A Study of Reuse and Plagiarism in LREC papers

Gil Francopoulo [1], Joseph Mariani [2], Patrick Paroubek [2]
1 LIMSI, CNRS, Université Paris-Saclay + Tagmatica (France)
2 LIMSI, CNRS, Université Paris-Saclay (France)

gil.francopoulo@wanadoo.fr, joseph.mariani@limsi.fr, pap@limsi.fr

Abstract

The aim of this experiment is to present an easy way to compare fragments of texts in order to detect (supposed) results of copy & paste operations between articles in the domain of Natural Language Processing (NLP). The search space of the comparisons is a corpus labeled as NLP4NLP gathering a large part of the NLP field. The study is centered on LREC papers in both directions, first with an LREC paper borrowing a fragment of text from the collection, and secondly in the reverse direction with fragments of LREC documents borrowed and inserted in the collection.

Keywords: Plagiarism Detection, Natural Language Processing

1. Introduction

Everything starts with a copy & paste and, of course the flood of documents that we see today could not exist without the practical ease of copy & paste. This is not new but what is new is that the availability of archives allows us to study a vast amount of papers in our domain (i.e. Natural Language Processing, NLP, both for written and spoken materials) and to figure out the level of reuse and plagiarism.

2. Context

Our work comes after the various studies initiated in the Workshop entitled: "Rediscovering 50 Years of Discoveries in Natural Language Processing" on the occasion of ACL's 50th anniversary in 2012 [Radev et al 2013] where a group of researchers studied the content of the corpus recorded in the ACL Anthology [Bird et al 2008]. Among these studies, one was devoted to reuse and it is worth quoting Gupta and Rosso [Gupta et al 2012]: "It becomes essential to check the authenticity and the novelty of the submitted text before the acceptance. It becomes nearly impossible for a human judge (reviewer) to discover the source of the submitted work, if any, unless the source is already known. Automatic plagiarism detection applications identify such potential sources for the submitted work and based on it a human judge can easily take the decision". Let's add that this subject is a specific and active domain ruled yearly by the PAN international plagiarism detection competition[1].

3. Objectives

Our aim is not to present the state-of-art or to compare the various metrics and algorithms for reuse and plagiarism detection, see [Hoad et al 2003] [HaCohen-Kerner et al 2010] for instance. In order to avoid any misunderstanding, we position our work as an extrinsic detection, the aim of which is to find near-matches between texts, as opposed to intrinsic detection whose aim is to show that different parts of a presumably single-author text could not have been written by the same author [Stamatatos et al 2011a], [Stein et al 2011], [Bensalem et al 2014].

In contrast, our main objective **is to deal with the entry level of the detection**. The main question is: Is there a meaningful difference in taking the verbatim raw strings compared with the result of a linguistic parsing? A secondary objective is to present and study a series of ascertainments about the practices of our field.

4. The corpus: NLP4NLP

The corpus is a large content of our own research field, i.e. NLP, covering both written and speech sub-domains and extended to a limited number of corpora, for which Information Retrieval and NLP activities intersect. This corpus was collected at IMMI-CNRS and LIMSI-CNRS (France) and is named NLP4NLP[2]. It currently contains 65,003 documents coming from various conferences and journals with either public or restricted access. This is a large part of the existing published articles in our field, apart from the workshop proceedings and the published books. The time period spans from 1965 to 2015. Broadly speaking, and aside from the small corpora, one third comes from the ACL Anthology[3], one third from the ISCA Archive[4] and one third from IEEE[5].

The corpus follows the organization of the ACL Anthology with two parts in parallel. For each document, on one side, the metadata is recorded with the author names and the title under the form of a BibTex file. On the other side, the PDF document is recorded on disk in its original form. Each document is labeled with a unique identifier, for instance "lrec2000_1" is reified on the hard disk as two files: "lrec2000_1.bib" and "lrec2000_1.pdf". When recorded as an image, the PDF content is extracted by means of Tesseract OCR[6]. The automatic test leading to the call (or not) of the OCR is implemented by means of some Apache PDFBox API calls[7]. For all the other documents, other PDFBox API calls are applied in order to extract the textual content.

