Multi-feature Plagiarism Detection

Jan Kasprzak, Šimon Suchomel, and Michal Brandejs

Faculty of Informatics, Masaryk University
{kas,suchomel,brandejs}@fi.muni.cz

1 General Approach

The approach Masaryk University team has used in PAN 2012 Plagiarism detection—detailed comparison sub-task is based on the same approach that we have used in PAN 2010 [2]. This time, we have used a similar approach, enhanced by several means

The algorithm evaluates the document pair in several stages:

- intrinsic plagiarism detection
- language detection of the source document
  - cross-lingual plagiarism detection, if the source document is not in English
- detecting intervals with common features
- post-processing phase, mainly serves for merging the nearby common intervals

2 Intrinsic plagiarism detection

Our approach is based on character n-gram profiles of the interval of the fixed size (in terms of n-grams), and their differences to the profile of the whole document [6]. We have further enhanced the approach with using gaussian smoothing of the style-change function [2].

For PAN 2012, we have experimented with using 1-, 2-, and 3-grams instead of only 3-grams, and using the different measure of the difference between the n-gram profiles. We have used an approach similar to [1], where we have compute the profile as an ordered set of 400 most-frequent n-grams in a given text (the whole document or a partial window). Apart from ordering the set we have ignored the actual number of occurrences of a given n-gram altogether, and used the value inveresly proportional to the n-gram order in the profile, in accordance with the Zipf's law [7].

This approach has provided more stable style-change function than than the one proposed in [6]. Because of pair-wise nature of the detailed comparison sub-task, we couldn't use the results of the intrinsic detection immediately, so we wanted to use them as hints to the external detection.

3 Cross-lingual detection

For cross-lingual plagiarism detection, our aim was to use the public interface to Google translate if possible, and use the resulting document as the source for standard intra-lingual detector. Should the translation service not be available, we wanted to use the

fall-back strategy of translating isolated words only, with the additional exact matching of longer words (we have used words with 5 characters or longer). We have supposed these longer words can be names or specialized terms, present in both languages.

We have used dictionaries from several sources, like `dicts.info`[1], `omegawiki`[2], and `wiktionary`[3]. The source and translated document were aligned on a line-by-line basis.

In the final form of the detailed comparison sub-task, the results of machine translation of the source documents were provided to the detector programs by the surrounding environment, so we have discarded the language detection and machine translation from our submission altogether, and used only line-by-line alignment of the source and translated document for calculating the offsets of text features in the source document.

4 Multi-feature Plagiarism Detection

Our pair-wise plagiarism detection is based on finding common passages of text, present both in the source and suspicious document. We call them *features*. In PAN 2010, we have used sorted word 5-grams, formed from words of three or more characters, as features to compare. Recently, other means of plagiarism detection have been explored: Stop-word n-gram detection is one of them [5].

We propose the plagiarism detection system based on detecting common features of various type, like word n-grams, stopword n-grams, translated words or word bigrams, exact common longer words from document pairs having each document in a different language, etc. The system has to be to the great extent independent of the specialities of various feature types. It cannot, for example, use the order of given features as a measure of distance between the features, as for example, several word 5-grams can be fully contained inside one stopword 8-gram.

We thus define *common feature* of two documents (susp and src) as the following tuple:

$$\langle \text{offset}_{\text{susp}}, \text{length}_{\text{susp}}, \text{offset}_{\text{src}}, \text{length}_{\text{src}} \rangle$$

In our final submission, we have used only the following two types of common features:

- word 5-grams, from words of three or more characters, sorted, lowercased
- stop-word 8-grams, from 50 most-frequent English words (including the possessive suffix 's), unsorted, lowercased, with 8-grams formed only from the seven most-frequent words (*the, of, a, in, to, 's*) removed

We have gathered all the common features for a given document pair, and formed *valid intervals* from them, as described in [3] (a similar approach is used also in [5]). The algorithm is modified for multi-feature detection to use character offsets only instead of feature order numbers. We have used valid intervals consisting of at least 5

[1] http://www.dicts.info/
[2] http://www.omegawiki.org/
[3] http://en.wiktionary.org/

common features, with the maximum allowed gap inside the interval (characters not belonging to any common feature of a given valid interval) set to 3,500 characters.

We have also experimented with modifying the allowed gap size using the intrinsic plagiarism detection: to allow only shorter gap if the common features around the gap belong to different passages, detected as plagiarized in the suspicious document by the intrinsic detector, and allow larger gap, if both the surrounding common features belong to the same passage, detected by the intrinsic detector. This approach, however, did not show any improvement against allowed gap of a static size, so it was omitted from the final submission.

5 Postprocessing

6 Further discussion

In the full paper, we will also discuss the following topics:

- language detection
- suitability of plagdet score[4] for performance measurement
- feasibility of our approach in large-scale systems
- other possible features to use, especially for cross-lingual detection
- discussion of parameter settings

References

1. Cavnar, W.B., Trenkle, J.M.: N-gram-based text categorization. In: In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval. pp. 161–175 (1994)
2. Kasprzak, J., Brandejs, M.: Improving the reliability of the plagiarism detection system. In: Notebook Papers of CLEF 2010 LABs and Workshops. Citeseer (2010)
3. Kasprzak, J., Brandejs, M., Kipa, M.: Finding plagiarism by evaluating document similarities. In: SEPLN'09: The 25th edition of the Annual Conference of the Spanish Society for Natural Language Processing (2009)
4. Potthast, M., Stein, B., Barrón-Cedeño, A., Rosso, P.: An Evaluation Framework for Plagiarism Detection. In: Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010) (to appear). Association for Computational Linguistics, Beijing, China (Aug 2010)
5. Stamatatos, E.: Plagiarism detection using stopword n-grams. Journal of the American Society for Information Science and Technology (2011)
6. Stamatatos, E.: Intrinsic plagiarism detection using character n-gram profiles. In: Proceedings of the SEPLN'09 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse. pp. 38–46 (2009)
7. Zipf, G.: The psycho-biology of language. (1935)