Miroslav Hlaváček, 359995, podzim 2012

Plagiarism - PAN 2011

**International Workshop on Plagiarism Analysis,
Authorship Identification, and Near-Duplicate Detection (PAN)**

"PA164 Machine learning and natural language processing" course essay

Abstract:
Web page (source 1) dedicated to the part of PAN 2011 dealing with plagiarism contains information about text plagiarism detection evaluation framework designed especially for purposes of PAN as well as results of participants of 2011's year.
The PAN plagiarism detection "framework" consists of corpus designated as learning corpus for plagiarism detection and performance measures providing objective comparison of plagiarism detectors.
Reasons for new corpus PAN-PC-10 (in comparison with academic papers available in year 2010):

- little papers focused on text documents plagiarism (most papers focused on plagiarism in code)
- most papers refer to a small corpus (most often 10^3 documents because of local collection of documents)
- lack of objective and general evaluation methods
- availability – authorship issues (need to have approval from both author and plagiarist)
- lack of focus on information retrieval (plagiarism case should be detected only once and in full length)

The methods of building corpus and incorporating plagiarism cases were discussed. Plagiarism cases can be then divided by several points of view – mainly long vs. short, intra-topic vs. inter-topic (documents are clustered by several topics), intrinsic (does not use external knowledge and tries to identify discrepancies in style within a suspicious document) vs. external, unobfuscated ("copy paste style") vs. obfuscated (plagiarized passage has the same meaning although different words or word ordering is used). Several methods for generating obfuscated plagiarism were used (examples can be found in source 2). These methods can be divided into:

- simulated – text is rewritten by human who is paid for the task (based on Amazon's Mechanical Turk project)
  (*problems:* is necessary to determine right amount of cash per task; these plagiarists were usually well educated, …)
- artificial – generated by computer (Random text operations, Semantic word variation, POS-preserving word shuffling)
  (*problems:* connected with the fact that computer does not understand the text)

Overview of PAN-PC-10 corpus can be found in source 2.
For evaluation were proposed new evaluation metrics:

- granularity – determines whether plagiarism case is detected only once (best possible value) or more often (number of detected plagiarism cases (R set) denoting the one (and only) real plagiarism case (S set) is the worst possible value of granularity); therefore granularity

determines plagiarism detection performance for the information retrieval part of plagiarism detection task

$$gran(S,R) = \frac{1}{|S_R|} \sum_{s \in S_R} |R_s| \qquad S_R = \{s \mid s \in S \wedge \exists r \in R : r \text{ detects } s\}$$
$$R_s = \{r \mid r \in R \wedge r \text{ detects } s\}.$$

- **plagdet** – new performance measure which combines recall, precision and granularity (where F is F-measure for precision and recall with parameter α and logarithm of granularity is used to decrease its influence to a reasonable level).

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

Web page (source 1) shows results determined by plagdet scores for both tasks of intrinsic (30% of corpus) and external plagiarism detection. An evaluation corpus is different for each year of PAN which means that winning performance may vary as well. Winner was awarded by money price of 500,- Euro. The best results for year 2011 are:

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:

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