[1] http://pan.webis.de

[2] www.nlp4nlp.org
[3] http://aclweb.org/anthology
[4] www.isca-speech.org/iscaweb/index.php/archive/online-archive
[5] https://www.ieee.org/index.html
[6] https://code.google.com/p/tesseract-ocr
[7] https://pdfbox.apache.org

The detail is presented in table 1, as follows:

short name	# docs	format	long name	language	access to content	period	# venues
acl	4264	conference	Association for Computational Linguistics Conference	English	open access *	1979-2015	37
acmtslp	82	journal	ACM Transaction on Speech and Language Processing	English	private access	2004-2013	10
alta	262	conference	Australasian Language Technology Association	English	open access *	2003-2014	12
anlp	278	conference	Applied Natural Language Processing	English	open access *	1983-2000	6
cath	932	journal	Computers and the Humanities	English	private access	1966-2004	39
cl	776	journal	American Journal of Computational Linguistics	English	open access *	1980-2014	35
coling	3813	conference	Conference on Computational Linguistics	English	open access *	1965-2014	21
conll	842	conference	Computational Natural Language Learning	English	open access *	1997-2015	18
csal	762	journal	Computer Speech and Language	English	private access	1986-2015	29
eacl	900	conference	European Chapter of the ACL	English	open access *	1983-2014	14
emnlp	2020	conference	Empirical methods in natural language processing	English	open access *	1996-2015	20
hlt	2219	conference	Human Language Technology	English	open access *	1986-2015	19
icassps	9819	conference	IEEE International Conference on Acoustics, Speech and Signal Processing - Speech Track	English	private access	1990-2015	26
ijcnlp	1188	conference	International Joint Conference on NLP	English	open access *	2005-2015	6
inlg	227	conference	International Conference on Natural Language Generation	English	open access *	1996-2014	7
isca	18369	conference	International Speech Communication Association	English	open access	1987-2015	28
jep	507	conference	Journées d'Etudes sur la Parole	French	open access *	2002-2014	5
lre	308	journal	Language Resources and Evaluation	English	private access	2005-2015	11
lrec	4552	conference	Language Resources and Evaluation Conference	English	open access *	1998-2014	9
ltc	656	conference	Language and Technology Conference	English	private access	1995-2015	7
modulad	232	journal	Le Monde des Utilisateurs de L'Analyse des Données	French	open access	1988-2010	23
mts	796	conference	Machine Translation Summit	English	open access	1987-2015	15
muc	149	conference	Message Understanding Conference	English	open access *	1991-1998	5
naacl	1186	conference	North American Chapter of the ACL	English	open access *	2000-2015	11
paclic	1040	conference	Pacific Asia Conference on Language, Information and Computation	English	open access *	1995-2014	19
ranlp	363	conference	Recent Advances in Natural Language Processing	English	open access *	2009-2013	3
sem	950	conference	Lexical and Computational Semantics / Semantic Evaluation	English	open access *	2001-2015	8
speechc	593	journal	Speech Communication	English	private access	1982-2015	34
tacl	92	journal	Transactions of the Association for Computational Linguistics	English	open access *	2013-2015	3
tal	177	journal	Revue Traitement Automatique du Langage	French	open access	2006-2015	10
taln	1019	conference	Traitement Automatique du Langage Naturel	French	open access *	1997-2015	19
taslp	6612	journal	IEEE/ACM Transactions on Audio, Speech and Language Processing	English	private access	1975-2015	41
tipster	105	conference	Tipster DARPA text program	English	open access *	1993-1998	3
trec	1847	conference	Text Retrieval Conference	English	open access	1992-2015	24
Total	67937					1965-2015	558

Table 1 Detail of NLP4NLP, with the convention that an asterisk indicates that the corpus is in the ACL Anthology

