





Text Reuse and Plagiarism Detection

Muhammad Sharjeel

PhD Student

SCC - Lancaster University

 @msharjeel  s.muhammad6@lancaster.ac.uk

School of Computing
& Communications



Introduction – Basic Concepts

- **Text reuse** is the process of creating [new texts] using [existing ones]
 - *Existing text = Source or original text*
 - *New text = Suspicious or derived or plagiarised text*
- Derived text(s) can be word-by-word copied, paraphrased or reused the idea of the existing text(s)
- The amount of text reused also varies from small phrases, sentences, paragraphs or even entire documents
- Text reuse has two types, mono-lingual and cross-lingual
- Mono-lingual text reuse: Source and derived texts are in the same language
- Cross-lingual text reuse: Both texts are from different languages

Introduction – Basic Concepts

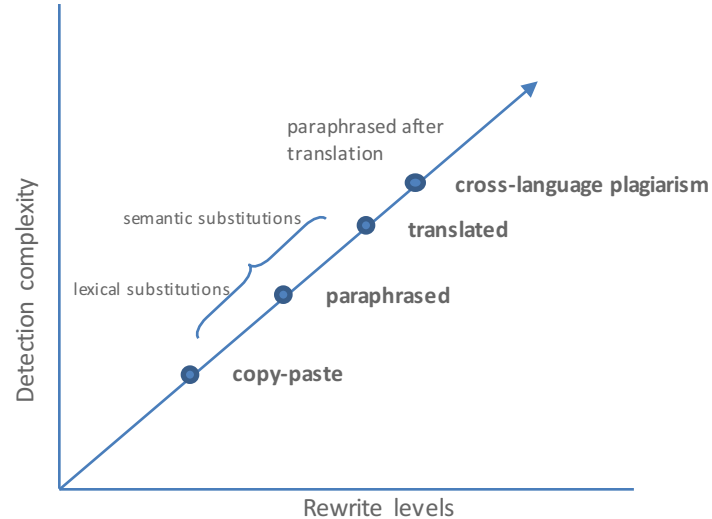
- **Plagiarism** by definition is;
- “to steal and pass off the ideas or words of another as one’s own”
- “giving incorrect information about the source of a quotation”
(from www.plagiarism.org and IEEE)
- Text plagiarism is a special case of text reuse, and similarly may cross language boundaries, called cross-language plagiarism

Introduction – Basic Concepts

- S1: Copying words or ideas from someone else without giving credit
- P1: Copying the words and ideas from someone else's text without giving credit
- P2: Copiar las palabras o ideas de alguien más sin darle crédito
- P3: کریڈٹ دیئے بغیر کسی اور کے الفاظ یا خیالات کو نقل کرنا
- NP: Changing words but copying the sentence structure of a source without acknowledgement
 - *S1 is a source text*
 - *P1 is a mono-lingual plagiarism example*
 - *P2 is a cross-lingual (English-Spanish) plagiarism example*
 - *P3 is a cross-lingual cross-script (English-Urdu) plagiarism example*
 - *NP is non-plagiarism example*

Introduction – Types of plagiarism

- copy-paste
- paraphrase
- idea
- source-code
- translated



Introduction – Why detect text reuse and plagiarism?

- Plagiarism is not only a serious academic offence but now exists across genres
- 61% of Spanish students admitted copying fragments from the Web, a large number are using translated text
- 80% of professors found plagiarism in their students works
- Two Portuguese journalists, New York Times columnist and a Time magazine journalist, admitted plagiarising their news articles
- Two German Ministers, Guttenberg and Schavan (2011, 2013) and a Romanian PM Ponta (2012) found guilty of plagiarised PhD dissertations
- In media, songs, lyrics and stories are reused without citing the corresponding source

Introduction – Why is plagiarism detection interesting?

