CAPS: A Cross-genre Author Profiling System
Ivan Bilan, Desislava Zhekova
Center for Information and Language Processing, LMU Munich, Germany
ivan.bilan@gmx.de, zhekova@cis.uni-muenchen.de

1. CAPS Overview

2. Preprocessing

- HTML, Bulletin Board Code removal
- normalization of Links ([URL]), Usernames e.g. @username ([USER])
- lemma and POS annotation via the TreeTagger (Schmid, 1994)
- duplicate sample removal:

3. Feature Extraction

- TF-IDF lemmas (uni-, bi- and trigrams), POS-tags (uni-, bi-, tri-, fourgrams), characters (trigrams)
- LDA topic modelling - 100 different topics
- dictionary-based (connective words, emotion words, contractions, family related words, collocations, abbreviations, acronyms, stops)
- POS-based – use of verbs, interjections, adjectives, determiners, conjunctions, plural nouns, lexical measure; includes a more complex F-Measure feature following Heylighen et al. (2002)
- text structure – e.g. type/token ratio, average word length, use of punctuation marks
- stylistic – frequency of use of adjectival and adverbial suffixes – e.g. -ly, -able, -ic, -il, -less, -ous etc.
- readability index – Automated Readability Index, SMOG Readability Formula, Flesch Reading Ease (not effective for the cross-genre setting)
- chi-square term selection for dimensionality reduction

4. Feature Scaling

- The sample length is rescaled relative to the lowest mean length of a text sample throughout all possible writing styles that could be represented in both training and test sets.
- The feature values are divided by this rescaled sample length.

5. Classification

- classify each text sample independent of the others
- classify the author class based on each text sample belonging to the author
- gender (LinearSVC) and age (Multinomial Logistic Regression) are also classified independently

6. Evaluation

6.1 Final PAN16 results for CAPS

- CAPS was ranked third from all 22 teams that participated in PAN16.
- CAPS achieved the second best score – 74.36% accuracy (with the best performing system reaching 75.64%) for gender identification on the official PAN16 test set for English.
- CAPS also proved to be very competitive to the state-of-the-art systems on both PAN15 and PAN14 datasets.
- CAPS achieved 81.69% for gender classification on the PAN15’s English dataset with the best PAN15 participating system reaching a performance of 85.92%.
- CAPS would profit from language specific features for all languages other than English, since currently the feature set is tailored to English.
- CAPS could also profit from text sample-author profile interrelation, which we plan to explore.

6.2 Results on the PAN15 Datasets

6.3 Results on the PAN14 Datasets

7. Conclusion and Future Work

- CAPS was ranked third from all 22 teams that participated in PAN16.
- CAPS achieved the second best score – 74.36% accuracy (with the best performing system reaching 75.64%) for gender identification on the official PAN16 test set for English.
- CAPS also proved to be very competitive to the state-of-the-art systems on both PAN15 and PAN14 datasets.
- CAPS achieved 81.69% for gender classification on the PAN15’s English dataset with the best PAN15 participating system reaching a performance of 85.92%.
- CAPS could profit from language specific features for all languages other than English, since currently the feature set is tailored to English.
- CAPS could also profit from text sample-author profile interrelation, which we plan to explore.