A semi-automatic cleaning process was applied on the metadata in order to avoid false duplicates concerning middle names (for X Y Z, is Y a second given name or the first part of the family name?) and for this purpose, we use the specific BibTex format where the given name is separated from the family name with a comma. Then typographic variants (e.g. “Jean-Luc” vs “Jean Luc” or “Herve” vs “Hervé”) were searched and false duplicates were normalized in order to be merged. The resulting number of different authors is 48,894. Figures are not extracted because we are unable to compare images. See [Francopoulo et al 2015] for more details about the extraction process as well as the solutions for some tricky problems like joint conferences management or abstract / body / reference sections detection.

The majority (90%) of the documents come from conferences, the rest coming from journals. The overall number of words is roughly 270M. Initially, the texts are in four languages: English, French, German and Russian. The number of texts in German and Russian is less than 0.5%. They are detected automatically and are ignored. The texts in French are a little bit more numerous (3%), so they are kept with the same status as the English ones. This is not a problem as our tool is able to process English and French.

The corpus is a collection of documents of a single technical domain which is NLP in the broad sense, and of course, some conferences are specialized in certain topics like written processing, speech processing, information retrieval or machine translation, but these specializations do not imply for an author the need to duplicate works accross conferences. It is an important issue as this statement does not apply when the corpus is multi-domains because the audience is then so different that publishing the same work in different places appears necessary.

5. Definitions

As the terminology is fuzzy and contradictory among the scientific literature, we need first to define four important terms in order to avoid any misunderstanding.

The term "**self-reuse**" is used for a copy & paste when the source of the copy has an author who belongs to the group of authors of the text of the paste and when the source is cited.
The term "**self-plagiarism**" is used for a copy & paste when the source of the copy has similarly an author who belongs to the group of authors of the text of the paste, but when the source is not cited.
The term "**reuse**" is used for a copy & paste when the source of the copy has no author in the group of authors of the paste and when the source is cited.
The term "**plagiarism**" is used for a copy & paste when the source of the copy has no author in the group of the paste and when the source is not cited.

Said in other words, the terms "self-reuse" and "reuse" qualify a situation with a proper source citation, on the contrary of "self-plagiarism" and "plagiarism". Let's note that in spite of the fact that the term "self-plagiarism" seems to be contradictory, we use this term because it is the usual habit within the community of the plagiarism detection. Some authors also use the term "recycling", for instance [HaCohen-Kerner et al 2010].

6. Directions

Another point to clarify concerns the expression "in LREC papers" within our title. As a convention, we call "focus" the corpus which is the center of the study: here LREC. The whole NL4NLP collection will be the search space. We examine the copy & paste operations in both directions. We study the configuration with an LREC paper borrowing a fragment of text from the NLP4NLP collection, in other words, a backward study. But we also study the reverse direction with fragments of LREC documents borrowed and inserted in the NLP4NLP collection, in other words, a forward study.

7. Algorithm

Comparison of word sequences has proven to be an effective method for detection of copy & paste [Clough et al 2002a] and in several occasions, this method won the PAN contest [Barron-Cedeno et al 2010], so we will adopt this strategy. The corpus is first processed with the deep NLP parser TagParser [Francopoulo 2007] to produce a Passage format [Vilnat et al 2010] with lemma and part-of-speech (POS) indications.

The algorithm is not very complex and is as follows:

- For each document of the focus (here LREC), all the sliding windows [8] of 7 lemmas (excluding punctuations) are built and recorded under the form of a character string key in an index locally to a document.
- An index gathering all these local indexes is built and is called the "focus index".
- For each document apart from the focus (i.e. outside LREC), all the sliding windows are built and **only the windows** contained in the focus index are recorded in an index locally to a document. This filtering operation is done to optimize the comparison phase, as there is no need to compare the windows out of the focus index.
- Then, the keys are compared to compute a similarity overlapping score [Lyon et al 2001] between documents D1 and D2, with the Jaccard distance: **score(D1,D2) = shared windows# / union# (D1 windows, D2 windows).** The pairs of documents D1 / D2 are then filtered according to a threshold of 0.04 to retain only significant scoring situations.

