PAN-2013

Author Identification & Author Profiling

Grammar Checker Features for Author Identification and Author Profiling

Overview

Hypothesis

The number and the types of **grammatical and stylistic errors** serve as indicators for a specific author or a group of people.

Challenges

- How to **analyse** the text and **identify** grammatical and stylistic errors?
- How to transform these errors into a **representation suitable for the machine** to fulfil the task?

Approach

- Prepossessing based on open-source **NLP** tools
- Open-source grammar checker as input
- Transformation into binary features & feature vectors
- Supervised **Machine Learning & Information Retrieval** techniques

The whole system is available as open-source:

https://www.knowminer.at/svn/opensource/projects/pan2013/trunk

Feature Extraction

LanguageTool Features

The central component of our authorship identification and profiling system is a component to detect grammatical errors within text, which has already been reported as suitable features for the task of author identification [1, 2]. Here we employ the open-source tool LanguageTool [3], 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.

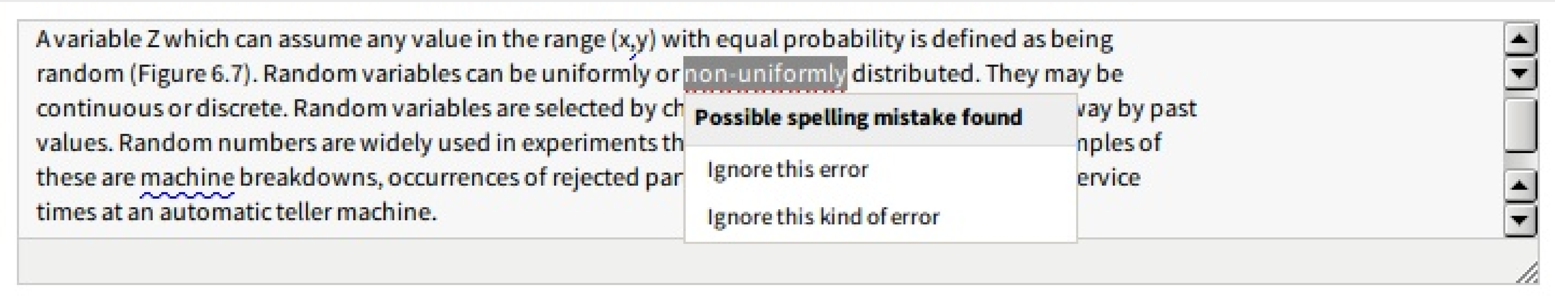

Figure 1: Example for a short snippet of text which contains 2 errors according to the LanguageTool. For the second annotated location, LanguageTool suggests: "Consider using a past principle here: 'machined'".

Basic Statistics Features

- Lines: Number of lines, number of characters per line, max line length, ...
- Sentence: Number of tokens per sentence, number of punctuations per sentence, ...
- Paragraph: Number of paragraphs, number of characters per paragraph, ...
- Document: Number of tokens, number of stop words, ratio of capital letters, ...

Vocabulary & Stylometric Features

- Hapax Legomena, Hapax Dislegomena, Yules K, Simpsons D, Sichels S, Honores H, Brunets W [4]
- Ratio of alphanumeric characters, ratio of white-space characters, ...

Stem Suffix Feature

The suffix of the words, which would be remove by a stemmer, i.e. the Snowball stemmer.

Slang Word Feature

Features generated out of the slang words contained within the text, where there are three lists of such words: Internet slang words, swear words and common smileys.

Sentence Structure Features

Features generated out of the Stanford Parser, which generates a parse tree and typed dependencies for the grammatical roles. There the depth of the parse tree and statistics of the types dependencies are taken as features. *(These features are by default disabled, as the parser component takes considerable time to compute.)*

For more information about the features please see [5], or the source code :)

Author Identification

Feature Spaces

Each set of features is transformed into a feature space. For each feature spaces the features of all documents are combined into a single meta feature space. The binary features of the meta feature space are then the comparison of the reference document and the text document: i) more than minimum, ii) less than maximum, iii) within minimum and maximum, and iv) about mean, which integrates the standard deviation.

The grammatical features are interpreted as sample of a probability distribution, each document being a single sample. This input data is smoothed and pair-wise compared between the reference documents and the test document. For the comparison the Kolmogorov–Smirnov test is used. Here the binary features are: i) same distribution for close matches, and ii) about the same distribution for less close matches.

Classification

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.

Results

Additionally to the Pan 2013 data set we also report on the results from a evaluation data set generated out of the Pan 2012 data.

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

Author Profiling

Feature Spaces

For author profiling we just use two feature spaces (for the submitted system): i) output of the style and grammar checker and ii) word tri-grams.

Algorithmic Approaches

We provide two main approaches: i) Language Models, and ii) the k-NN classification algorithm. Our system allows to play with any combination of algorithm and features spaces, where the algorithms are applied in sequence and the first algorithm which provides a score (ignoring ties) is taken as final result.

Language Model

Training: For each group (genders, age groups) a single Language Model is build from the training documents for P(feature|group).

Classification: Iterate over all features: $score_{group}(feature) = \sum \frac{P(feature|group)}{P(feature)}$

k-NN Classifier

Training: Each document is treated a single instance, with the gender and age group of the author stored alongside.

Classification: Combine the similarity score from the top 3 nearest neighbours.

Results

We report the results on the training data set, which has been randomly split into 70% used for training and 30% for testing:

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**

References

[1] Koppel, M., Schler, J.: Exploiting Stylistic Idiosyncrasies for Authorship Attribution. pp. 69–72. No. 2000 (2003)
[2] Stamatatos, E.: A survey of modern authorship attribution methods. Journal of the American Society for Information Science 60(3), 538–556 (2009)
[3] Homepage of the LanguageTool: http://www.languagetool.org/
[4] Tweedie, F., Baayen, H.: How variable may a constant be? Measures of lexical richness in perspective. In: Computers and the Humanities. pp. 323–352 (1998)
[5] 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

Roman Kern
Know-Center GmbH
rkern@know-center.at

Acknowledgements: The Know-Center is funded within the Austrian COMET Program - Competence Centers for Excellent Technologies - under the auspices of the Austrian Federal Ministry of Transport, Innovation and Technology, the Austrian Federal Ministry of Economy, Family and Youth and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG.