The Short Stories Corpus
Notebook for PAN at CLEF 2015

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

[1]University of Sheffield, United Kingdom,
[2]King Fahd University of Petroleum & Minerals, Saudi Arabia.
{falvi1, mark.stevenson, p.d.clough}@sheffield.ac.uk

Abstract In this work we describe the construction of a plagiarism detection/text reuse corpus submitted for the PAN-2015 Evaluation Lab. Our corpus consists of four different text reuse scenarios namely, (1) no-plagiarism, (2) story-retelling, (3) synonym-replacement and (4) character-substitution. Among these scenarios the most interesting one is story retelling - through it we find patterns of textual similarity between story retellings. We use Grimm brothers fairy tales as described in the Project Gutenberg as the source of our documents. The corpus consists of 200 pairs of documents, with 50 document pairs for each type of text reuse. Empirical observation shows interesting patterns of textual similarity within the corpus. Furthermore, plagiarism detection using various approaches shows the difficulty of detection of various groups within the corpus.

1 Introduction

The PAN Lab Evaluation Series [10] has been active in conducting experimental evaluation of plagiarism detection approaches since the past several years. This year in PAN-2015 [9], the corpus construction task has been initiated for the text alignment task - i.e., the participants have to submit a corpus with annotated passages involving real and/or artificially generated samples of text reuse or plagiarism.

Text reuse detection has been a well researched area in the news domain [4]. In this work we explore textual similarity in short stories. For this task, we submitted a collection of document pairs that have annotated passages classified within four groups. The source documents' passages have been taken from various translations of Grimms' fairy tales and are available on the Project Gutenberg [1] website.

The corpus consists of 200 document pairs with 50 document pairs each within the following four different groups namely, (1) no-plagiarism, (2) (human) story-retelling, (3) synonym-replacement and (4) character-substitution. The no-plagiarism group contains completely different short stories that may share some genre-specific terms leading to minor textual overlap. The story retelling group describes pairs of story fragments taken from two different retellings by human writers. The third group, synonym replacement, describes story fragment pairs with replacement of words and phrases with their synonyms. Finally, character substitution refers to technical disguise, where letters in words are replaced with their look-alike unicode equivalent characters. Empirical observation reveals interesting similarity patterns within the corpus.

2 Corpus Construction

The corpus consists of documents from the Grimms fairy tales as available on Project Gutenberg website. The corpus is small in comparison to other PAN corpora. This is because the number of tales available within the Grimm's collection ranges from a maximum of 200 in some editions to less than 50 within other editions. In order to have a balanced collection of documents within each group, our corpus consists of 200 document pairs, with 50 pairs for each group. Some statistics related to the passage length within the corpus documents are shown in table 1.

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**

Here, a passage is defined as a contiguous maximal-length sequence of characters (or text) that consists of similar text between two versions. For corpus construction, we selected passages from two versions of a tale that correspond to the same events, since different versions of the same story may sometimes differ in details of events.

2.1 No plagiarism

The no-plagiarism group consists of stories that are completely different but may have minor textual overlap due the occurrence of some genre-specific words. For this group, we found ferret [6] trigram similarity for the the entire Grimms collection and chose 50 document pairs such that there was no other similarity within them.

2.2 Story Retelling

(Story) retelling is defined as "*a new, and often updated or retranslated, version of a story.*"[1]. In this context, a question that appears on some internet forums or websites is: "*Is retelling an old fairy tale (or a short story) considered plagiarism?*"[2].

In this work we do not address this question definitively i.e., "*Does story retelling involve text reuse and/or plagiarism?*". Therefore, we use the term 'textual similarity' when referring to similar passages of text within two story retellings. However, from the literature we see that Clough et al [4] remark that, "*Of course, reusing language is as old as the retelling of stories...*". This suggests a link between retelling of stories

[1] http://dictionary.reference.com/browse/retelling [Last Accessed: 07-June-2015]

[2] http://www.answers.com/Q/Is_rewriting_an_old_fairy_tale_considered_plagiarism [Last Accessed: 15-July-2015]

and reuse of language; consequently a story retelling *may* involve text reuse from the original story, however we cannot claim this for a particular retelling.

In figure 1 we show fragments from two retellings of the story 'King Thrushbeard' (also called 'King Grisly Beard'). Here we give a correspondence between sentence fragments found between the two retellings:

1. (Fragment 1) The wedding of the King's eldest son was to be celebrated.
   (Fragment 2) The king's eldest son was passing by, going to be married.
2. (Fragment 1) She thought of her lot with a sad heart.
   (Fragment 2) She bitterly grieved.
3. (Fragment 1) She put in her jars to take home.
   (Fragment 2) She put into her basket to take home.

We see that in pair 1, fragment 2 corresponds to a change of voice from fragment 1 with some modification; likewise in pair 2, fragment 2 corresponds to a summarization of fragment 1, while the two fragments in pair 3 are exactly the same with minor word replacement. It may be relevant to mention here that the original Grimm's tales are in the German language, with these two retellings being two different translated versions in English. Earlier, Barzilay et al [3] have also extracted paraphrases from multiple English translations of the same source text in English.

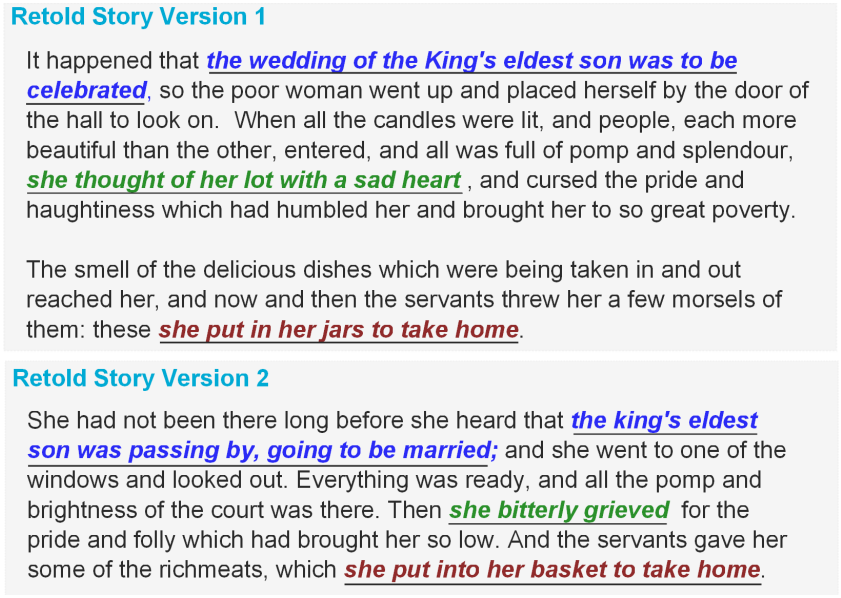

Figure 1. Example of textual similarity in story retellings

2.3 Synonym Replacement

The third group in our corpus is synonym replacement. This refers to replacement of words (and some phrases) with synonymous words and equivalents. For this purpose, we initially used Wordnet [8] by searching for the synset corresponding to a given word in the text and returning back the first available synonym. However the resulting text was too far from the original in some cases. This technique can be improved by incorporating the context in which a particular word appears as well as incorporating word sense disambiguation. We plan to incorporate these changes to the corpus at a later stage.

Since the size of the corpus is not large and words belong to a particular domain, we created a customized list of synonyms for commonly occurring words and phrases in the documents. While this approach certainly produced meaningful texts, it is may not be scalable for large number of documents and/or corpora involving diverse topics. In addition to these, we also removed some articles (a, an, the), and replaced alternate occurences of some pronouns with proper nouns. Below we give an example of 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...

2.4 Character Substitution or Technical Disguise

Substitution of characters with their unicode equivalents in order to exploit the weakness of a plagiarism detection approach is known as technical disguise [7]. In this work we used a simple replacement of two of the most frequently occuring letters 'a' and 'e' with their Cyrillic equivalents [5]. Table 2 shows the correspondence between an ASCII letter and its unicode cyrillic equivalent.

