JOURNAL OF INTEGRATED OMICS
A METHODOLOGICAL JOURNAL
HTTP://WWW.JIOMICS.COM

ORIGINAL ARTICLE | DOI: 10.5584/jiomics.v8i4.230

Recovering Damaged Documents to Improve Information Retrieval Processes

Ángel Luis Garrido[1*], Alvaro Peiró[2]

[1] SID Research Group, University of Zaragoza, Pedro Cerbuna 12, 50009 Zaragoza, Spain ; [2] InSynergy Consulting S.A., Calle de López de Hoyos141, planta 2º - 28002 Madrid, Spain.

Received: 31 December 2017 **Accepted:** 22 October 2018 **Available Online**: 19 December 2018

ABSTRACT

Although computer forensics is frequently related to the investigation of computer crimes, it can also be used in civil procedures. An example of case of use is information retrieval from damaged documents, where words have undergone alterations, either accidentally or intentionally. In this paper, we present a new tool able to retrieve information from large volumes of documents whose contents have been damaged. We have designed a new approach to recover the original words, composed of two stages: a text cleaning filter, able to remove non relevant information, and a text correction unit, which gather a general purpose spell checker with a N-gram based spell checker built specifically for the domain of the documents. The benefits of using this combined approach are two-fold: on the one hand, the general spell checker allows us to leverage all the general purpose techniques that are usually used to perform the corrections; on the other hand, the use of an N-gram based model allows us to adapt them to the particular domain we are tackling exploiting text regularities detected in successfully processed domain documents. The result of the correction allows us to improve automatic information retrieval tasks of from the texts. We have tested it using a real data set by using an information extraction tool based on semantic technologies in collaboration with the Spanish company InSynergy Consulting.

Keywords: Information Recovery, Text Cleaning, Spell Check, N-gram.

1. Introduction

Computer forensics is a specialty within forensic science disciplines. The goal of computer forensics is the search for evidence in digital data, involving the preservation, identification, storage media extraction, documentation, and interpretation [1]. Computer forensics can also be used in civil procedures, and information recovery is one of these scenarios. In fact, correcting texts, also known as "text cleaning", is a well-known task in the field Natural Language Processing [2] and it has been widely used in many related works [3]. We can find different cases of use for the application of text cleaning inside the field of computer forensics, especially regarding activities related to the reconstruction of documents for legal, police, administrative or industrial purposes. In this work, we propose a text cleaning system for automatically recovering data from large documents, whose functioning is composed of two main steps: 1) a Text Cleaning Filter, able to remove non relevant information, as headings, page numbers, stamps, etc., and 2) a Text Correction Unit, which combines a general purpose spell checker with a N-gram based spell checker built specifically for the domain of the documents we are dealing with. The main contribution of our approach is the experimental study of how the application of N-gram together with the grammar checker can contribute to the

***Corresponding author:** Dr. Ángel Luis Garrido; e-mail: garrido@unizar.es. SID Research Group. Department of Computer Science and Systems Engineering. School of Engineering and Architecture. Ed. Ada Byron, Maria de Luna, 1, 50018, Zaragoza (Spain) Phone: +34 654 833 616. E-mail address: garrido@unizar.es.

improvement of a text cleaning task. For that, we have integrated our implementation in AIS [6,7], an information extraction system within the domain of legal texts, where extraction process is guided by a specific domain ontology for the typology of the document.

2. Material and Methods

In real-world scenarios, systems must deal many times with documents which are actually scanned images. In such cases, to extract the text contents out of the documents, a broadly adopted solution is to apply an Optical Character Recognition (OCR) system over the documents before any information retrieval task. Its effectiveness clearly influences the quality of the results, but even when the OCR is perfect, spelling problems or noisy words can be found. So, we propose a two-stage system with the aim of improving the text quality before a hypothetical data retrieval process:

1) After the OCR process, the first stage for the potentially damaged documents consists of removing non relevant information from the input to get the text as clean as possible for the correction step. The list of non-relevant items controlled by our system is as follows: document headers and foots, page numbers, stamps, spots, noisy characters, and signatures. All them difficult the retrieval process. These elements are removed by using regular expressions on the text.

2) The second stage is in charge of correcting the misspelled words which appear in the text. For this purpose, we advocate to apply a general purpose spell checker and, whenever it is possible, to enrich it with an N-gram based spell checker built specifically for the domain of the documents we are dealing with. We have used Hunspell [10], a powerful spell checker and morphological analyzer which offers a good multi-language support (e.g., English, French, Spanish, etc.), and we have developed our own N-gram libraries. For each detected error for a given word (wi), 1) we get the spell checker suggestions (si), and, 2) we assign each of them a score based on an adaptation of the Needleman-Wunsch [8] distance and computed with the actual word, and we reinforce them with the probabilities of the N-grams suggestions as follows:

$$wordScore(wi, si) = gSP(wi, si) * NG(wi)^{\alpha} \quad (1)$$

where gSP is the score based on the metric distance, and NG is the probability of wi being the next word in the domain where it appears. Both values are normalized in the 0..1 range. The α is used to assign a weight to the NG(wi) value in order to give more or less relevance to the suggested word.