- Plagiarism is considered as one of the biggest problems in publishing, science, and education
- Text plagiarism is observed at an unprecedented scale with the advent of the World Wide Web (billions of texts, source codes, images, sounds, and videos easily accessible)
- The manual analysis of text with respect to plagiarism becomes infeasible on a large scale
- Plagiarism detection, the automatic identification of plagiarism and the retrieval of the original sources, is researched and developed as a possible countermeasure to plagiarism

Introduction – text reuse and plagiarism detection task

- Automatic text reuse and plagiarism detection task is dissected into two tasks
- ***Extrinsic plagiarism detection***: compare a suspicious text against a collection of possible source(s) to identify the passages that have been copied
- ***Intrinsic plagiarism detection***: To check whether the entire (suspicious) text was written by a single author

Introduction – text reuse and plagiarism detection task

Intrinsic plagiarism detection

- Let d_q be presumably written by A . Determine whether the contents in d_q were actually written by A . If not, extract those fragments (potentially written by a A_0)

Extrinsic plagiarism detection

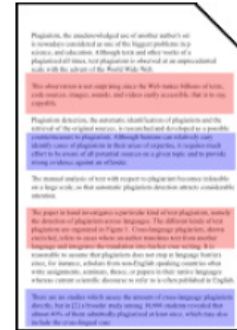
- Let d_q be a suspicious text and D be a set of potential source texts. Determine whether d_q contains borrowed text from a specific $d \in D$.

Cross language extrinsic plagiarism detection

- $d_q \in L, d' \in L' (L \neq L')$

Intrinsic plagiarism detection

- An expert is often able to detect plagiarism by reading a document
- Insertion of text from a different author into d_q causes style and complexity irregularities
- Quantification can be made by measuring;
 - Vocabulary richness (types/tokens ratio)
 - Basic statistics (avg. sentence length, avg. word length)
 - n -grams profiles (character level statistics)



-
- The background of the slide is a collage of numerous overlapping document icons, each representing a page from a book or a set of guidelines. In the center, one document icon is highlighted with a black border and contains the following text:
- National Curriculum Framework for School Education**
- This curriculum framework provides the structure and content for the school curriculum. It is designed to ensure that all students receive a high-quality education that is relevant to their lives and the world around them. The framework is based on the National Education Policy (NEP) 2020 and is intended to guide the development of school curricula across all levels of education.
- The framework is organized into three main sections: **Foundational Stage**, **School Education**, and **Higher Education**. Each section contains a set of guidelines that outline the key concepts, skills, and values that should be taught at that level. The framework is designed to be flexible, allowing schools to adapt the curriculum to their own needs and the needs of their students.
- The framework is also designed to be inclusive, ensuring that all students, regardless of their background or abilities, can access and benefit from the education. The framework is a living document, and it will be updated as needed to reflect changes in the field of education.

- [Potthast et al., 2009]

Text reuse and plagiarism corpora

- Three main type
 - Real, Simulated, Artificial
- METER (**ME**asuring **TE**xt **R**euse)
- PAN Corpora (**P**lagiarism, **A**uthorship and **N**ear Duplicate)
- SAC (**S**hort **A**nswer **C**orpus), P4P (**P**araphrase **4** **P**lagiarism)

Text reuse and plagiarism corpora

- Example source-suspicious pair from METER Corpus

Source: The waterlogged conditions that ruled out play yesterday still prevailed at Bourda this morning, and it was not until mid-afternoon that the match restarted. Less than three hours play remained, and with the West Indies still making their first innings reply to England's total of 448, there was no chance of a result. At tea the West Indies were two for 139.

Rewrite: Waterlogged conditions ruled out play this morning, but the match resumed with less than three hours play remaining for the final day. The West Indies are making a first innings reply to England's total of 448. At tea the West Indies were 139 for two, but there's no chance of a result.

Text reuse and plagiarism @UCREL (Lancaster)

- UPPC – **U**rdParaphrase **P**lagiarism **C**orpus

http://www.lrec-conf.org/proceedings/lrec2016/pdf/364_Paper.pdf

- 20 source, 140 suspicious documents (75 paraphrased, 65 non-paraphrased)
- Simulated plagiarism cases

- COUNTER – **C**orpus of **U**rdNews **T**ext **R**euse

- Real text reuse cases from the field of journalism
- 600 source, 600 suspicious documents (135 Wholly Derived, 288 Partially Derived, 177 Non Derived)