8. Algorithm comments and evaluation

In a first implementation, we compared the raw character strings with a segmentation based on space and punctuation. But, due to the fact that the input is the result of PDF formatting, the texts may contain variable caesura for line endings or some little textual variations. Our objective is to compare at a higher level than hyphen variation (there are different sorts of hyphens), caesura (the sequence X/-/endOfLine/Y needs to match an entry XY in the lexicon to distinguish from an hyphen binding a composition), upper/lower case variation, plural, orthographic variation ("normalise" vs "normalize"), spellchecking (particularly useful when the PDF is an image and when the extraction is of low quality) and abbreviation ("NP" vs "Noun Phrase" or "HMM" vs "Hidden Markov Model"). Some rubbish sequence of characters (e.g. a series of hyphens) are also detected and cleaned.

Given that a parser takes all these variations and cleanings into account, we decided to apply a full linguistic parsing. The syntactic structures and relations are ignored. Then a module for entity linking is called in order to bind different names referring to the same entity, a process often labeled as "entity linking" in the literature [Guo et al 2011][Moro et al 2014]. This process is based on a Knowledge Base called "Global Atlas" [Francopoulo et al 2013] which comprises the LRE Map [Calzolari et al 2012]. Thus "British National Corpus" is considered as possibly abbreviated to "BNC", as well as less regular names like "ItalWordNet" possibly abbreviated to "IWN". Each entry of the Knowledge Base has a canonical form, possibly associated with different variants: the aim is to normalize into a canonical form to neutralize proper noun obfuscations based on variant substitutions. After this processing, only the sentences with at least a verb are considered.

We examined the differences between the two strategies concerning all types of copy & paste situations above the threshold, as mentioned Table 2, with the last column adding the two other columns without the duplicates produced by the couples of the same year.

[8] Also called "n-grams" in some NLP publications.

Strategy	Backward study document pairs#	Forward study document pairs#	Backward+forward document pairs# after duplicate pruning
Raw text	438	373	578
Linguistic processing (LP)	559	454	736
Difference (LP-raw)	121	81	158

Table 2 Levels of source with the same parameters

The strategy based on linguistic processing provides more pairs (158) and we examined these differences. Among these pairs, the vast majority (80%) concerns caesura: this is normal because most conferences demand a double column format, so the authors frequently use caesura to save place[9]. The other differences (20%) are mainly caused by lexical variations and spellchecking. Thus, the results show that using raw texts gives a more "silent" system. The drawback is that the computation is much longer[10], but we think that it is worth the value.

9. Tuning parameters

There are three parameters that had to be tuned: the window size, the distance function and the threshold. The main problem we had was that we did not have any gold standard to evaluate the quality specifically on our corpus and the burden to annotate a corpus is too heavy. We therefore decided to start from the parameters presented in the articles related to the PAN contest. We then computed the results, picked a random selection of pairs that we examined and tuned the parameters accordingly.

In the PAN related articles, different window sizes are used. A window of five is the most frequent one [Kasprzak et al 2010], but our results shows that a lot of common sequences like "the linguistic unit is the" overload the pairwise score. After some trials, we decided to select a size of seven.

Concerning the distance function, the Jaccard distance is frequently used but let's note that other formulas are applicable and documented in the literature. For instance, some authors use an approximation with the following formula: score(D1,D2) = shared windows# / min(D1 windows#, D2 windows#) [Clough et al 2009], which is faster to compute, because there is no need to compute the union. Given that computation time is not a problem for us, we kept the most used function which is the Jaccard distance.

Concerning the threshold, we tried thresholds of 0.03 and 0.04 and we compared the results. The last value gave more significant results, as it reduced noise, while still allowing to detect meaningful pairs of similar papers.

