Vive la petite différence! Exploiting small differences for gender attribution of short texts

Filip Graliński, Rafał Jaworski, Łukasz Borchmann, Piotr Wierzchoń

Adam Mickiewicz University in Poznań, Poland

Abstract. This article describes a series of experiments on gender attribution of Polish texts. The research was conducted on the publicly available corpus called "He Said She Said", consisting of a large number of short texts from the Polish version of Common Crawl. As opposed to other experiments on gender attribution, this research takes on a task of classifying relatively short texts, authored by many different people.

For the sake of this work, the original "He Said She Said" corpus was filtered in order to eliminate noise and apparent errors in the training data. In the next step, various machine learning algorithms were developed in order to achieve better classification accuracy.

Interestingly, the results of the experiments presented in this paper are fully reproducible, as all the source codes were deposited in the open platform *Gonito.net*. Gonito.net allows for defining machine learning tasks to be tackled by multiple researchers and provides the researchers with easy access to each other's results.

Keywords: gender attribution, text classification, corpus, Common Crawl, reseach reproducibility

1 Introduction

Gender classification of written language has been a subject of linguistic studies for decades now. Note, for instance, a ground-breaking book [8], describing characteristic features of women's language. In more recent years, linguists and socio-linguists researching this subject have been aided by statisticians, see for instance: [11] and [10]. Furthermore, development of social media opened a possibility of building large-scale text corpora, annotated with meta information regarding the author's gender. This resulted in numerous projects aimed at automatic gender classification based on training data acquired from the Web. The annotated text resources used to build the corpora were often taken from blogs, e.g. [10].

However, we believe that gender annotated corpora scraped from the selected Web sources are prone to some flaws. Firstly, they may suffer from thematic bias – women tend to write about different subjects than men, see [11]. Secondly, there might be significantly more text written by authors of either gender. This is due to the fact that if a corpus is a collection of gender annotated items such as blog entries, these items might differ significantly in length. And lastly – the volume of the corpus may not be sufficient for reliable statistical analysis. All these flaws result in a situation, where gender

classification is heavily biased by the subject women and men write about and not the language they are using.

In this paper we describe an experiment on gender classification of texts based on a custom built corpus which tries to avoid the flaws mentioned above. The corpus itself is publicly available. Gender classification mechanism is designed using statistical algorithms. It is also worth noting that the results of the described experiment are fully reproducible, as the research was conducted on the *Gonito.net* platform.

Section 2 of the paper describes similar experiments on automatic gender classification. The corpus used in our experiment is described in Section 3. Section 4 describes the classification mechanism itself, while Section 6 lists the conclusions.

2 Related work

2.1 Argamon et al.

Work on gender classification has been described in [1]. Its authors carried out an experiment on data from the British National Corpus. The corpus contained documents labeled as being authored either by a woman or a man. Total volume of the corpus reached 604 documents, 302 for each gender, coming from different fiction and non-fiction genres (such as science, arts, commerce). Average number of words per document was 42 000 words, which brought the overall word count of the corpus to over 25 million words. Furthermore, the corpus was POS-tagged.

The authors used this data to train a custom version of an algorithm referred to as EG, described in [7]. For the sake of the training a feature set of 1000 features was prepared. It consisted of 467 function words and over 500 most frequent part-of-speech n-grams.

The experiment determined that female and male languages differ in terms of the use of pronouns and certain types of noun modifiers – females tend to use more pronouns, while males use more noun specifiers. Generally speaking, the authors describe female writing style as “involved” and male writing as “informative”.

The work by Argamon et al. laid ground for research on automatic statistical gender classification. Some methods used by the authors, specifically the classification algorithm and feature selection procedure, could be and were improved in subsequent experiments.

2.2 Gender classification of blog entries

In the first decade of the 21st century many solutions to the automatic gender classification problem exploited textual data from blog entries. Main advantages of blog corpora were considerable sizes and relatively reliable gender annotations (thanks to metadata about the authors of the blogs). Disadvantages, on the other hand, included significant thematic bias.

Nevertheless, a large number of classification algorithms was proposed. They relied on features such as content words, dictionary based content analysis or part-of-speech tags, see for instance [12], [2] or [14].