<http://ucrel.lancs.ac.uk/textreuse/index.php>

Similarity estimation methods

- Syntactic methods
 - computing similarity based on surface-level
 - n-gram Overlap, Vector Space Model, Longest Common Subsequence , Greedy String Tiling
- Semantic methods
 - computing similarity based on meaning or concept
 - Latent semantic analysis, Explicit semantic analysis, Word Embedding

Methods – n -gram overlap

- An n -gram is an adjacent string of tokens (characters or words)
- “ n ” is the length of gram (or word)
- 1-gram (unigram)
- 2-gram (bigram)
- 3-gram (trigram)
- 4-gram (fourgram) and so on
- Based on set theoretic principles, used for pairwise comparison
- Two step process;
 1. Generate set of n -grams for each document
 2. Compute similarity between sets of n -grams

Methods – n -gram overlap (uni-gram)

- Source: 'a dog bites a man'
- Suspicious: 'a man was bitten by a dog'
- Generate sets of unigrams
- source_unigram: {a, dog, bites, a, man}
- suspicious_unigrams: {a, man, was, bitten, by, a, dog}
- Sim. Score = $4/\min(5,7) = 4/5 = 0.8$

Methods – n -gram overlap

- Degree of overlap, between two sets of n -grams (A is set of source n -grams and B is the set of suspicious n -grams), can be quantified using a number of measures
- Jaccard: $S = |A \cap B| / |A \cup B|$
- Dice: $S = 2 \times |A \cap B| / |A| + |B|$
- Overlap: $S = |A \cap B| / \min |A|, |B|$
- Containment: $S = |A \cap B| / |A|$

Methods – n -gram overlap (bi-gram)

- Source: 'a dog bites a man'
- Suspicious: 'a man was bitten by a dog'
- Generate sets of bigrams
- Source_bigram: {a dog, dog bites, bites a, a man}
- Suspicious_bigrams: {a man, man was, was bitten, bitten by, by a, a dog}
- Compute Similarity
- $\text{Sim} = 2/\min(4,6) = 2/4 = 0.5$

Methods – *n*-gram overlap (stop-words and lemmatisation)

- Source: 'a dog bites a man'
- Suspicious: 'a man was bitten by a dog'

- Source_unigrams (without stop-words): {dog, bites, man}
- Suspicious_unigrams (without stop-words): {man, bitten, dog}
- Sim. score = $2/3 = 0.66$

- Source (without stop-words + lemmas): {dog, bite, man}
- Suspicious (without stop-words + lemmas): {man, bite, dog}
- Sim. score = $3/3 = 1.0$

Methods – *n*-gram overlap

- Source: 'a dog bites a man'
- Suspicious: 'a young person was bitten by a big hound'
- Source_unigrams: {dog, bites, man}
- Suspicious_unigrams: {young, person, bitten, big, hound}
- Sim. Score = $0/3 = 0.0$
- Hard to detect heavily paraphrased text (or identify semantic similarity)

Query Expansion

- Same concept can be expressed using different terms
- Expand suspicious/source text with synonymous words from a lexicon/dictionary
- Lexicons
 - WordNet
 - Paraphrase Lexicon – generated using automatic paraphrase generation system
 - UMLS Meta Thesaurus (Biomedical Text)
- Query - car
- Expanded Query (w is weight)
 - $\text{car automobile}^w \text{ motorcar}^w$

Methods – n -gram overlap (with query expansion)

- Source: 'a dog bites a man'
- Suspicious: 'A young person was bitten by a big hound'
- Source_unigrams: {dog, bites, man}
- Suspicious_unigrams: {young, **teenager**, person, **man**, **male**, bitten, **cut**, big, hound, **dog**, **animal**}
- Sim. Score = $2/3 = 0.66$

Methods – word embedding

- Source: This is my blue car
- Suspicious: This is my blue automobile
- Source tokens: this, blue, car
- Suspicious tokens: this, blue, automobile

<i>Google n-gram model</i>	this	blue	automobile
this	1.0	0.01	0.05
blue	0.01	1.0	0.06
car	0.02	0.06	0.58

- Sim. Score = $1.0 + 1.0 + 1.0 / 3.0 = 1.0$ (*threshold value 0.5*)

Methods – word embedding

- Source: He was saddened by the news
- Suspicious: The news depressed him
- Sim. Score = 0.62

