Presented to the
2018 PAN Workshop
on Author Profiling

FEMALE AND MALE LANGUAGE

Jussi Karlgren, Chantal Gratton, Lewis Esposito, Pentti Kanerva

KTH & GAVAGAI; STANFORD; BERKELEY

SEPTEMBER 10-11, 2018 CLEF 2018

Male And Female Language Is Different, How, Why?

By CHANTAL GRATTON AND LEWIS ESPOSITO

There are many claims of gender differences in language, in popular culture, in popular science, news media, and in academic research. Many of these claims, hypotheses, and tentative explanations have little or no empirical support; some are even clearly contradicted by data *(e.g. "women talk more than men", "women hedge more than men", "women use 'like' more than men").* There *are* observable differences between male and female language use in many contexts: understanding how those differences can be explained and which of them can be expected to sustain generally across situations is a research challenge.

Topic is not the interesting difference

By JUSSI KARLGREN

The gender of Twitter authors is from previous studies known to be distinguishable with a precision of around 80% mostly by lexical cues. This is mostly because female and male authors write about different topics. This is a result which cannot be expected to generalise to other situations, genres, and time periods. Topical variation is interesting in itself, but not necessarily a reliable gender identifier, and not likely to be of utility for downstream tasks.

Authors in Semantic Space

By PENTTI KANERVA AND JUSSI KARLGREN

We represent authors as unweighted sums of text vectors in a high-dimensional random indexing semantic vector space. We use the following features to compute the text vectors:

(1) All words used by an author, frequency weighted Some terms (*game* (70% ♂), *win (65% ♂)*, *birthday* (67% ♀)) can fairly be called topical. Others reflect more stylistic or attitudinal usage (*happy* (63% ♀), *love* (67% ♀), *wrong (67% ♂)*, sure *(69% ♂))*. *Terms such as* stuff *(63% ♂)*, while referential, simultaneously reveal volumes about the authors attitude to the topic under treatment.

(2) Part of speech sequences Each sentence was represented by POS (Penn Treebank) label triples.

(3) Constructional and stylistic features of interest First person pronouns (83% ♀); profanity (69% ♂); interjections (*lol, omg, hey, oh, wtf, ...*) (63% ♀); amplifiers (esp. anomaly amplifiers) (64% ♀); hedges (72% ♂); passives (67% ♂); progressives (60% ♀).

(4) Non-topical condition To reduce the topical content of the experimental material, nouns, verbs, and adjectives were replaced with their POS tag. This means adjective comparation, verb tense, and noun number is preserved, but the referential meaning of the word was taken out.

Results

The submitted results were computed using feature sets (2), (3), and (4), removing topical referents. This resulted in an underwhelming accuracy. Some computational issues remain to be addressed but more importantly, some of the underlying hypotheses about gender and language need to be formulated appropriately.

MOST CRUCIAL FUTURE RESEARCH QUESTIONS

1. What hypotheses on gender differences do we assume to hold?
2. Can we assume those differences to hold over time and situation?
3. Are they useful for downstream task?
4. Should the gender be represented explicitly or indirectly, by representing authors?
5. Should the author features be weighted according to discriminative power?
6. How many neighbouring authors should be used to establish the gender of an unknown author?
7. Precision differs across genders. One tentative but likely explanation is that there are more than two styles, and that there are more female styles than male styles among them in this material. How many categories (rather than two genders) would be most appropriate?
8. Can the difference between topical and other referential expressions be determined from the data itself?