Plagiarism - PAN 2011

PA164 Machine learning and natural language processing

Miroslav Hlaváček
(podzim 2012)

Reasons for new evaluation framework

- little papers focused on text documents plagiarism
- usually dealing with a small corpus (most often 10^3)
- lack of objective and general performance evaluation methods
- availability – authorship issues
- lack of focus on information retrieval

Plagiarism cases in text

- long vs. short
- unobfuscated vs. obfuscated
- obfuscated – simulated vs. artificial (both has advantages and disadvantages)
- PAN-PC-10 corpus
  - intrinsic (30%) vs. external
  - intra-topic vs. inter-topic.

Obfuscation examples

Original Text
The quick brown fox jumps over the lazy dog.

Manual Obfuscation (by a human)
Over the dog which is lazy jumps quickly the fox which is brown.
Dogs are lazy which is why brown foxes quickly jump over them.
A fast auburn vulpine hops over an idle canine.

Random Text Operations
over The. the quick lazy dog <*context word*> jumps brown fox
over jumps quick brown fox The lazy. the
brown jumps the. quick dog The lazy fox over

Semantic Word Variation
The quick brown dodger leaps over the lazy canine.
The quick brown canine jumps over the lazy canine.
The quick brown vixen leaps over the lazy puppy.

POS-preserving Word Shuffling
The brown lazy fox jumps over the quick dog.
The lazy quick dog jumps over the brown fox.
The brown lazy dog jumps over the quick fox.

PAN-PC-10 corpus overview

Document Statistics		
Document Purpose		
source documents		50%
suspicious documents		
– with plagiarism		25%
– w/o plagiarism		25%
Intended Algorithms		
external detection		70%
intrinsic detection		30%
Plagiarism per Document		
hardly	(5%-20%)	45%
medium	(20%-50%)	15%
much	(50%-80%)	25%
entirely	(>80%)	15%
Document Length		
short	(1-10 pp.)	50%
medium	(10-100 pp.)	35%
long	(100-1000 pp.)	15%

Plagiarism Case Statistics		
Topic Match		
intra-topic cases		50%
inter-topic cases		50%
Obfuscation		
none		40%
artificial		
– low obfuscation		20%
– high obfuscation		20%
simulated (AMT)		6%
translated ({de,es} to en)		14%
Case Length		
short	(50-150 words)	34%
medium	(300-500 words)	33%
long	(3000-5000 words)	33%

Evaluation – measures

- S and R sets
- precision (micro vs. macro version)
- recall (micro vs. macro version)
- granularity [1,|R|]

$$gran(S,R) = \frac{1}{|S_R|} \sum_{s \in S_R} |R_s|$$

$$S_R = \{s \mid s \in S \wedge \exists r \in R : r \text{ detects } s\}$$

$$R_s = \{r \mid r \in R \wedge r \text{ detects } s\}.$$

- plagdet

$$plagdet(S,R) = \frac{F_\alpha}{\log_2(1 + gran(S,R))}$$

Evaluation - results

- testing corpus for each year
- plagdet score
- winner – 500,- Euro by Yahoo!

External Plagiarism Detection Performance

Rank	Plagdet	Recall	Precision	Granularity	Participant
1	0.5563430	0.3965569	0.9368736	1.0022487	J. Grman and R. Ravas SVOP Ltd., Slovakia
2	0.4153395	0.3376925	0.8119867	1.2167900	C. Grozea* and M. Popescu° *Fraunhofer Institute FIRST, Germany °University of Bucharest, Romania
3	0.3468605	0.2257937	0.9116530	1.0611984	G. Oberreuter, G. L'Huillier, S A. Ríos, and J.D. Velásquez Universidad de Chile. Chile

Intrinsic Plagiarism Detection Performance

Rank	Plagdet	Recall	Precision	Granularity	Participant
1	0.3254817	0.3397965	0.3123243	1.0000000	G. Oberreuter Universidad de Chile, Chile
2	0.1679779	0.4279112	0.1075817	1.0329386	M. Kestemont, K. Luyckx, and W. Daelemans University of Antwerp, Belgium
3	0.0841286	0.1277831	0.0664302	1.0549085	N. Akiva Bar Ilan University, Israel

Sources

- http://www.webis.de/research/events/pan-11
- http://www.uni-weimar.de/medien/webis/publications/papers/stein_2010p.pdf

Thank you for your attention.