SU@PAN'2015: Experiments in Author Verification

Stanimir Nikolov,[1] Dobrinka Tabakova,[1] Stefan Savov,[1] Yasen Kiprov,[1] Preslav Nakov [2]

[1] Sofia University, FMI [2] Qatar Computing Research Institute, HBKU

Introduction

We discuss the participation of the Sofia University team, kiprov15, in the 2015 edition of the Author Verification task, part of the PAN 2015. Team kiprov15 experimented with an SVM classifier using variety of features extracted from publicly available resources.

Data

The training set contains examples from 100 authors for each language: English, Dutch, Greek, Spanish.

Pipeline

1. GATE ANNIE Tokenizer
2. GATE ANNIE Sentence Splitter
3. Groovy script for adding features
4. Groovy script for adding n-grams

Method

We extracted the above-described features, and as a result, for each document we obtained a feature vector. Given a problem, i.e, a set of known documents and a questioned document, we aggregated these feature vectors for all known documents. Similarly, we built a feature vector for the questioned document (but this time there was nothing to aggregate as it is only one). Finally, we produced a 10-dimensional vector for the (known set, questioned document) pair as follows: for the first seven features (i.e., excluding the n-grams), we just subtracted them, and for the n-gram features, we calculated separately the cosine similarity for the 1-grams, the 2-grams and the 3-grams, and we used the values as eighth, ninth and tenth features. Then, we scaled the real values to the [0;1] range, and we saved the scaling factors. We further added a class label: same or different (author). On testing, we produced the 10-dimensional vectors in the same way, except that we reused the scaling factors from training.

Features

- Average sentence length to character count ratio;
- Average sentence length to word count ratio;
- Average word length;
- Average paragraph length to word count ratio;
- Average paragraph length to sentence count ratio;
- Punctuation to word count ratio;
- Sentence count to word count ratio;
- Word based n-grams of sizes 1,2,3;
- Number of tweets starting with a user mention.

Results

Language	AUC	C1	Final Score	Runtime	Placement
Greek	0.7086	0.6400	0.4535	00:01:01	9/15
English	0.4926	0.5243	0.2582	00:01:35	15/18
Spanish	0.2802	0.3400	0.0953	00:01:09	17/17
Dutch	0.2560	0.3476	0.0890	00:00:47	17/17

Table 1. Official results for the Author Identification task for our team *kiprov15*.

Notes

- We chose a C-SVM type of classifier with radial basis function (RBF) kernel. We further set the following parameter values: C = 1, gamma = 0:5, epsilon = 0:001;
- Our solution was among the fastest, but it did not perform very well in terms of AUC and C1.

The project source code can be found on BitBucket:
https://bitbucket.org/StanimirNikolov/pan-author-identication

Future Work

- Experiments with other features such as lists of stopwords, language-specific resources, character n-grams, part-of-speech n-grams, etc.
- Artificially expand the training data by using some of the examples in the known set as questioned examples.