Feature Exploration for Authorship Attribution of Lithuanian Parliamentary Speeches

Jurgita Kapočiūtė-Dzikienė[1], Andrius Utka[1], and Ligita Šarkutė[2]

[1] Vytautas Magnus University, K. Donelaičio 58, LT-44248, Kaunas, Lithuania
[2] Kaunas University of Technology, K. Donelaičio 73, LT-44029 Kaunas, Lithuania
j.kapociute-dzikiene@if.vdu.lt, a.utka@hmf.vdu.lt, ligita.sarkute@ktu.lt

Abstract. This paper reports the first authorship attribution results based on the automatic computational methods for the Lithuanian language. Using supervised machine learning techniques we experimentally investigated the influence of different feature types (lexical, character, and syntactic) focusing on a few authors within three datasets, containing transcripts of the parliamentary speeches and debates. Due to our aim to keep as many interfering factors as possible to a minimum, all datasets were composed by selecting candidates having the same political views (avoiding ideology-based classification) from the overlapping parliamentary terms (avoiding topic classification task).
Experiments revealed that content-based features are more useful compared with the function words or part-of-speech tags; moreover, lemma n-grams (sometimes used in concatenation with morphological information) outperform word or document-level character n-grams. Due to the fact that Lithuanian is highly inflective, morphologically and vocabulary rich; moreover, we were dealing with the normative language; therefore morphological tools were maximally helpful.

Keywords: Authorship attribution, supervised ML, Lithuanian.

1 Introduction

Authorship attribution is a process based on an “writeprint” notion. Due to this view, each individual possess his/her own unique idiosyncratic way to express thoughts, which can distinguish him among the others. Authorship attribution analysis is relevant to such applications as author verification (based on the decision if the text is written by the certain author), plagiarism detection (based on finding similarities between the texts written by the different authors), author characterization (based on the extraction of meta information about the authors: i.e. his/her gender, age, education, personality, emotional state, etc.), etc.

But typically authorship attribution task (for review see [18]) is formulated as a task of assigning a text of unknown authorship to one of the candidate authors, when the text samples of those candidates are available. Hence, from the machine learning perspective this problem can be assumed as a supervised multi-class single-label text classification task [17].

The early works on the authorship attribution done for the Lithuanian language date 1971, when the concept of idiolect (individual’s distinctive and unique use of language) was disputed for the first time by Pikčialingis [16]. Since then lots of

P. Sojka et al. (Eds.): TSD 2014, LNAI 8655, pp. 93–100, 2014.
This is preprint prepared by Proceedings editor for Springer International Publishing Switzerland.

descriptive linguistic studies have been done in this field focusing on more specific applications such as forensic linguistics or the analysis of e-mail messages [21]. Despite these important linguistic works, we still need more results on the automatic methods. Consequently, this research is the first attempt at finding a good method to perform authorship attribution task on the Lithuanian texts. In this work we focus on the exploration of the various feature types (lexical, character and syntactic) used together with the supervised machine learning methods. The task is complex due to the couple of reasons. First, the datasets used in our task contain text transcripts of the Lithuanian parliamentary speeches and debates. Since the task is referred to the task of predicting the authorship according to the speaking style, popular orthographic and typographic stylometric features are not valid. Besides, we have to deal with the Lithuanian language, which is highly inflective; ambiguous (47% of all words and their forms are ambiguous); has rich morphology, vocabulary (0.5 million headwords), and word derivation system (e.g. 78 suffixes for diminutives and hypocoristic words).

2 Methodology

2.1 The Datasets

Our experiments were carried out on three datasets to make sure that findings generalize over the different domains. All datasets were composed of the text transcripts of the Lithuanian parliamentary speeches and debates [9], thus represent formal spoken, but normative Lithuanian language. The text transcripts are from the regular parliamentary sessions and cover the period of 7 parliamentary terms starting from March 10, 1990 and ending with December 23, 2013.

Due to the fact that we want to find the best set of features for the authorship attribution, the other interfering factors that determine variation in the text and could facilitate our task must be kept to a minimum, therefore:

- Selected author candidates were from the overlapping (but not necessary from absolutely the same) parliamentary terms (thus avoiding topic classification task).
- Selected author candidates had the same political views (thus avoiding ideology-based classification task).

