Grammar Checker Features for Author Identification and Author Profiling

Notebook for PAN at CLEF 2013

Roman Kern

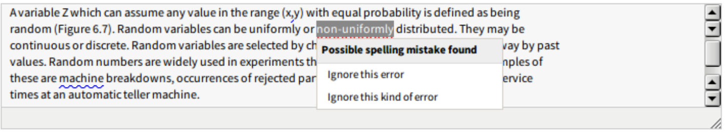
Know-Center
rkern@know-center.at

Abstract Our work on author identification and author profiling is based on the question: Can the number and the types of grammatical errors serve as indicators for a specific author or a group of people? In order to detect the grammatical errors we base our approach on the output of the open-source library LanguageTool. In the case of the author identification we transform the problem into a statistical test, where an unknown document is written by another author when the distribution of grammatical errors deviated from documents of a reference corpus. For author profiling we implemented an instance based classification approach, namely a k-NN classifier, in combination with a Language Model where a text is assigned to a specific age or gender group where the according reference corpus contains the closest match. In the evaluation we found that for both scenarios grammatical errors do perform better than the baseline and do capture an aspect of a writing style, which is not contained in more traditional features, like stylometric features or word n-grams.

1 Introduction

The task of author identification and author profiling can be seen as similar problems. Author identification is the task to find out whether a previously unseen text document has been authored by the same person as a number of reference documents. Therefore the problem can be reformulated to: Does a given text match a specific writing style of a single person. In the case of author profiling one tries to infer certain characteristics of an author from given piece of text. Again the problem can be phrased as: Does a given text match a specific writing style of a group of people. A overview of the tasks in the context of the PAN 2013 is given in [4].

In both cases one can assume that in the general case the content of the text cannot be seen as a reliable indicator for a match. An overview of stylometric features and main approaches is given in [5]. Using lexical errors and syntactic errors for authorship identification has already been proposed in the past [3]. The authors state that this approach is similar to some extend to the way how humans assess the authorship of text document. One downside of such a approach is that tools to detect those writing errors do not deliver the necessary performance and heavy post-processing seems unavoidable. We follow the same intuition for our approach and study the effectiveness of a contemporary grammar checking tool for authorship identification and profiling.



A variable Z which can assume any value in the range (x,y) with equal probability is defined as being
random (Figure 6.7). Random variables can be uniformly or non-uniformly distributed. They may be
continuous or discrete. Random variables are selected by ch
values. Random numbers are widely used in experiments th
these are machine breakdowns, occurrences of rejected par
times at an automatic teller machine.

vay by past
nples of
ervice

Possible spelling mistake found

Figure 1. Example for a short snippet of text which contains 2 errors according to the LanguageTool. For the second annotated location, LanguageTool suggests: “Consider using a past principle here: ’machined”’.

2 Approach

The central component of our authorship identification and profiling system is a component to detect grammatical errors within text. Here we employ the open-source tool LanguageTool[1], which is a style and grammar checker. It works for 20 different languages and can be easily be extended to include additional rules. To illustrate the output of the LanguageTool library an example is depicted in figure 1, where two different types of errors are detected, where the example is directly taken from the PAN 2013 authorship identification data-set. Additionally to the feature generated out of the LanguageTool grammar checker, we integrated more traditional stylometric features into our system.

Author Identification The task of author identification is transformed into statistical test, where the input is a set of reference documents from a single author and an unknown document. The documents are processed independently from each other, where each document is fed through the feature extraction pipeline. The pipeline consists of two stages, where in the first stage a number of feature spaces are filled, and in the second stage the feature spaces of the reference document are merged into a single meta feature-space. The feature spaces for the first stage are: i) stylistic and grammatical errors, ii) basic statistics, e.g. number of lines, iii) stylometric statistics, e.g. hapax legomena, iv) stem suffixes, v) slang words, and vi) sentence structure. The last feature space is optional and not enabled by default, as the run-time increases dramatically, which is due to the use of a sophisticated parser component - the Stanford Parser [2]. All but the first feature space have already been used for Authorship Attribution by our system [1].

For all the feature spaces of the reference documents are then aggregated and compared to corresponding the feature spaces of the unknown document. Out of the comparison a final meta feature space is generated. The binary features of the meta feature space are for the majority of feature spaces: i) more than minimum, ii) less than maximum, iii) within minimum and maximum, and iv) about mean, which integrates the standard deviation. For the grammatical features, a more sophisticated route is taken. Here the probability distribution of individual style and grammar error types are smoothed and pairwise compared between all documents, including the reference documents as well as the reference document. For the comparison the Kolmogorov–Smirnov

[1] http://www.languagetool.org/

Table 1. Performance of our Authorship Identification system, where the F_1 performance measure is used.