Table 2. The letters 'a' and 'e' with their cyrillic equivalents

Ansi Character	Unicode Value	Unicode Equivalent	Unicode Value
a	92	а (Cyrillic)	U + 0430
e	97	е (Cyrillic)	U + 0435

Most word n-gram based approaches might fail to detect this type of obfuscation since a unit of similarity in these approaches is a word. In the example shown below we replace an 'e' with an '**е**' and an 'a' with an '**а**' to emphasize the change that happens.

1. (Sentence 1) Now there lived in the country two brothers, sons of a poor man, who declared themselves willing to undertake the hazardous enterprise.
2. (Sentence 2) Now th**е**r**е** liv**е**d in th**е** country two brothers, sons of **а** poor m**а**n, who d**е**cl**а**r**е**d th**е**ms**е**lv**е**s willing to und**е**rt**а**k**е** th**е** h**а**z**а**rdous **е**nt**е**rpris**е**.

3 Results and Discussion

We tested the corpus using the simple PAN Baseline approach as well as our hashing and merging based approach [2] submitted to PAN-2014. Numerical results of precision, recall, granularity and overall plagdet score are given in Table 3, while a comparison of the two approaches performance on various groups is given in Fig 2.

Table 3. Results of PAN Baseline and our hashing/merging approach on the corpus

	Approach Used	Overall	No Plag	Story Retelling	Synonym Replace	Character Substitut
Plagdet	PAN Baseline	0.02593	1.00000	0.00492	0.07125	0.00000
	Hash/Merge	0.26221	1.00000	0.10150	0.56686	0.00886
Precision	PAN Baseline	0.99632	1.00000	1.00000	0.99598	0.00000
	Hash/Merge	0.98663	1.00000	0.99788	0.99923	0.00446
Recall	PAN Baseline	0.01313	1.00000	0.00246	0.00369	0.00000
	Hash/Merge	0.15119	1.00000	0.05347	0.39565	0.61392
Granularity	PAN Baseline	1.00000	1.00000	1.00000	1.00000	1.00000
	Hash/Merge	1.00000	1.00000	1.00000	1.00000	1.00000

From figure 2, we see that the two approaches performed very well in detecting no-plagiarism and synonym replacement. However the overall score was low in case of story retelling and is close to zero in case of character substitution.

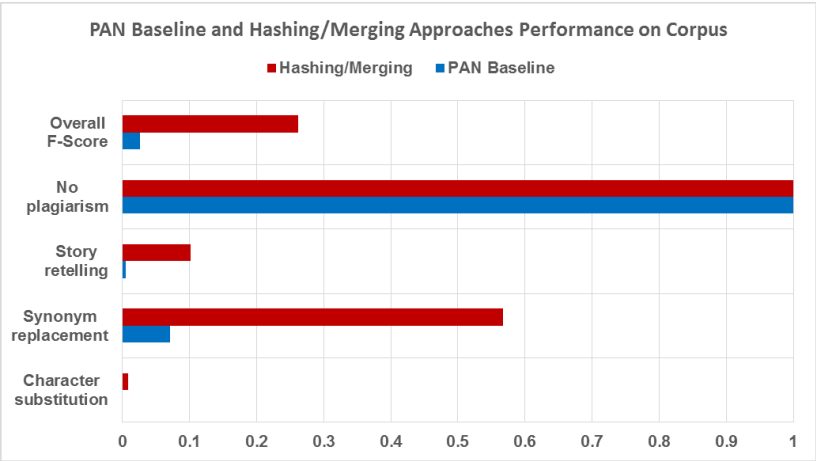

Figure 2. Visual performance of various approaches based on the Plagdet Scores

4 Conclusion and Future Work

In this work we constructed a corpus in PAN format based on story retelling. We used various translations of Grimms' fairy tales in the construction of this corpus. The question whether a story retelling involves text reuse has not been addressed in this work comprehensively. However, given the nature of story retelling, it can be expected that text reuse *may* happen in case of story retelling. Apart from story retelling our other strategies were synonym replacement and character substitution in the form of technical disguise.