There is no consensus about the minimum text length appropriate for the authorship attribution. Some of the researchers argue that 2,500 words is the optimal length independent of the random noise [12], some of them obtain promising results with the text fragments of 500 words [6], the others achieve reasonable results with the extremely short texts (with an average length of only 60 words) [10], [15].

The majority of the parliamentary speeches are rather short, thus we could hardly follow the recommendations of e.g. 2,500 words. Besides, we wanted to test the robustness of our methods dealing with the short texts, therefore 150 words was chosen as the minimum text length and all shorter text samples were filtered out.

After previously described pre-processing steps three datasets were created (see Table 1)[3]. Each of them refers to the different Lithuanian party groups (*Social Democrats*,

[3] We did not balance our datasets, because it can cause negative influence on the results [13].

Conservatives, and *Liberals*) and contains parliamentary speeches and debates of three different parliamentarians (e.g. *Social Democrats* contains speeches and debates of parliamentarians with id *SD*1, *SD*2, and *SD*3).

Table 1. Statistics about all datasets: categories; number of texts; number of (distinct) tokens, and lemmas; average text lengths; random and majority baselines.

Dataset	Category	Numb. of texts	Numb. of tokens	Numb. of dist. tokens	Numb. of dist. lemmas	Avg. text length	Random baseline	Majority baseline
Social Democrats	SD1	3,932	925,864	58,673	19,263	235.47	0.3685	0.4863
	SD2	2,130	451,212	36,621	12,524	211.84		
	SD3	2,024	400,807	35,719	12,987	198.03		
	TOTAL	8,086	1,777,883	84,484	26,845	219,87		
Conser-vatives	C1	5,093	1,075,880	57,728	18,616	211.25	0.3738	0.4974
	C2	2,654	455,712	29,612	11,893	171.71		
	C3	2,491	484,446	32,220	12,609	194.48		
	TOTAL	10,238	2,016,038	77,885	25,986	196.92		
Liberals	L1	2,539	456,501	31,239	11,371	179.80	0.4602	0.6209
	L2	930	187,501	24,594	8,599	201.61		
	L3	619	125,523	20,995	7,863	202.78		
	TOTAL	4,088	769,525	50,566	16,581	188.24		

2.2 Experimental Setup

Before the experiments all three datasets were preprocessed:

- *Tokenized* (see Table 1),
- *Lemmatized* (see Table 1). Documents were processed using Lithuanian morphological analyzer and lemmatizer "Lemuoklis" [20,3], which replaces all numbers with the special tag and transforms generic words into the lowercase. This preprocessing technique reduced the number of tokens by ~32–33%.
- *Part-of-speech tagged*. Documents were processed using "Lemuoklis", which also performs coarse-grained (identifies main 18 part-of-speech categories, such as noun, verb, etc.) and fine-grained (identifies 12 morphological categories, such as case, gender, tense, etc.) part-of-speech tagging.

We explored the wide range of the individual and compound features that covered lexical, character and syntactic levels[4]:

- *fwd* – the content-free lexical feature which involves only function words[5]. This feature type by consensus is considered as the topic-neutral and was proved to be a relatively good identifier of the author writing style [1].

[4] All typographic and orthographic features were skipped, because we are dealing with the speaking (not writing) style of the author.

[5] Considering the Lithuanian language specific, the function words are: prepositions, pronouns, conjunctions, particles, interjections, and onomatopoeias.

- *lex* – the most popular content-based lexical feature which involves word unigrams or their interpolation (up to *n*=3 in our experiments).
- *lem* – content-based lexical feature, especially recommended for the highly inflective languages, which involves lemmas based on the word unigrams or their interpolation (up to *n*=3 in our experiments).
- *chr* – character feature which involves document-level character n-grams (*n*=[2, 7] in our experiments) – i.e. successions of *n* characters including spaces and punctuation marks. This feature type was proved to surpass other types for Dutch in authorship attribution [11]. Moreover, it gave the most accurate results for the Lithuanian topic classification task [5].
- *pos* – content-free syntactic feature which involves coarse-grained part-of-speech tags based on the word unigrams or their interpolation (up to *n*=3 in our experiments). This feature type is not among the most accurate, but it usually selected for the comparison purposes or used in concatenation with the lexical features.
- *lexpos*, *lempos*, *lexmorf*, *lemmorf* – aggregated features which involves unigrams of concatenated lexical and syntactic features or their interpolation (up to *n*=3 in our experiments) (e.g. *žodis_dktv* (word_noun) is the example of *lexpos* feature). Feature *morf* – determines single string of the concatenated fine-grained morphological category values (e.g. *morf*(*esanti* (existent))="*Noun_Nominative_Feminine_Singular_Present_Active_Non-reflexive_Non-pronominal_Positive*").