The method proposed by [10] is based on two novel ideas. The first is a new feature group – variable length POS sequence patterns, while the second is a novel feature selection algorithm. The authors report the accuracy of their system to reach 88.56%, while the previously mentioned systems by [12], [2] and [14] only score 77.86%, 79.63% and 68.75% respectively.

2.3 Other interesting solutions

The work presented in [11] explores different statistical approaches to gender classification. Interestingly, it is one of very few and possibly the first publication to acknowledge the problem of topic bias. The authors tried to remove the topic and genre bias completely and experiment with an evenly balanced corpus. Furthermore, they take the challenge of cross-topic and cross-genre gender classification. They try to find gender-specific features in language alone, with the help of stylometric techniques. After performing the experiment, the authors concluded that previous results in gender classification might have been overly optimistic due to the topic bias issues. Among various machine learning techniques used to perform classification on the bias-free corpus, simple character based models surprisingly proved the most robust.

Another interesting approach is presented in [3]. The authors managed to reproduce the results of [10] by using neural networks. Considering that [3] present merely preliminary research results, the deep learning approach may have a significant potential in the gender classification task.

3 Corpus used in the experiment

3.1 Original corpus

The corpus used in our experiment is based on the "He Said She Said" Polish corpus (referred to as HSSS) described in detail in [5]. While the corpora used in research on gender-related differences have been based on *metadata* supplied manually, for example, statements of gender by text authors, or gender tags added manually by the corpus creators, the creators of the HSSS corpus propose a different approach: to look in the text itself for *gender-specific first-person expressions*. Not all languages have these (almost none of them to be found in English), but they are quite frequent in the Slavic languages and they also occur, to some extent, in the Romance languages. The procedure of preparing the HSSS corpus was to take Common Crawl-based Web corpus[1] of Polish [4] and grep for lines containing gender-specific first-person expressions to create a gender-specified subcorpus. The procedure was applied to Polish, a language in which the frequency of gender-specific first-person expressions is particularly high. The resulting HSSS corpus contains short fragments of texts, annotated as authored by a female or a male, accompanied by the URL pointing to the place where the fragment was published. Texts came from 92 834 different websites, by which we conclude that the HSSS corpus is likely to contain texts authored by 50-100K different people.

[1] `http://data.statmt.org/ngrams/raw/`

The HSSS corpus differs from other gender annotated corpora as it contains short fragments of texts coming from a wide variety of thematic domains. Other researchers performed their experiments on narrow domain corpora (such as thematic blogs) which could make the classification task significantly easier.

3.2 Filtering the corpus

Even though the HSSS corpus is an interesting linguistic dataset, it is not flawless. For example, it is not free from the topic bias problem. Most male texts come from websites presumably visited by men (such as technical, engineering or sport portals), while majority of female texts comes from more stereotypically women-like portals (e.g. sites about pregnancy). For the sake of our experiment, the original HSSS corpus underwent filtering procedures. Firstly, male/female balance was introduced by using the following principle – within fragments coming from one web domain (website) there should be an equal number of female and male texts. To ensure this, within each website, all the fragments that made one gender outnumber the other, were discarded. For example, if the corpus contained 3 male and 100 female texts from a parenting portal, only the 3 male and the first 3 female texts were taken into consideration. And if a site contained texts of only one gender, it was discarded completely. This procedure was performed in order to reduce the effects of the topic bias problem.

Another type of problems in the HSSS corpus are so called "leaks". We refer as leaks to gender-specific expressions which were not turned into male versions during the creation of the HSSS corpus. Examples of leaks we managed to filter out were some forms of adjectives and participles, which in Polish have genders (e.g. "green" in Polish is "zielona" in the feminine form but "zielony" in masculine).

Errors were also found in annotation of texts, which contained gender-specific first-person expressions, but the gender of the author could not be determined based on these expressions alone. For example, the name of the popular TV series "How I Met Your Mother" in Polish is "Jak poznałem waszą matkę" and the word "poznałem" is a male-specific first-person verb. Therefore, in the HSSS corpus we found reviews of the series written by females automatically annotated to be written by males. In order to overcome this problem, we filtered out all the texts containing the name of this and a few other titles causing the same problem.

Some noise was also observed in the HSSS corpus. It included texts which were clearly not written by Polish native speaker, but were a result of a very poor, probably automatic translation. In some of these texts we observed an oddly high frequency of the term "negacja logiczna" (logical negation), which we attributed to a machine translation mechanism, trying to translate the simple word "not" into Polish. As a result, we decided to discard all the texts containing the phrase "negacja logiczna".

