Author Identification

Uras Mutlu & E.Burak Dündar

2017

Boğaziçi University

Introduction

- Problem:
  - Text -----> Author
  - Gender,age
  - Text to Text similarity
- Features: word frequencies, characters
- Plays of Shakespeare,19th century
- “Federalist Papers”
  - The most influential work
- Stylometric features, until the 1990s

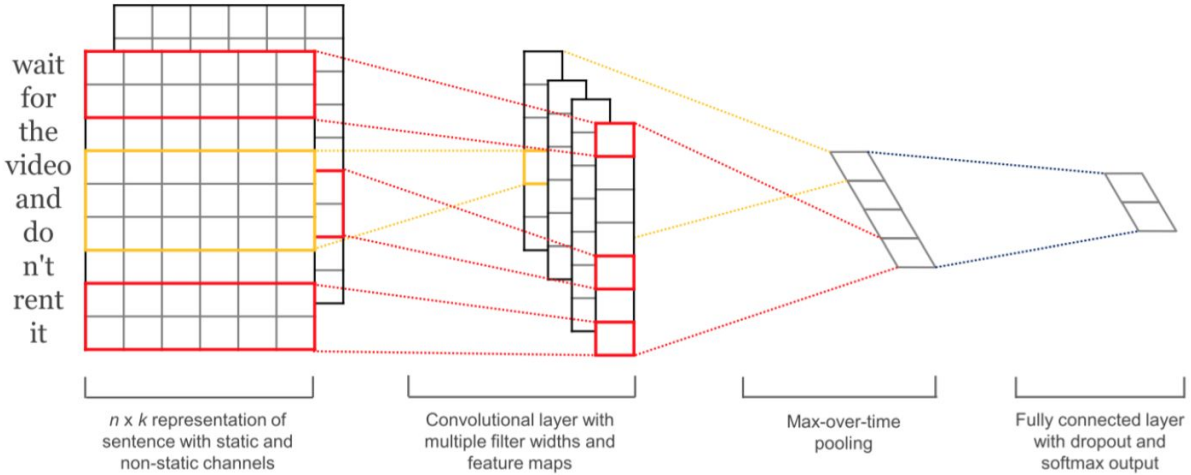
Features

- Character features ? ^ ; -
  - Character n-grams
- Lexical features
  - Bag of words, word n-grams, frequencies of words, functional words
- Syntactic features
  - Syntactic patterns for capturing style
- Semantic features
  - Meaning of a text
- Application-specific features
  - Authors on same theme, different keywords

Methods

- Bayesian approach
  - $$P(v|a_1,\ldots,a_n) = \frac{P(v)\cdot P(a_1,\ldots,a_n|v)}{P(a_1,\ldots,a_n)}$$
- Compression approach:
  - Similarity= C(text+unseenText) - C(text)
- SVM outperforms Decision Tree and Neural Netwoks
  - 70-90% accuracy
- Convolutional Neural Network
  - State-of-the-art

Methods - State of the art

- 

EXAMPLE USE-CASES

- Intelligence
  - Authorship identification for online communication messages of terrorist organizations.
- Crime investigation
  - Verifying the authenticity of digital evidence. Such as e-mails, suicide notes, electronic journals.
- Cyber-crimes
  - Identifying the developers who wrote malicious software by capturing their coding style.

Tools and Systems

- Turnitin,1997
  - online plagiarism prevention system
- Online authorship attribution system
  - AICBT: Canadian R&D company
  - Supports 2 Authors
  - NLTK and Sci-kit Python Libraries

Results

Model	Emails		IMDb	Blogs		Twitter		reddit		Ave
Number of authors	10	50	62	10	50	10	50	10	50	-
SVM+Stems	83.0	72.9	88.3	36.5	29.7	91.7	81.3	35.1	21.2	60.0
SCAP	83.1	69.0	**94.8**	48.6	41.6	91.3	82.5	46.5	30.3	65.3
Imposters	52.0	32.9	76.9	35.4	22.6	71.4	52.5	32.1	16.3	43.6
LDAH-S	82.0	39.1	72.0	52.5	18.3	90.0	38.3	43.0	14.2	49.9
CNN-word	89.7	82.5	84.3	59.0	43.0	96.2	80.5	36.2	20.1	65.7
CNN-word-word	90.3	81.0	82.0	58.8	41.4	95.7	79.3	39.6	18.3	65.2
CNN-char	**93.1**	**88.1**	91.7	59.7	48.1	**97.5**	**86.8**	**58.8**	**37.2**	**73.4**
CNN-word-char	90.3	84.9	90.2	**61.2**	**49.4**	97.2	84.9	53.1	27.7	70.1
CNN-word-word-char	90.1	83.8	88.4	**61.2**	47.0	95.9	84.0	56.1	27.0	70.4

Conclusion

- In last 15-20 years, with the advancements in information retrieval, machine learning, and natural language processing, there has been a lot of studies involving authorship identification.
- Most systems work well with closed sets and small number of authors, but there is still some progress needed for identifying the works of large number of authors.
- Better NLP tools that give less noisy data when analyzing texts syntactically and semantically can be very helpful to capture writing styles and achieve better accuracies.

References

1. Character-level and Multi-channel Convolutional Neural Networks for Large-scale Authorship Attribution. Retrieved from: https://arxiv.org/pdf/1609.06686.pdf
2. Authorship Analysis Studies: A Survey. Retrieved from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.429.570&rep=rep1&type=pdf
3. Severyn, Aliaksei, and Alessandro Moschitti. "Twitter sentiment analysis with deep convolutional neural networks." *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2015.
4. Stamatatos E. (2009) A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology*, 60(3), pages 538-556.
5. Abbasi, Ahmed, and Hsinchun Chen. "Applying authorship analysis to extremist-group web forum messages." *IEEE Intelligent Systems* 20.5 (2005): 67-75.
6. Chaski, Carole E. "Who's at the keyboard? Authorship attribution in digital evidence investigations." *International journal of digital evidence* 4.1 (2005): 1-13.
7. Frantzeskou, Georgia, Stefanos Gritzalis, and Stephen G. MacDonell. "Source code authorship analysis for supporting the cybercrime investigation process." *Handbook of Research on Computational Forensics, Digital Crime, and Investigation: Methods and Solutions* (2004): 470-495.

Thank you!