Taking into account inflective character of the Lithuanian language and the fact that the normative language texts are used in all our datasets, we formulated our main hypothesis which states that authorship attribution should significantly benefit from the morphological information, in particular, from lemmatization. Besides, the results should be even more improved using as much morphological information as possible, in particular, lexical features in concatenation with the syntactic.

2.3 Classification

In this paper we focus on the supervised machine learning techniques [7] applied to the text categorization [17]) and used for the authorship attribution [18].

The aim of our task is to find a method, which could distinguish authors from each other by creating a model the best approximating the "writeprints" of each individual author speaking style.

We explored the features described in the previous Section using two different machine learning approaches:

- *Support Vector Machine* (SVM) [2] – discriminative instance-based approach is the most popular technique for the text classification, because it can cope with the high dimensional feature spaces (e.g. 84,484 word features in *Social Democrats* dataset; 77,885 in *Conservatives* dataset; and 50,566 in *Liberals* dataset) and the sparseness of the feature vector (only ~220 non-zero feature values among 84,484 in *Social Democrats* dataset instances; ~197 in among 77,885 in *Conservatives*; and ~188 among 50,566 in *Liberals*).

- *Naive Bayes Multinomial* (NBM) [8] – generative profile-based approach was selected for the comparison purposes, in particular, because it is very fast, performs especially well when the number of features is large, and sometimes surpass SVM.

In our experiments we used chi-squared feature extraction method, SMO polynomial kernel (it gave the highest accuracy in our preliminary control experiments) with SVM and NBM implementations in WEKA [4] machine learning toolkit, version 3.6 [19]. All remaining parameters were set to their default values.

3 Results

We performed experiments based on the stratified 10-fold cross-validation with SVM and NBM methods using feature types, described in Section 2.2, but for clarity reasons (not to overload with the information) only the best results of each type are reported (see Table 2 and Table 3). All obtained results are reasonable, because they outperform random and majority baselines.

Table 2. Accuracies, macro-averaged and micro-averaged F-scores for all datasets with different feature types and SVM. The best obtained authorship attribution results are underlined.

Feature types	**Social Democrats**			**Conservatives**			**Liberals**		
	acc.	**microF**	**macroF**	**acc.**	**microF**	**macroF**	**acc.**	**microF**	**macroF**
fwd	0.8260	0.8250	0.8150	0.7931	0.7930	0.7783	0.8312	0.8300	0.7773
chr3	0.9129	0.9130	0.9090	0.9204	0.9200	0.9173	0.9261	0.9260	0.9027
lex1	0.9374	0.9370	0.9347	0.9302	0.9300	0.9277	0.9354	0.9350	0.9193
lex2	0.9375	0.9370	0.9347	0.9363	0.9360	0.9340	0.9291	0.9290	0.9100
lem1	0.9440	0.9440	0.9407	0.9400	0.9400	0.9380	0.9511	0.9510	0.9357
lem2	0.9489	0.9490	0.9470	0.9384	0.9380	0.9360	0.9550	0.9550	0.9417
pos3	0.8167	0.8160	0.8080	0.8272	0.8270	0.8160	0.8378	0.8380	0.7963
lexpos3	0.9387	0.9390	0.9353	0.9318	0.9320	0.9290	0.9359	0.9360	0.9187
lempos3	0.9535	0.9530	0.9513	0.9334	0.9330	0.9310	0.9496	0.9500	0.9367
lexmorf2	0.9400	0.9400	0.9373	0.9355	0.9360	0.9330	0.9293	0.9290	0.9103
lexmorf3	0.9389	0.9390	0.9360	0.9326	0.9330	0.9300	0.9247	0.9250	0.9040
lemmorf2	0.9436	0.9440	0.9410	0.9326	0.9330	0.9300	0.9418	0.9420	0.9270
lemmorf3	0.9427	0.9430	0.9397	0.9333	0.9330	0.9310	0.9406	0.9410	0.9260
Random baseline:	*0.3685*			*0.3738*			*0.4602*		
Majority baseline:	*0.4863*			*0.4974*			*0.6209*		

4 Discussion

Zooming into the results presented in Table 2 and Table 3, allows as to report the following statements.