3.3 Train, dev and test sets

The training, development and test sets for the classification task were not selected from the corpus in a completely random manner, but instead the division respected the websites, i.e. one set of websites was denoted the training set, another – development set and the third – test set. Thus, the classification task became even more challenging,

as the mechanism is to be trained on different thematic domains than it would be tested on. Importantly, all the gender-specific expressions identified during the creation of the corpus were turned to male versions, even in the female texts.

For these reasons, a classifier prepared for this task should in fact recognize gender-specific features in female and male texts and not base its judgements on simple topical differences.

4 Classification problem formulation and solution

The modified version of the HSSS corpus was turned into a text classification challenge on Gonito.net (which is an open platform for research competition, cooperation and reproducibility). The training and developments sets as well as the input for the test set are publicly available there as a Git repository.[2] Accuracy was chosen as an evaluation metric.

The accuracy for the null model (always returning male or female authors) is of course 50% {86dd91}.

(For each classification method described in this paper, the output files and all the source codes are available as submissions to the Gonito.net platform. Git commit SHA1 prefixes are always given here in curly brackets. In the electronic edition of this paper, the above commit number prefix is clickable as `http://gonito.net/q/86dd914ad99a4dd77ba1998bb9b6f77a6b076352` – alternatively, the ID `86dd91` could be entered manually at `http://gonito.net/q` or directly pasted into a link (`http://gonito.net/q/86dd91`) – and leads to a submission summary with an URL to a publicly accessible Git repository.)

Using a hand-crafted regular expression for known "leaked" feminine first-person expressions has an accuracy of only 51.03% {f98f7d}

4.1 Logistic regression with Vowpal Wabbit

A logistic regression model was trained with the Vowpal Wabbit open-source learning system [9]. Lower-cased tokens (no other normalisation was done) were used as features. With this simple set-up, we obtained accuracy of 67.54% {36ff5b}. Using a simple sigmoidal feedforward network with 6 hidden units yields a slightly better result (68.32%, {d96cfc}), whereas adding prefixes {8f9557} or suffixes {8e0e25} as features does not improve much.

4.2 Language models

We used KenLM language modelling toolkit [6] to create two separate 3-gram language models: one for male texts and one for female texts. During the classification, the class with the higher probability is simply chosen. Even though this method was substantially different from the Vowpal Wabbit classifier the results were quite similar (67.98%,

[2] `git://gonito.net/petite-differance-challenge.git`

{85317f}). That's why we decided to combine both methods with another layer of neural network. This way, the best result so far was achieved: 71.06% {12b6d8}.[3]

4.3 Morphosyntactic tags only

In order to completely avoid topical imbalance in vocabulary, we also trained classifiers using only morphosyntactic tags. TreeTagger [13] with a Polish model was used to tag the texts, i.e. each word was assigned its part of speech and other tags such as person, number, gender, tense.

A classifier was trained with Vowpal Wabbit with the following features:

- part-of-speech 3-grams,
- morphosyntactic tags for a given word treated as one feature,
- morphosyntactic tags for a given word treated as separate features.

The accuracy on the test set for this classifier was 59.17% {b4e142}.

A similar, but slightly better result (60.58%) was obtained with 6-gram language models trained on morphosyntactic tags {edfdfc}.

We find these results quite satisfactory as they were obtained on short tags using only morphosyntactic tags.

5 Back to the corpus

The models described in Section 4 were analysed to find the most distinctive features. It turned out that a large number of them are actually "leaks" (expressions which should have been identified as gender-specific and normalised when HSSS was created but were not):

- gender-specific inflected forms of verbs absent from the lexicon of inflected forms,
- frequent inflected forms written with a spelling mistake (in particular, without a diacritic),
- verb *być* (*be*) with a longer adjective phrase (e.g. *jestem bardzo zadowolony/ zadowolona* = *I am very glad*),
- the word *sam*/*sama* (= *myself*, which has a different masculine and feminine form and which could mean *himself*/*herself*),

Some other problems were also identified in the balanced corpus: automatically generated spam not filtered out and gender-specific forms found in the film titles.