Note that our system only accepts words that are suggested by the spell checker, and gets their probability from the N-gram suggestion list. Besides, wordScore is never 0 as gSP(wi) always returns a value greater than 0, as we add perplexity to our N-gram model using the add-one Laplace Smoothing method [9].

3. Results

For the experiments, we have used a dataset formed by 250 notarial purchase documents in Spanish. These documents have been selected randomly from the private repository of InSynergy Consulting (ISYC), a company devoted to document management. These 250 documents have between 100 and 200 pages and each document was manually revised by expert professionals in the field to detect text errors. We have trained our domain N-gram model using 7,500 already refined notarial documents in Spanish, also available in the private repository of ISYC. After testing with a 5-Gram model, we finally have used in the experiments a 3-Gram model because of its better behavior, comprising 1,213,920 different sequences and a vocabulary of 174,831 different words. As we wanted to evaluate how our domain N-gram influenced the quality of the text correction on the ISYC dataset, we carried out a set of experiments in order to: 1) evaluate if a domain N-gram contributes to the performance of text correction in the presence of a general spell checker, and 2) select the best combination of Hunspell and the domain N-gram model by varying the α parameter.

4. Discussion

In Table 1, we show the obtained results, where α = 0 corresponds to the case where the system only considers the information provided by Hunspell. We analyzed the results

Table 1. Results obtained using Hunspell and domain 3-Gram over ISYC dataset. For Hits, greater is better; for Misses and FP, lower is better, and for FN, greater is worse.

	α value						
	0	**0.01**	**0.02**	**0.03 - 0.09**	**0.1**	**0.2**	**0.3**
HITS	617	623 ▲0.97%	632 ▲1.46%	**634** **▲1.78%**	634 ▲1.78%	630 ▲1.13%	630 ▲1.13%
MISSES	816	810 ▼0.86%	806 ▼1.23%	**798** **▼1.47%**	802 ▼0.98%	802 ▼0.98%	802 ▼0.98%
FP	366	288 ▼13.11%	288 ▼13.11%	**288** **▼13.11%**	288 ▼13.11%	288 ▼13.11%	288 ▼13.11%
FN	239	240 ▲0.42%	240 ▲0.42%	**240** **▲0.42%**	240 ▲0.42%	240 ▲0.42%	240 ▲0.42%

considering the typical matrix of confusion: Hits (the system correction is equal to the real word), Misses (the system correction is different to the real word), False Positives (FP) (the system corrects a word that was already correct), and False Negatives (FN) (the system does not correct a word that was incorrect). We can see that the optimal value of α in Equation 1 according to our test data must be in the range (0.03 - 0.1), and the results are stabilized from 0.1 and onwards. Regarding misses, FP and FN, they are mainly due to Hunspell failures.

5. Concluding Remarks

In this paper, we have described the design of a text cleaning system for correcting errors of damaged scanned documents. Also, we have enhanced the correction step by using an N-gram model that weights all the suggestions made by a well-known spell checker. The word sequences which frequently appear in a particular document typology make our system to be able to perform a highly adapted word-level correction for that kind of documents. As a future work, we want to improve the correction process by adding new methods and tools, for example, with the incorporation of a different spell-check, a fine grained machine learning system and natural language processing techniques like rule based grammars in the correction step. We also plan to test the behavior in other scenarios different from the scope of legal documents, since the initial requirements are the same.

Acknowledgements

This research work has been supported by projects TIN2013-46238-C4-4-R, TIN2016-78011-C4-3-R (AEI/FEDER, UE), and DGA/FEDER. We want to thank ISYC, Dr. Carlos Bobed, and María G. Buey for their collaboration in this work.

References

1] W. G. Kruse II, J. G. Heiser, Computer forensics: incident response essentials. Pearson Education (2001). ISBN 0-201-70719-5.

2] C.D. Manning, H. Schütze, Foundations of statistical natural language processing. MIT Press (1999). doi: 10.1017/S1351324902212851

3] C. Biemann, A. Mehler, Text Mining: From Ontology Learning to Automated Text Processing Applications. Springer (2014) ISBN 978-3-319-12654-8. doi: 10.1007/978-3-319-12655-5

4] M. G. Buey, A.L., Garrido, C. Bobed, S. Ilarri, The AIS Project: Boosting Information Extraction from Legal Documents by using Ontologies. Proceedings of the 8th International Conference on Agents and Artificial Intelligence (2016) 438-445. doi: 10.5220/0005757204380445

5] M. G. Buey, C. Roman, A. L. Garrido, C. Bobed, and E. Mena, Automatic Legal Document Analysis: Improving the Results of Information Extraction Processes using an Ontology. Studies of Big Data, Springer (to be published in 2018)

6] S. B. Needleman, and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of molecular biology, 48(3) (1970) 443-453. doi: 10.1016/0022-2836(70)90057-4

7] J. H. Martin, and D. Jurafsky. Speech and language processing. International Edition, (2000).710. ISBN 978-0-13-187321-6

8] http://hunspell.github.io/ Last accessed: December 2017.