The Short Stories Corpus

Faisal Alvi[1, 2], Mark Stevenson[1] and Paul Clough[1]

[1] University of Sheffield, United Kingdom & [2] King Fahd University of Petroleum and Minerals, Saudi Arabia.

1. Introduction

The PAN-2015 Text Alignment task invited data submissions – the corpus construction task. Participants were asked to construct a corpus that contains "*real world instances of text reuse or plagiarism*" and/or generated passages of "*reused or plagiarized text* by applying *obfuscation means of your choosing*".

2. The Short Stories Corpus

We constructed a corpus using various translations of Grimms' fairy tales as our data source. Some of the reasons for this choice are:

- Story retellings provide an interesting data source for text alignment algorithms. According to [1] "*Independent translations of fairy tales will provide for an interesting challenge for text alignment algorithms.*"
- Grimms' fairy tales are in the public domain, and have a good number of versions to provide for variety in storytelling.
- However, some versions contain archaic language, no longer in use today.

3. Corpus Details

- The corpus consists of 200 documents, with 50 documents spread across four text reuse/textual similarity strategies.
- Lengths of aligned passages in our corpus are similar to lengths of aligned passages in earlier PAN corpora, as shown below.

Table 1. Statistics of passage sizes in the corpus (number of characters)

Passage Length	No Plagiarism	Story Retelling	Synonym Replace	Character subst
Number of Docs.	50	50	50	50
Maximum Length	none	1160	765	729
Minimum Length	none	285	259	220
Average Length	none	**590**	**497**	**455**

- There are four strategies in the corpus as follows:

1. **No Plagiarism**: This group consists of stories that are completely different but may have minor textual overlap due the occurrence of some genre-specific words.
2. **Story Retelling**: Story retelling is defined as [2] "*a new, and often updated or retranslated, version of a story.*" This group consists of aligned passages from different translations of Grimms' fairy tales. These passages are embedded into other, unrelated fairy tales as the source and suspicious documents, thereby providing contextual similarity.
3. **Synonym Replacement**: This refers to replacement of words (and phrases) with synonymous words and equivalents. We created and used a customized list of synonyms for commonly occurring terms and phrases in the text – effectively a domain specific thesaurus.
4. **(UTF) Character Substitution**: Substitution of characters with their unicode equivalents (in order to exploit the weakness of a plagiarism detection approach) is known as technical disguise [4]. We used a simple replacement of two of the most frequently occuring letters 'a' and 'e' with their Cyrillic equivalents.

4. Examples from the Corpus

- **Story Retelling:**

Retold Story Version 1

It happened that ***the wedding of the King's eldest son was to be celebrated***, so the poor woman went up and placed herself by the door of the hall to look on. When all the candles were lit, and people, each more beautiful than the other, entered, and all was full of pomp and splendour, ***she thought of her lot with a sad heart*** , and cursed the pride and haughtiness which had humbled her and brought her to so great poverty.

The smell of the delicious dishes which were being taken in and out reached her, and now and then the servants threw her a few morsels of them: these ***she put in her jars to take home***.

Retold Story Version 2

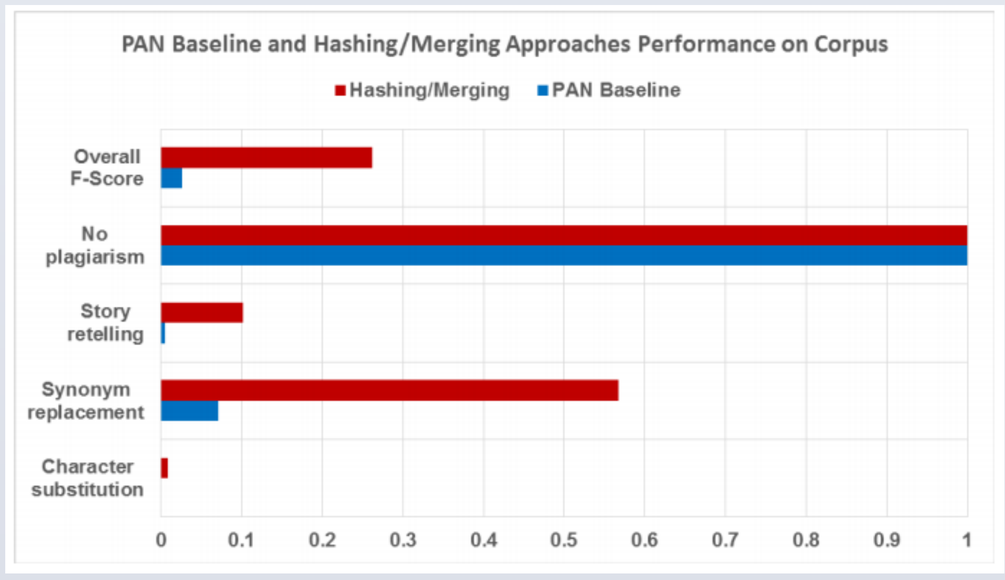
She had not been there long before she heard that ***the king's eldest son was passing by, going to be married;*** and she went to one of the windows and looked out. Everything was ready, and all the pomp and brightness of the court was there. Then ***she bitterly grieved*** for the pride and folly which had brought her so low. And the servants gave her some of the richmeats, which ***she put into her basket to take home***.

- **Synonym Replacement:**

1. (Fragment 1) The King, who had a bad heart, and was angry...
2. (Fragment 2) the monarch, who had a worse heart, and was enraged...

5. Results

We used PAN Baseline approach and our hashing and merging based approach [3] to obtain plagdet scores.



- The results indicate that these algorithms performed very well in detecting no plagiarism and synonym replacement strategies.
- However, the performance was low in case of story retelling and the plagdet score was close to zero in case of character substitution, suggesting that technical disguise was the most difficult to detect.
- According to [1], our corpus displayed "*typical typical detection performances among state-of-the-art text alignment approaches*".

References

1. Martin Potthast, Steve Göring, Paolo Rosso, and Benno Stein. **Towards Data Submissions for Shared Tasks: First Experiences for the Task of Text Alignment**. In Working Notes Papers of the CLEF 2015 Evaluation Labs, CEUR Workshop Proceedings, September 2015. CLEF and CEUR-WS.org. ISSN 1613-0073
2. http://dictionary.reference.com/browse/retelling [Last Accessed: 07-June-2015]
3. Alvi, F., Stevenson, M., Clough, P.D.: **Hashing and Merging Heuristics for Text Reuse Detection**. In: Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014. pp. 939–946 (2014).
4. Meuschke, N., Gipp, B.: **State-of-the-art in Detecting Academic Plagiarism**. International Journal for Educational Integrity 9(1) (2013)