- Source: Niagara Falls is viewed by thousands of tourists every year
- Suspicious: Each year, thousands of people visit Niagara Falls
- Sim. Score = 0.68

Methods – cross-language character n -gram overlap

- Source: Copying words or ideas from someone else without giving credit
- Suspicious: Copiar las palabras o ideas de alguien más sin darle crédito
- Source: ['C', 'o', 'p', 'y', 'i', 'n', 'g', '_', 'w', 'o', 'r', 'd', 's', '_', 'o', 'r', '_', 'i', 'd', 'e', 'a', 's', ...]
- Suspicious : ['C', 'o', 'p', 'i', 'a', 'r', '_', 'l', 'a', 's', '_', 'p', 'a', 'l', 'a', 'b', 'r', 'a', 's', '_', 'o', 'r', '_', 'i', 'd', 'e', 'a', 's', ...]
- Sim. Score (uni-grams) = 0.933
- Sim. Score (bi-grams) = 0.356
- Sim. Score (tri-grams) = 0.1551

Methods – n -gram overlap (with bi-lingual dictionary)

- Source: 'Cheap electricity will be generated through coal in Pakistan with the Chinese investment'
- Suspicious: 'چین کی سرمایہ کاری سے پاکستان میں کوئلے کے ذریعے سستی ترین بجلی پیدا کرینگے'
- Source_unigrams: {Cheap, electricity, generated, through ,coal, Pakistan, Chinese, investment}
- Suspicious_unigrams: {چین، سرمایہ کاری، پاکستان، کوئلے، ذریعے، سستی ترین، بجلی، پیدا کرینگے}

Methods – *n*-gram overlap (with bi-lingual dictionary)

Source with POS	Suspicious with POS	Suspicious translation with POS
cheap <Adjective, JJ>	چین <Noun, NNP>	china <Noun, NNP>
electricity <Noun, NN>	سرمایہ کاری <Noun, NN>	investment <Noun, NN>
generate <Verb, VBN>	پاکستان <Noun, NNP>	pakistan <Noun, NNP>
through <Preposition, IN>	کوئلے <Noun, NN>	coal <Noun, NN>
coal <Noun, NN>	ذریعے <Noun, NN>	through <Preposition, IN>
pakistan <Noun, NNP>	سستی ترین <Adjective, JJ>	cheap <Adjective, JJ>
china <Noun, NNP>	بجلی <Noun, NN>	power <Noun, NN>
investment <Noun, NN>	پیدا کرینگے <Verb, VBN>	create <Verb, VBN>

- Sim. Score = $6/8 = 0.75$

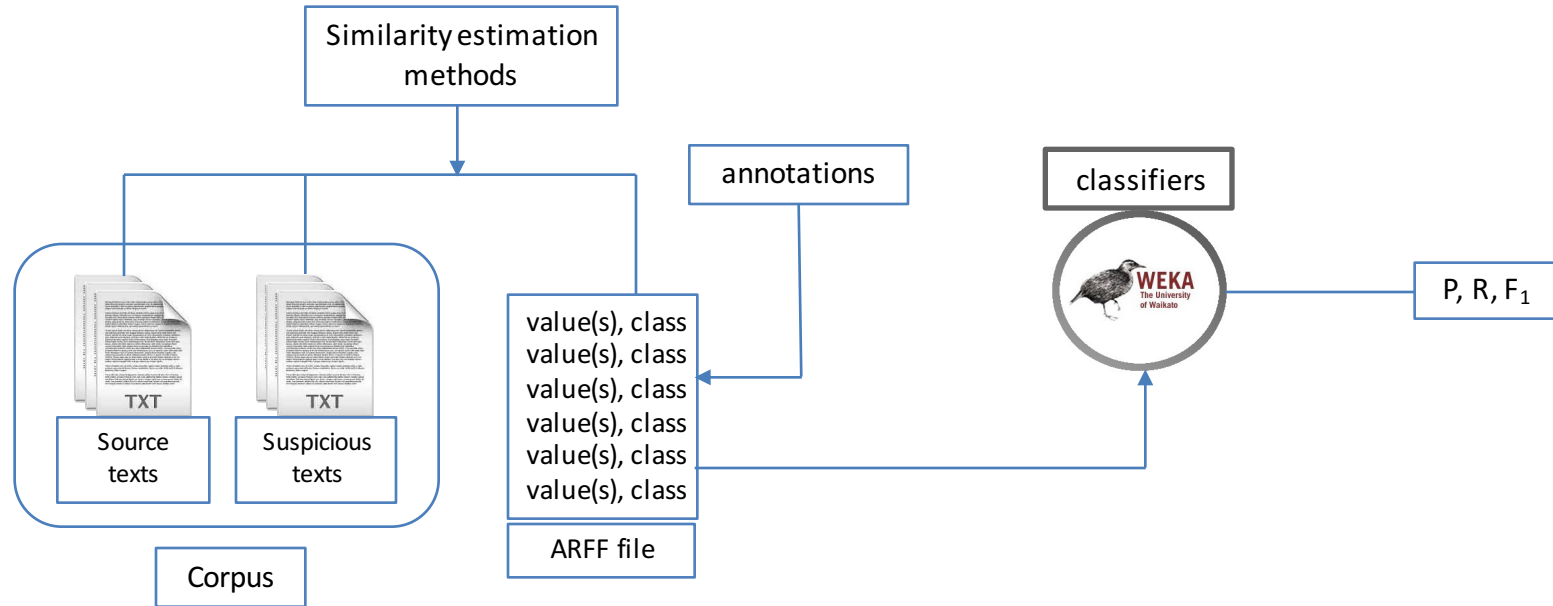
Evaluation

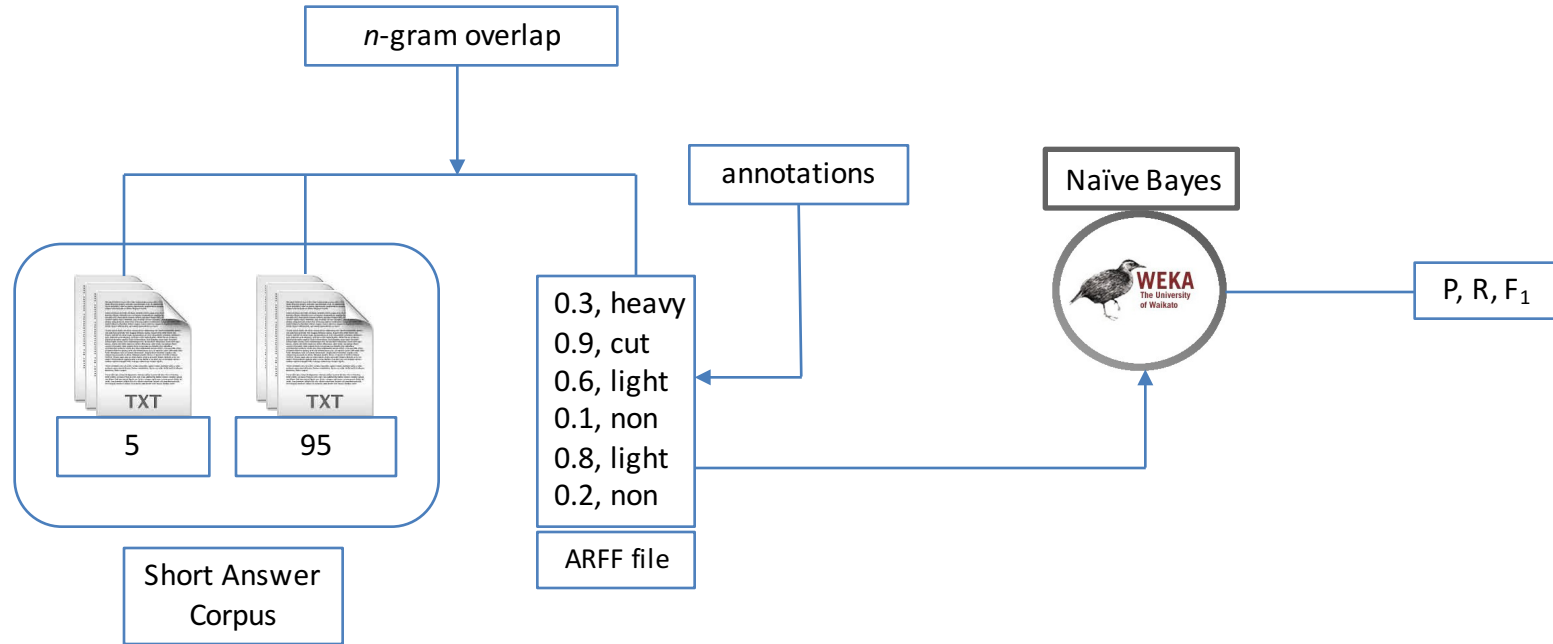
- Distinguishing between different levels of text reuse or plagiarism is a supervised classification task
 - Binary (Derived – Non Derived)
 - Ternary (Wholly – Partially – Non Derived) [multi]
- Precision, recall and F_1 scores are computed for each class using n-fold cross-validation
- WEKA implementation of the classifiers are used