In future, this corpus can be improved upon by enhancing synonym replacement with a comprehensive automatic paraphrasing strategy. Another possible area of exploration could be to extend the domain of story retelling to modern short stories. Some of the stories in our corpus contain archaic language or words, however this is to be expected since we used versions of fairy tales that, unlike modern short stories, are in the public domain.

References

1. Books: Grimm (sorted by popularity) - Project Gutenberg. http://www.gutenberg.org/ebooks/search/?query=grimm, Last Accessed: 2015-07-15
2. 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)
3. Barzilay, R., McKeown, K.R.: Extracting Paraphrases from a Parallel Corpus. In: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics. pp. 50–57. Association for Computational Linguistics (2001)
4. Clough, P.D., Gaizauskas, R.J., Piao, S.S., Wilks, Y.: Measuring text reuse. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA. pp. 152–159 (2002)
5. Gillam, L., Marinuzzi, J., Ioannou, P.: Turnitoff–Defeating Plagiarism Detection Systems. In: Proceedings of the 11th Higher Education Academy-ICS Annual Conference. Higher Education Academy (2010)
6. Lane, P., et al.: UH Ferret: Implementation of a copy-detection tool. Software (2011), http://uhra.herts.ac.uk/handle/2299/12041
7. Meuschke, N., Gipp, B.: State-of-the-art in Detecting Academic Plagiarism. International Journal for Educational Integrity 9(1) (2013)
8. Miller, G.A.: WordNet: A Lexical Database for English. Communications of the ACM 38(11), 39–41 (1995), http://doi.acm.org/10.1145/219717.219748
9. Potthast, M., Hagen, M., Göring, S., Rosso, P., Stein, B.: 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, CLEF and CEUR-WS.org (Sep 2015), http://www.clef-initiative.eu/publication/working-notes
10. Potthast, M., Stein, B., Eiselt, A., Barrón-Cedeño, A., Rosso, P.: Overview of the 1st International Competition on Plagiarism Detection. In: Stein, B., Rosso, P., Stamatatos, E., Koppel, M., Agirre, E. (eds.) SEPLN 09 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09). pp. 1–9. CEUR-WS.org (Sep 2009)

A Peer Review of Submitted Corpora

Detailed reviews of various corpora have been submitted earlier. Due to space limitations, here we give a summary of corpora review as shown in table 4. Regarding the errors,

- *Synonym errors* refer to an error in replacement of word by its synonym. Such a replacement may make the resulting text somewhat incomprehensible.
- *Demarcation Errors* refer to errors in the alignment of the source and the suspicious text. For example, the source text may be marked a few characters earlier or later than it should have been. However, this is a minor error and might have also been observed due to programming errors in creation of XML files, or in the viewer program used for reviewing the corpora.

The average errors per document figure is given as a range, since definition of synonym error is not precise – likewise demarcation errors might not be present at all, but may have been observed due to the character set used.

Due to lack of expertise in language and/or non-availability of language resources, review of non-English corpora could not be successfully carried out.

Table 4. Tabular Summary of Peer Review for Submitted Corpora

Corpus Name	Mono/Bi Lingual	Method of Construction	Documents Reviewed	Average Errors Per Document	Observations and Errors
alvi-15	Mono (English)	Refer to Notebook	-	-	-
cheema-15	Mono (English)	Possibly Automatic + Manual	≈ 75	≥ 0	• Meaningful names for obfuscation strategies needed. • Synonym errors may be present.
mohtaj-15	Mono (English)	Possibly Automatic	≈ 75	≥ 1	• Synonym errors and demarkation errors may be present.
palkovskii-15	Mono (English)	Possibly Automatic	≈ 75	≥ 1	• Some grammatically incorrect sentences.
asghari-15	Bi (Eng-Persian)	No basis for judgment	< 10	N/A	• N/A
hanif-15	Bi (Eng-Urdu)	No basis for judgment	< 10	N/A	• N/A
khoshnava-15	Mono (Persian)	No basis for judgment	< 10	N/A	• N/A
kong-15	Mono (Chinese)	No basis for judgment	< 10	N/A	• N/A