Data-Set	English	Spanish	Greek
Pan 2012 - Small	0.727	-	-
Pan 2012 - Medium	0.727	-	-
Pan 2012 - Large	0.800	-	-
Pan 2013 - Train	0.800	1.000	0.583
Pan 2013 - Test	0.533	0.560	0.500

test is used. Here the binary meta features are: i) same distribution for close matches, and ii) about the same distribution for less close matches. None of the the involved threshold have been extensively evaluated and were set in a ad-hoc manner.

For the final decision the binary of the meta feature space are combined: $\frac{|F_{true}|}{|F_{true}|+|F_{false}|}$, where F_{true} is the set of all meta features with a positive value. If this ratio excess $.35$ the unknown document is assumed to be sufficiently similar to the reference documents.

Author Profiling For the author profiling the task is to identify the age group and the gender of the author of a given text document. For this task we combined two algorithmic approaches and two difference feature types. The two algorithmic approaches are: i) Language Models, and ii) a k-NN classification algorithm. In terms of feature types we again used the output of the style and grammar checker, as well as word tri-grams. The system is build in a flexible way which allows to freely combine features and algorithms. In the training phase the reference corpus is processed and the Language Models and the k-NN lookup index are build. For all of the groups within the reference data-set a separate Language Model is build, which captured how often a specific feature is used within the document associated with the specific group. For the k-NN classifier, a single Apache Lucene[2] index is build, where the user groups are stored are separate fields.

When a previously unseen document is processed, the results from the Language Models and the k-NN classifier can be combined. In the case of the Language Models, for each group a score is computed by iterating over all features: $score_{group}(feature) = \sum \frac{P(feature|group)}{P(feature)}$, where $P(feature|group)$ is the probability of feature for a given group. In the case of the k-NN classifier, the index is search by using the features of the unseen document as query. The top three results are then examined and the score from the search engine are summed to give a final ranking of groups. When more than one algorithmic approach are used, they are processed in sequence. The first approach which provides a score, instead of no result or a tie, is then taken as final decision.

3 Evaluation

To assess the performance of our system for Authorship Identification we report the performance numbers not only for the PAN 2013 data-sets, but also for three datasets, which we assembled out of the PAN 2012 data-set. In table 1 the performance

[2] http://lucene.apache.org/

Table 2. Performance of our Authorship Profiling system on the PAN 2013 data-set for three selected configurations, where the F_1 is used as performance measure.

Configuration	**Language**	**Age: 10s**	**Age: 20s**	**Age: 30s**	**Gender: Male**	**Gender: Female**
k-NN + Trigrams (knn-tri)	English	0.263	0.543	0.701	0.613	**0.605**
Language Model + Grammar (lm-lt)	English	0.005	0.031	**0.721**	**0.643**	0.375
knn-tri + lm-lt (default)	English	**0.266**	**0.527**	0.700	0.618	0.603
k-NN + Trigrams (knn-tri)	Spanish	**0.105**	0.601	**0.478**	0.567	0.554
Language Model + Grammar (lm-lt)	Spanish	0.000	**0.721**	0.134	**0.642**	0.596
knn-tri + lm-lt (default)	Spanish	0.011	0.651	0.458	0.619	**0.598**

of our system for the available data-sets for the three languages is reported. To assess the performance of our system for Author Profiling, we took the PAN 2013 data-set as provided by the organisers and split it into two parts. The first part, which contains 70% of all conversations is used for training and the remaining conversations are used as testing data-set. In table 2 the performance for three selected configurations is reported.

4 Conclusions

We studied the effectiveness of style and grammar errors for Authorship Identification and Author Profiling. Therefore we build a system which combines the output of a grammar checker tool with stylometric features, which have been used for Authorship Attribution already in the past. We found that these features derived from the grammatical errors does help in such scenarios and that they capture different aspect of the writing style then the remaining stylometric features. We found that further tuning of our system is necessary as the performance figures do vary considerably between different data-sets. In the future we further plan to use stylistic and grammatical errors as indicators for authorship, especially as any improvements in detecting these errors will also be beneficial for our approach.

References

1. Kern, R., Klampfl, S., Zechner, M.: Vote/veto classification, ensemble clustering and sequence classification for author identification. CLEF 2012 Evaluation Labs and Workshop – Working Notes Papers 2012, 09–20 (2012)
2. Klein, D., Manning, C.D.: Accurate unlexicalized parsing. Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - ACL '03 pp. 423–430 (2003)
3. Koppel, M., Schler, J.: Exploiting Stylistic Idiosyncrasies for Authorship Attribution, pp. 69–72. No. 2000 (2003)
4. Potthast, M., Gollub, T., Hagen, M., Tippmann, M., Kiesel, Johannes Stamatatos, E.R.P., Stein, B.: Overview of the 5th International Competition on Plagiarism Detection (2013)
5. Stamatatos, E.: A survey of modern authorship attribution methods. Journal of the American Society for Information Science 60(3), 538–556 (2009)