Developing High-Resolution Universal Multi-Type N-Gram Plagiarism Detector

Notebook for PAN at CLEF 2014

Yurii Palkovskii, Alexei Belov

Zhytomyr Ivan Franko State University, Institute of Foreign Philology
SkyLine LLC, Plagiarism Detector Project
vb-expert@yandex.ru cargo-base@yandex.ru

Abstract. Abstract. This paper describes approaches used for the Plagiarism Detection task used in PAN 2014 International Competition on Uncovering Plagiarism, Authorship, and Social Software Misuse, that scored 1-st place with plagdet score (0.907) for test corpus no.3 and 3-rd place score (0.868) for test corpus no. 2. In this work we aggregated all the previously researched experience from PAN12 and PAN 13 research works and thus further improved previously developed methods of detecting plagiarism, with the help of n-gram based fingerprinting that includes: contextual n-grams, surrounding context n-grams, named entity based n-grams, odd-even skip n-grams, functional words frame based n-grams, TF-IDF sentence level similarity index and noise sensitive clusterization algorithm, focused summary type detection heuristics, combined into a single model to mark similarity sections and thus effectively detect different types of obfuscation techniques.

1 Introduction

Each year PAN competition pushes forward the baseline of plagiarism detection effectiveness thus making it harder and harder to compete with the top plagdet score becoming close to the absolute values irrespective of even more demanding plagiarism types included into the new corpora. This year we tried to incorporate all the knowledge we could aggregate via already published PAN works and try to figure out how to make most of the algorithms that already proved to achieve best performance during the last years. The introduction of the lately developed TIRA platform has greatly boosted the software deployment and thus made it easier both to participate in the competition, synchronize the local result and the result in the test environment.

2 Methods

The main goal of this year research was to aggregate all the most efficient methods that have already been applied to the task of plagiarism detection. We did such analysis and came up with the following methods of n-gram generation:

1. Regular n-grams
2. Variable length stop-word n-grams
3. Named Entity based n-grams
4. Most Frequently used words n-grams

N-grams expansion by:

1. Odd-even generation on a sentence level
2. Resulting n-gram set stemming
3. A single word deletion on a fingerprint level
4. N-grams alphasorting

Text preprocessing included special symbols removal and space trimming. The input text parsing was made on 2 levels: sentence level and word level.

We applied angled ellipse based graphical clustering algorithm to define clusters of shared fingerprints. The main approach was to detect what kind of plagiarism is dominating within the document pair - by applying special kind of analysis that included global noise level detection, single stage analysis for the existence of verbatim plagiarism clusters, shared fingerprints sequence analysis, diagonal density analysis, summary type presence and then finally selecting which analysis preset will be used for final analysis. Our software selects one of 4 possible parameter presets for the final detection strategy:

1. Verbatim Plagiarism
2. Random Plagiarism
3. Summary type Plagiarism
4. Undefined type

3 Evaluation

The developed software scored 1-st place with plagdet score (0.907) for test corpus no.3 and 3-rd place score (0.868) for test corpus no. 2. At the moment we are not able to assess and explain the difference in results as long as both corpora are not yet publically downloadable and we do not know the difference between the test corpora versions, we are waiting for the test data release to further research the difference in the achieved Recall and Granularity.

Corpus:	Plagdet:	Recall:	Precision	Granularity	Runtime:
Test 3	0.90779	0.88916	0.92757	1.00027	00:57:15
Test 2	0.86806	0.82637	0.92227	1.00580	01:10:04

4 Conclusions and Future Work

The result achieved at this year PAN competition shows really tight competition for every percent. Comparing our previous results to this year results we may conclude that we made really good progress landing between the best competing teams at PAN. Unfortunately, due to many different factors that go beyond the scope of the abovementioned research, we were not able to fully optimize the developed system. The majority of the parameters used were heuristically set to most optimal values we come with initially. This fact shows large space for future improvements. Applying genetic algorithm for global parameter tuning will definitely give much better results and will allow much better tuning for each specific corpus.

Acknowledgments. To the PAN RnD team, for their fantastic job making deploying and running software on TIRA platform possible, for their prompt responses and discrete guidance both on the competition website and Google groups. We would like to thank our competing teams as well, as our research was heavily based on and dramatically boosted by the knowledge and experience they shared in their previously published works and live workshop discussions at previous PANs. We also would like to thank the organizers for providing support and the official letter of invitation to be able to append PAN workshop aiding us with visa acquisition for the conference.

References

1. Martin Potthast, Benno Stein, Alberto Barrón-Cedeño, and Paolo Rosso. An Evaluation Framework for Plagiarism Detection. In 23rd International Conference on Computational Linguistics (COLING 10), August 2010. Association for Computational Linguistics.
2. Martin Potthast, Matthias Hagen, Tim Gollub, Martin Tippmann, Johannes Kiesel, Paolo Rosso, Efstathios Stamatatos, and Benno Stein. Overview of the 5th International Competition on Plagiarism Detection. In Pamela Forner, Roberto Navigli, and Dan Tufis, editors, Working Notes Papers of the CLEF 2013 Evaluation Labs, September 2013. ISBN 978-88-904810-3-1.
3. Rodríguez-Torrejón, D.A., Martín-Ramos, J.M.: "Detección de plagio en documentos: sistema externo monolingüe de altas prestaciones basado en n-gramas contextuales" (Plagiarism Detection in Documents: High Performance Monolingual External Plagiarism Detector System Based on Contextual N-grams). Procesamiento del Lenguaje Natural. N. 45 (2010).
4. Rodríguez-Torrejón D.A., Martín-Ramos J.M.: CoReMo System (Contextual Reference Monotony) A Fast, Low Cost and High Performance Plagiarism Analyzer System: Lab Report for PAN at CLEF 2010. In Braschler M., Harman D., Pianta E.,

editors. Notebook Papers of CLEF 2010 LABs and Workshops, 22-23 September, Padua, Italy, 2010.
5. Tim Gollub, Martin Potthast, Anna Beyer, Matthias Busse, Francisco Rangel, Paolo Rosso, Efstathios Stamatatos, and Benno Stein. Recent Trends in Digital Text Forensics and its Evaluation. In Pamela Forner, Henning Müller, Roberto Paredes, Paolo Rosso, and Benno Stein, editors, Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. 4th International Conference of the CLEF Initiative (CLEF 13), September 2013. Springer. ISBN 978-3-642-40801-4.
6. Šimon Suchomel, Jan Kasprzak, and Michal Brandejs. Three Way Search Engine Queries with Multi-feature Document Comparison for Plagiarism Detection—Notebook for PAN at CLEF 2012. In Forner et al. [6]. ISBN 978-88-904810-3-1. URL http://www.clef-initiative.eu/publication/working-notes
7. Efstathios Stamatatos. Plagiarism Detection Using Stopword n-Grams. JASIST, 62(12):2512–2527, 2011. doi: http://dx.doi.org/10.1002/asi.21630.
8. Yurii Palkovskii and Alexei Belov. Applying Specific Clusterization and Fingerprint Density Distribution with Genetic Algorithm Overall Tuning in External Plagiarism Detection—Notebook for PAN at CLEF 2012. In Forner et al. [6]. ISBN 978-88-904810-3-1. URL http://www.clef-initiative.eu/publication/working-notes.