note that the Elaboration discourse relation is distinctive in the first two experiments, while the usage of Contrast relation becomes relevant to gender and book identification. These features are not helpful in the PANLiterary experiment, where discourse patterns were not found in the small dataset. The discourse tree width and the subordinate clause width are distinctive in the author identification experiment, while they are not in the other experiments. This is likely because they can serve as indicators of the structural complexity of a text and thus of the idiosyncrasy of a writing style of an individual – as punctuation marks such as periods and commas, which are typical stylistic features. Discourse markers, words with positive sentiment, first person plural pronouns, Wh-Adverbs and modal verbs are distinctive features in the gender identification experiment. The fact that the usage of positive words is only relevant in the gender identification experiment could be caused by the differences in the expressiveness/emotiveness of the writings of men and women. Punctuation marks become very distinctive in the book identification experiment, where the usage of colons, semicolons, parentheses, commas, periods, exclamations and quotation marks are among the most relevant features of the experiment. Syntactic shape features are distinctive in the author identification and PANLiterary experiments while not as impactful in the rest of the experiments.

Author	Gender	Book	PANLiterary
pronouns	AMOD	semicolons	quotations
VC	discourse markers	colons	charsperword
AMOD	pronouns	VB	firstS
commas	firstP	PRP	commas
PRD	VC	MD	hyphens
discourse width	ADV	OBJ	NNP
P	MD	acronyms	subordinate depth
TO	Elaboration	VC	DT
Elaboration	TO	IM	CC
present verbs	OPRD	sentence STD	determiners
subordinate width	PRT	parentheses	PRP
quotations	Contrast	commas	discourse markers
OBJ	PRP	periods	VC
CC	Manner-means	stopwords	VBZ
sentence STD	RBR	OPRD	CONJ
nouns	positive words	AMOD	firstP
OPRD	OBJ	Contrast	PUT
PRP$	WRB	exclamations	LOC-OPRD
HMOD	present verbs	PRP$	coordinate width
periods	sentence range	quotations	adverbs

Table 3: 20 features with the highest information gain in all the experiments

5 Conclusions and Future Work

We have shown that syntactic dependency and discourse features, which have been largely neglected in state-of-the-art proposals so far, play a significant role in the task of gender and author identification and author verification. With more than 88% of accuracy in both gender and author identification within the literary genre, our models that uses them beats competitive baselines. In the future, we plan to experiment with further features and other traits of author profiling.

[5]The features starting with a capital are discourse relations; 'sentence range' is defined as the difference between the minimum and maximum value of words per sentence. 'STD': standard deviation, 'firstP': first person plural pronouns, 'AMOD': Adjective/adverbial modifier f(requency), 'VC': Verb Chain f, 'PRD': Predicative complement f, 'ADV': General Adverbial f, 'P': Punctuation f, 'MD': Modal Verb f, 'TO': Particle *to* f, 'OPRD': Predicative Complement of raising/control verb f, 'PRT': Particle dependent on the verb f, 'OBJ': Object f, 'PRP': Adverbial of Purpose or Reason f, 'CC': Coordinating Conjunction f, 'RBR': Comparative Adverb f, 'PRP$': Possessive Pronoun f, 'WRB': Wh-Adverb f, 'HMOD': Dependent on the Head of a Hyphenated Word f., 'NNP': Singular proper noun f, 'DT': Determiner f, 'VBZ': 3rd person singular present verb f, 'CONJ': Second conjunct (dependent on conjunction) f, 'PUT': Complement of the verb put f, 'LOC-OPRD': non-atomic dependency that combines a Locative adverbial and a predicative complement of a control verb f.

References

Ahmed Abbasi and Hsinchun Chen. 2005. Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems*, 20(5):67–75.

Refat Aljumily. 2015. Hierarchical and non-hierarchical linear and non-linear clustering methods to "shakespeare authorship question". *Social Sciences*, 4(3):758–799.

