Complexity Measures and POS N-grams for Author Identification in Several Languages

SINAI AT PAN@CLEF 2018

Rocío López-Anguita Arturo Montejo-Ráez Manuel C. Díaz-Galiano

SINAI Research Group, Centro de Estudios Avanzados en TIC, Universidad de Jaén (Spain)

Objective

Given a set of documents (known *fanfics*) by a small number (up to 20) of candidate authors, to identify the authors of another set of documents (unknown *fanfics*). All documents are in the same language that may be English, French, Italian, Polish, or Spanish.

Strategies

We have applied a supervised learning approach, where features are constructed according to two different strategies:

1. Use of several measures of the complexity of the fanfics texts for each candidate
2. Analysis the fanfics of each candidate by applying a Part-Of-Speech Tagger and a n-gram based vector space model

Conclusions

Complexity metrics considered are not very helpful to identify the author of a text. This could be explained because all these smaller characteristics (rare words, punctuation marks, sentence length...) into a final index of readability or complexity may have nothing to do with author style or characterization.
Our second approach, the use of **n-grams of POS-tags**, seems a better approach to the problem, although results are from very bad (in the case of Polish, Problem00007) to very good (English, Problem00002). These results need of further analysis.

Measure-based features

A vector of numerical features is generated, where each value corresponds to one of the following complexity-readability formulas:

Complexity measure	EN	ES	FR	IT	PL
Punctuation Marks	✓	✓	✓	✓	✓
Sentence Complexity Index	✓	✓	✓	✗	✗
Automated Readability Index	✓	✓	✓	✓	✓
μ Readability	✓	✓	✓	✗	✗
Dependency Tree Height	✓	✓	✗	✗	✗
FOG	✓	✗	✗	✗	✓
Flesch	✓	✗	✗	✗	✓
Flesch-Kincacid	✓	✗	✗	✗	✓
SMOG	✓	✗	✗	✗	✗
Lexical Complexity	✗	✓	✗	✗	✗
Spaulding Readability	✗	✓	✗	✗	✗
Fernández-Huerta Readability	✗	✓	✗	✗	✗
Flesch-Szigrist Readability	✗	✓	✗	✗	✗
Gutierrez Readability	✗	✓	✗	✗	✗
Minimum Age of Readability	✗	✓	✗	✗	✗
SOL	✗	✓	✓	✗	✗
Crawford	✗	✓	✗	✗	✗
Kandel-Models	✗	✗	✓	✗	✗
Dale Chall	✗	✗	✓	✗	✗
Flech-Vaca	✗	✗	✗	✓	✗
Gulpease	✗	✗	✗	✓	✗
Pisarek	✗	✗	✗	✗	✓

POS N-gram features

We obtain the POS tags for each fanfic text and apply the TF on POS N-grams and SVM for final classification. For English, Spanish and French we have used the *Freeling* tool to process texts. For Italian and Polish we used the *NLTK* Python library, and for Polish we used TreeTagger to get the POS tags for each text. For both we have applied Python's *SciKit-Learn* libraries for automatic learning (SVC classifier).
The experiments we have conducted are parametrized on options like how the normalization of the vectors is computed (L2 over samples or over features), the weighting scheme used (TF or TF.IDF) or the maximum length of POS n-grams considered (from 2 up to 4).

Official results

Problem	Macro F1
Problem 00001	0.110
Problem 00002	0.202
Problem 00003	0.078
Problem 00004	0.235
Problem 00005	0.102
Problem 00006	0.109
Problem 00007	0.052
Problem 00008	0.276
Problem 00009	0.032
Problem 00010	0.296

Dev results

The highest results we obtained on the training set are with the POS-TF (1,2,3,4-grams) with L2-normalization per sample.

Problem	Language	F1 (Measures)	F1 (POS N-grams)
00001	English	0.035	0.435
00002	English	0.143	0.838
00003	French	0.038	0.475
00004	French	0.27	0.605
00005	Italian	0.09	0.365
00006	Italian	0.299	0.512
00007	Polish	0.023	0.123
00008	Polish	0.358	0.447
00009	Spanish	0.021	0.373
00010	Spanish	0.287	0.404

Contact information

Web: http://sinai.ujaen.es
EMail: amontejo@ujaen.es

Acknowledgements

This work has been partially funded by the Spanish Government under the REDES Project (TIN2015-65136-C2-1-R).

Sistemas Inteligentes de Acceso a la Información