The content information is very important to achieve high authorship attribution accuracy: i.e. content-free feature types (based on the function words or part-of-speech tags) are easily beaten.

Table 3. Authorship attribution results with NBM.

Feature types	**Social Democrats**			**Conservatives**			**Liberals**		
	acc.	**microF**	**macroF**	**acc.**	**microF**	**macroF**	**acc.**	**microF**	**macroF**
fwd	0.7857	0.7870	0.7780	0.7125	0.7090	0.6893	0.7774	0.7820	0.7353
chr3	0.7833	0.7840	0.7770	0.7232	0.7210	0.7087	0.7432	0.7520	0.7240
lex1	0.8865	0.8850	0.8773	0.7744	0.7750	0.7730	0.7740	0.7830	0.7763
lex2	0.8910	0.8900	0.8823	0.7719	0.7720	0.7677	0.7630	0.7740	0.7660
lem1	0.8700	0.8700	0.8640	0.7740	0.7760	0.7753	0.7833	0.7910	0.7787
lem2	0.8922	0.8910	0.8850	0.7701	0.7700	0.7670	0.7725	0.7820	0.7713
pos3	0.6838	0.6840	0.6787	0.7131	0.7110	0.6950	0.7277	0.7390	0.6993
lexpos3	0.8909	0.8900	0.8823	0.7678	0.7670	0.7620	0.7561	0.7660	0.7547
lempos3	0.8957	0.8950	0.8887	0.7690	0.7690	0.7647	0.7605	0.7720	0.7613
lexmorf2	0.8918	0.8900	0.8823	0.7712	0.7710	0.7670	0.7625	0.7730	0.7617
lexmorf3	0.8925	0.8910	0.8833	0.7680	0.7670	0.7620	0.7571	0.7670	0.7563
lemmorf2	0.8957	0.8940	0.8867	0.7762	0.7760	0.7727	0.7605	0.7720	0.7593
lemmorf3	0.8978	0.8960	0.8890	0.7757	0.7750	0.7697	0.7544	0.7670	0.7550
Random baseline:	*0.3685*			*0.3738*			*0.4602*		
Majority baseline:	*0.4863*			*0.4974*			*0.6209*		

Document-level character n-grams reaching the peak at *n*=3 are surpassed by the content-based lexical features. Thus, we can assume that for this authorship attribution task character n-grams are not robust enough to capture the patterns of the Lithuanian inflection system intrinsically as it can be done with the lemmas. Lemmas, reducing number of features by ~32–33%, also reduce the sparseness of the feature vector which, in turn, results in a more robust classification model creation.

The best results with NBM are obtained using concatenated lemmas & fine-grained part-of-speech tags based on the interpolation of unigrams to trigrams with *Social Democrats* dataset; concatenated lemmas & fine-grained part-of-speech tags based on the interpolation of unigrams & bigrams with *Conservatives* dataset; and lemmas based on the token unigrams with *Liberals* dataset. SVM method is much more accurate compared to NBM. Thus, the best results with SVM and the best overall results are obtained using concatenated lemmas & coarse-grained part-of-speech tags based on the interpolation of unigrams & bigrams with *Social Democrats* dataset; lemmas based on the token unigrams with *Conservatives* dataset; and lemmas based on the interpolation of token unigrams & bigrams with *Liberals* dataset. McNemar [14] test with one degree of freedom applied on the results with the different feature types revealed that differences between features based on lemmas or used in concatenation with lemmas in most of the cases are not statistically significant, but mostly are statistically significant when compared with the other feature types. Hence, it allows as to confirm our hypothesis that lemmatization and morphological information indeed improves the results. Moreover, the analysis of words unrecognized by the lemmatizer allows assuming that the results would probably be even better, if lemmatizer could cope with the shortened endings (often used in the spoken Lithuanian language) and deal with a bigger set of the international words and named entities.

5 Conclusions and Future Work

In this paper we report the first authorship attribution results on the normative Lithuanian language texts (transcripts of the Lithuanian parliamentary speeches and debates) obtained with the supervised machine learning techniques.

We formulated and experimentally confirmed our hypothesis, that such highly inflective and morphologically rich language as Lithuanian (especially when dealing with the normative texts) mostly benefits from the morphological information, in particular, lemmatization; besides lemmas supplemented with the part-of-speech information can even more boost the performance.

In the future research we are planning to expand the number of authors in the datasets; to experiment with the different domains (e.g. blog data, tweets, etc.) and language types (not only normative Lithuanian).

Acknowledgments. This research was funded by a grant (No. LIT-8-69) from the Research Council of Lithuania.

References

1. Argamon, S., Levitan, S.: Measuring the usefulness of function words for authorship attribution. In: the 2005 joint Conference of the Association for Computers and Humanities and the Association for Literary and Linguistic Computing, pp. 1–3. (2005)
2. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning, vol. 20(3), pp. 273–297. (1995)
3. Daudaravičius, V., Rimkutė E., Utka A.: Morphological annotation of the Lithuanian corpus. In: Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies (ACL'07), pp. 94–99 (2007)
4. Hall, M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten, I. H.: The WEKA Data Mining Software: An Update. In: SIGKDD Explorations, vol. 11(1), pp. 10–18. (2009)
5. Kapočiūtė-Dzikienė, J., Vaassen F., Daelemans W., Krupavičius, A.: Improving Topic Classification for Highly Inflective Languages. In: 24th International Conference on Computational Linguistics (COLING 2012), pp. 1393–1410 (2012)
6. Koppel, M., Schler, J., Bonchek-Dokow, E.: Measuring Differentiability: Unmasking Pseudonymous Authors. Journal of Machine Learning Research, vol. 8, pp. 1261–1276 (2007)
7. Kotsiantis, S. B.: Supervised Machine Learning: A Review of Classification Techniques. Informatica, vol. 31, pp. 249–268 (2007)
8. Lewis, D. D, Gale, W. A.: 1994. A sequential algorithm for training text classifiers. In: 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR-94), pp. 3–12 (1994)
9. Lithuanian Parliament official page, `http://www3.lrs.lt/pls/inter/w5_sale.kad_ses`
10. Luyckx, K.: Authorship Attribution of E-mail as a Multi-Class Task – Notebook for PAN at CLEF 2011. In: Petras, V., Forner P., Clough P. (eds.) Cross-Language Evaluation Forum (Notebook Papers/Labs/Workshop) (2011)
11. Luyckx, K., Daelemans, W.: The effect of author set size and data size in authorship attribution. Literary and Linguistic Computing, vol. 26(1), pp. 35–55. (2011)

12. Maciej, E.: Does size matter? Authorship attribution, small samples, big problem. Literary and Linguistic Computing (2013)
13. Manning, Ch. D., Schütze, H.: Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA, USA (1999)
14. McNemar Q. M.: Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika, vol. 12(2), pp. 153–157. (1947)
15. Mikros, G. K., Perifanos K.: Authorship identification in large email collections: Experiments using features that belong to different linguistic levels – Notebook for PAN at CLEF 2011. In: Petras, V., Forner P., Clough P. (eds.) Cross-Language Evaluation Forum (Notebook Papers/Labs/Workshop) (2011)
16. Pikčilingis, J. Kas yra stilius?[What is style?]. Vaga, Vilnius. (in Lithuanian) (1971)
17. Sebastiani, F.: Machine Learning in Automated Text Categorization. ACM Computing Surveys, vol. 34, pp. 1–47 (2002)
18. Stamatatos, E.: A Survey of Modern Authorship Attribution Methods. Journal of the Association for Information Science and Technology, vol. 60(3), pp. 538–556 (2009)
19. WEKA Machine Learning Toolkit, `http://www.cs.waikato.ac.nz/ml/weka/`
20. Zinkevičius, V.: Lemuoklis – morfologinei analizei [Morphological analysis with Lemuoklis]. Gudaitis, L. (ed.) Darbai ir dienos, vol. 24, pp. 246–273. (in Lithuanian) (2000)
21. Žalkauskaitė, G.: Idiolekto požymiai elektroniniuose laiškuose [Idiolect signs in the e-mails]. PhD dissertation, Vilnius University, Lithuania. (2012)