Large-scale Author Verification
Temporal and Topical Influences

Michiel van Dam and Claudia Hauff
Web Information Systems
Delft University of Technology
the Netherlands

1. Introduction

Author verification task:
Has author X (with known reference texts $K_1..K_i$) written text T?

Limitations of publicly available corpora such as PAN [1]:
* very **small** (a few dozen test cases),
* **few languages** covered,
* texts $K_1..K_i$ and T are **long**, and **matched** in time & genre

Goals:
* pipeline to create large-scale corpora **automatically**,
* investigate **factors** of **time** and **topic** on the accuracy of author verification

2. Research questions

Topic hypothesis:
Short and topically diverse reference documents make the verification problem more difficult.

Common **topic-based assumption**:
Two texts about similar topics are biased towards being recognized as from the same author.

Temporal hypothesis [3]:
Authors' writing style changes over time. Texts written within a short period of time are more aligned than texts written at very different times.

40% English Wikipedia
2,891 authors with 4+ comments

Topical similarity

Test set	# Test cases	Accuracy
All annotated	3950	0.598
Similar	316	0.642
Different	2406	0.580
Matching	1544	0.618
Nonmatching	1240	0.588

Semantic similarity of respective Wikipedia articles; balanced test sets.

Temporal similarity

Test set	# Test cases	Accuracy
All annotated	1368	0,633
Similar (<1wk)	684	0,665
Different (>3yr)	684	0,588
Matching	684	0,684
Nonmatching	684	0,580

Temporal differences based on the comments' timestamps.

3. Corpus

Required:
* many authors,
* many topics,
* extended period of time

Wikipedia Revision History
↓
Wikipedia Talkpages
↓
Comment Extraction
(2,500-10,000 chars, denoisified)
↓
Test Set Creation
(1 reference text per test case)

`Similar` K_1&T	`Different` K_1&T
`Author match` K_1&T	`No author match` K_1&T

4 test cases per author

4. Results

Implementation: *Common N-gram approach* [2]; writing style profile as vector of character 3-grams.

Topical hypothesis could be verified (`Different < Similar`).

Common **topic-based assumption** could not be verified (`Similar` achieves highest acc.).

topic words act as style markers

Temporal hypothesis could be verified (`Similar > Different`).

time is an important dimension

Additional experiments in French, German, Spanish and Greek: the trends hold!

5. Future Work

* In-depth analysis across languages
* Comparison of different author verification algorithms

References

[1] PAN: http://pan.webis.de/ (benchmark series starting in 2009, still ongoing)

[2] V. Keselj, F. Peng, N. Cercone, and C. Thomas. N-gram-based author profiles for authorship attribution. Pacic Association for Computational Linguistics, pages 255-264, 2003.

[3] F. Can and J. M. Patton. Change of Writing Style with Time. Computers and the Humanities, 38(1):6182, 2004.