Faith Amuchi, Ameer Al-Nemrat, Mamoun Alazab, and Robert Layton. 2012. Identifying cyber predators through forensic authorship analysis of chat logs. In *Cybercrime and Trustworthy Computing Workshop (CTC), 2012 Third*, pages 28–37, Ballarat, Australia, October. IEEE Computer Society.

Shlomo Argamon, Moshe Koppel, James W. Pennebaker, and Jonathan Schler. 2009. Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2):119–123.

Bernd Bohnet. 2010. Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 89–97, Beijing, China, August. Association for Computational Linguistics.

John D. Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1301–1309, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Jill Burstein, David Marcu, and Kevin Knight. 2003. Finding the write stuff: automatic identification of discourse structure in student essays. *IEEE Intelligent Systems*, 18(1):32–39.

Na Cheng, Xiaoling Chen, R. Chandramouli, and K. P. Subbalakshmi. 2009. Gender identification from e-mails. In *Computational Intelligence and Data Mining, 2009. CIDM '09. IEEE Symposium on*, pages 154–158, Nashville, TN, USA, March. IEEE Computer Society.

David Crystal and Derek Davy. 1969. *Investigating English style*. Indiana University Press Bloomington.

Chrysanne DiMarco and Graeme Hirst. 1993. A computational theory of goal-directed style in syntax. *Computational Linguistics*, 19(3):451–499.

Michael Gamon. 2004. Linguistic correlates of style: authorship classification with deep linguistic analysis features. In *Proceedings of Coling 2004*, pages 611–617, Geneva, Switzerland, Aug 23–Aug 27. COLING.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177, Seattle, WA, USA, August. ACM.

Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. 2002. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text - Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Pashutan Modaresi and Philipp Gross. 2014. A language independent author verifier using fuzzy c-means clustering. In *CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers*, pages 877–897, Sheffield, UK, September. CEUR-WS.org.

Frederick Mosteller and David L. Wallace. 1963. Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed federalist papers. *Journal of the American Statistical Association*, 58(302):275–309.

Arjun Mukherjee and Bing Liu. 2010. Improving gender classification of blog authors. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 207–217, Cambridge, MA, October. Association for Computational Linguistics.

Sindhu Raghavan, Adriana Kovashka, and Raymond Mooney. 2010. Authorship attribution using probabilistic context-free grammars. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 38–42, Uppsala, Sweden, July. Association for Computational Linguistics.

Ruchita Sarawgi, Kailash Gajulapalli, and Yejin Choi. 2011. Gender attribution: Tracing stylometric evidence beyond topic and genre. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 78–86, Portland, Oregon, USA, June. Association for Computational Linguistics.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W. Pennebaker. 2006. Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 199–205, Palo Alto, California, March. AAAI.

Efstathios Stamatatos, Walter Daelemans, Ben Verhoeven, Martin Potthast, Benno Stein, Patrick Juola, Miguel A. Sanchez-Perez, and Alberto Barrón-Cedeño. 2014. Overview of the Author Identification Task at PAN 2014. In *CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers, 15-18 September, Sheffield, UK*, pages 877–897, Sheffield, UK, September. CEUR-WS.org.

Mihai Surdeanu, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim Nivre. 2008. The conll-2008 shared task on joint parsing of syntactic and semantic dependencies. In *Proceedings of the*

Twelfth Conference on Computational Natural Language Learning, pages 159–177, Manchester, UK, August. Association for Computational Linguistics.

Mihai Surdeanu, Tom Hicks, and Marco Antonio Valenzuela-Escarcega. 2015. Two practical rhetorical structure theory parsers. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 1–5, Denver, Colorado, June. Association for Computational Linguistics.

Hamed Zamani, Hossein Nasr Esfahani, Pariya Babaie, Samira Abnar, Mostafa Dehghani, and Azadeh Shakery. 2014. Authorship identification using dynamic selection of features from probabilistic feature set. In *CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers*, pages 877–897, Sheffield, UK, September. CEUR-WS.org.