	Predicted Positive	Predicted Negative
Positive cases	TP	FN
Negative cases	FP	TN

- Recall (R) = $TP / TP + FN$
- Precision (P) = $TP / TP + FP$
- $F_1 = 2PR / P + R$

Experimental setup





Practice lab

- We'll use **Short Answer Corpus (SAC)**, ***n*-gram overlap** and **Naïve Bayes classifier**
- SAC contains 5 source texts and 95 suspicious texts
 - near copy = 19, light revision = 19, heavy revision = 19, non-plagiarised = 38
- Naïve Bayes classifier is appropriate as it operates on numeric features generated by the *n*-gram overlap method
- Similarity scores for each suspicious-source text pair are used as features for the Naive Bayes classifier (WEKA)
- The corpus and python code can be downloaded from the following link;
- <https://github.com/muhmmadsharjeel/text-reuse-ss16>

References

- Alberto Barrón-Cedeño. "On the mono-and cross-language detection of text reuse and plagiarism." In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, pp. 914-914. ACM, 2010.
- Hermann A. Maurer, Frank Kappe, and Bilal Zaka. "Plagiarism-A Survey." J. UCS 12, no. 8 (2006): 1050-1084.
- Comas-Forgas, Rubén, and Jaume Sureda-Negre. "Academic plagiarism: Explanatory factors from students' perspective." Journal of Academic Ethics 8, no. 3 (2010): 217-232.
- Sousa-Silva, R., 2014. Investigating academic plagiarism: A forensic linguistics approach to plagiarism detection. International Journal for Educational Integrity, 10(1).
- Chapman, K.J. and Lupton, R.A., 2004. Academic dishonesty in a global educational market: A comparison of Hong Kong and American university business students. International Journal of Educational Management, 18(7), pp.425-435.
- Benno Stein, Sven Meyer zu Eissen, and Martin Potthast. "Strategies for retrieving plagiarized documents." In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 825-826. ACM, 2007.
- Zu Eissen, S.M. and Stein, B., 2006, April. Intrinsic plagiarism detection. In *European Conference on Information Retrieval* (pp. 565-569). Springer Berlin Heidelberg.
- Stamatatos, Efstathios. "Intrinsic plagiarism detection using character n-gram profiles." threshold 2 (2009): 1-500.
- Gaizauskas, Robert, Jonathan Foster, Yorick Wilks, John A rundel, Paul Clough, and Scott Piao. "The METER corpus: a corpus for analysing journalistic text reuse." In Proceedings of the corpus linguistics 2001 conference, pp. 214-223. 2001.
- Clough, P. and Stevenson, M., 2011. Developing a corpus of plagiarised short answers. Language Resources and Evaluation, 45(1), pp.5-24.
- Barrón-Cedeño, A., Vila, M., Martí, M.A. and Rosso, P., 2013. Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection. *Computational Linguistics*, 39(4), pp.917-947.
- Broder, Andrei Z. "On the resemblance and containment of documents." In *Compression and Complexity of Sequences 1997. Proceedings*, pp. 21-29. IEEE, 1997.
- Manning, Christopher D., and Hinrich Schütze. Foundations of statistical natural language processing. Vol. 999. Cambridge: MIT press, 1999.

Thanks