[3] The output files and source codes are available for inspection and reproduction at Git repository `git://gonito.net/petite-difference-challenge`, branch `submission-00085`.

Acknowledgements

Work supported by the **Polish Ministry of Science and Higher Education** under the **National Programme for Development of the Humanities**, grant 0286/NPRH4/H1a/83/2015: "50 000 słów. Indeks tematyczno-chronologizacyjny 1918-1939".

6 Conclusions

The paper presented research on gender classification performed on a corpus created in a different way than in the work described so far. Gender annotations in the corpus were obtained by exploiting certain linguistic features of the Polish language, rather than by relying on meta-data. Furthermore, for the needs of the experiments the corpus was balanced by websites, in order to minimize the effect of gender and topic bias. Training data prepared in this manner is unique at least for the Polish language.

Developed classification algorithm achieved a maximum gender prediction accuracy of 71.06%. The algorithm relied on language modelling (KenLM toolkit) and the Vowpal Wabbit machine learning system. The two methods were combined using a neural network. Classification results revealed some noise in the training data that can and should be filtered out. Nonetheless, the prediction accuracy of above 70% can be viewed as a success, considering the competitiveness of the task.

Future work plans include further filtering of the corpus based on the information obtained during the gender classification task. Classification algorithms themselves will also be further optimized. This work will be facilitated by the Gonito.net platform.

References

1. Argamon, S., Koppel, M., Fine, J., Shimoni, A.R.: Gender, genre, and writing style in formal written texts. TEXT 23, 321–346 (2003)
2. Argamon, S., Koppel, M., Pennebaker, J.W., Schler, J.: Mining the blogosphere: Age, gender and the varieties of self-expression. First Monday 12(9) (2007), http://pear.accc.uic.edu/ojs/index.php/fm/article/view/2003
3. Bartle, A., Zheng, J.: Gender Classification with Deep Learning (2015)
4. Buck, C., Heafield, K., van Ooyen, B.: N-gram counts and language models from the common crawl. In: Proceedings of the Language Resources and Evaluation Conference. Reykjavk, Icelandik, Iceland (May 2014)
5. Graliński, F., Borchmann, L., Wierzchoń, P.: "He Said She Said" – Male/Female Corpus of Polish. In: Proceedings of the Language Resources and Evaluation Conference LREC 2016 (2016)

6. Heafield, K.: KenLM: faster and smaller language model queries. In: Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation. pp. 187–197. Edinburgh, Scotland, United Kingdom (July 2011), http://kheafield.com/professional/avenue/kenlm.pdf
7. Kivinen, J., Warmuth, M.K.: Additive versus exponentiated gradient updates for linear prediction. In: Proceedings of the Twenty-seventh Annual ACM Symposium on Theory of Computing. pp. 209–218. STOC '95, ACM, New York, NY, USA (1995), http://doi.acm.org/10.1145/225058.225121
8. Lakoff, R.: Language and woman's place. Harper colophon books, Harper & Row (1975), https://books.google.pl/books?id=0dFoAAAAIAAJ
9. Langford, J., Li, L., Zhang, T.: Sparse online learning via truncated gradient. In: Advances in neural information processing systems. pp. 905–912 (2009)
10. Mukherjee, A., Liu, B.: Improving Gender Classification of Blog Authors. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. pp. 207–217. EMNLP '10, Association for Computational Linguistics, Stroudsburg, PA, USA (2010), http://dl.acm.org/citation.cfm?id=1870658.1870679
11. Sarawgi, R., Gajulapalli, K., Choi, Y.: Gender attribution: Tracing stylometric evidence beyond topic and genre. In: Proceedings of the Fifteenth Conference on Computational Natural Language Learning. pp. 78–86. CoNLL '11, Association for Computational Linguistics, Stroudsburg, PA, USA (2011), http://dl.acm.org/citation.cfm?id=2018936.2018946
12. Schler, J., Koppel, M., Argamon, S., Pennebaker, J.: Effects of Age and Gender on Blogging. In: Proc. of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs (Mar 2006)
13. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: Proceedings of the international conference on new methods in language processing. vol. 12, pp. 44–49 (1994)
14. Yan, X., Yan, L.: Gender classification of weblog authors. In: In Proceedings of the AAAI Spring Symposia on Computational Approaches. pp. 27–29 (2006)