10. Special considerations concerning authorship and citations

As previously explained, our aim is to distinguish a copy & paste fragment associated with a citation compared to a fragment without any citation. To this end, we proceed with an approximation: we do not bind exactly the anchor in the text, but we parse the reference section and consider that, globally to the text, the document cites (or not) the other document. Due to the fact, that we have proper author identification for each document, the corpus forms a complex web of citations. We are thus able to distinguish self-reuse vs self-plagiarism and reuse vs plagiarism. We are in a situation slightly different from METER where the references are not linked. Let's recall that METER is the corpus usually involved in plagiarism detection competitions [Gaizauskas et al 2001][Clough et al 2002b].

11. Precision about the anteriority test

Given the fact that some papers and drafts of papers can circulate among researchers before the official published date, it is impossible to verify exactly when a document is issued; moreover we do not have any more detailed time indication than the year, as we don't know the date of submission. This is why we also consider the same year within the comparison.

12. Resulting files

The program computes a detailed result for each interesting pair of documents as an HTML page with the common fragments displayed as red highlighted snippets with HTML links back to the original documents[11]. The program produces also a global result, with the convention that only the corpora of the same language of the focus (here English) are presented.

13. LREC papers from and to LREC papers

Concerning the similarity of each LREC paper with other LREC papers that were published earlier or in the same year, we didn't find any case of plagiarism. Considering self-reuse and self-plagiarism, Only 66 couples of papers were detected (1.5% of papers published at LREC), ranging from a similarity ratio of 4% up to 46%. Half of the papers cite the source paper. In many cases, the similar parts are related to the presentation of a method, a program, a project, a problem or a resource shared by the two papers. In one case, the coverage is extensive and the difference is primarily in the name of the system being presented, while the description is almost the same! It appears that 20% of those papers were published in the same LREC conference and that 68% concern the (self-)reuse of a paper published at the previous conference. Only 12% reuse material from a longer period prior, as shown in Table 3.

[9] Concerning this specific problem, for instance, PACLIC and COLING which are one column formatted give much better extraction quality than LREC and ACL which are two columns formatted.

[10] It takes 25 hours instead of 3 hours on a mid-range mono-processor Xeon E3-1270 V2 with 32G of RAM.

[11] But the space limitations do not allow to present these results in lengthy details. And we do not want to display personal results.

(self-) reused \ (self-) reusing	1998	2000	2002	2004	2006	2008	2010	2012	2014	Total
2000	3	2								5
2002		4	1							5
2004		1	3	3						7
2006			2	10	2					14
2008			1		6	1				8
2010				1	1	5				7
2012						1	9	2		12
2014				1				5	2	8
Total	3	7	7	15	9	7	9	7	2	66

Table 3 LREC papers borrowing/being borrowed by LREC papers

14. Results: LREC papers borrowing NLP4NLP papers, i.e. backward study

The focus of the computation being here the LREC papers, all the LREC papers of a given year (let's say year Y) will be compared with the other papers older or equal to this year Y within the whole NLP4NLP collection, including LREC itself. The results appear in Table 4.

If we want to study the influence a given conference (or journal) has on another, we must recall that these figures are raw figures in terms of number of documents, and we must not forget that some conferences (or journals) are much bigger than others, for instance ISCA is a conference with more than 18K documents compared to LRE which is a journal with 308 documents. **When we look at the five top sources (marked in the extreme right column), we see that the main sources of "inspiration" for LREC papers are ISCA and LREC itself. Then come COLING, ACL and LTC[12].**

It appears that only two cases of possible plagiarism were detected, but we found after checking manually that, in the two cases, both papers referred with the same wording to the content of a third previous paper which described the method they used and that they both properly acknowledged. A set of 554 documents (about 12% of the papers published at LREC) have been reused by their authors with or without citing the source paper. Only 37% of the papers cite the source paper. The maximum degree of similarity is 89%, corresponding to the description of the same research center in two different conferences. Forty percent of the document pairs (224 papers) involve papers published in the same year: similar papers that may have been simultaneously submitted at different conferences, or additionally in a journal. 49% fall within a window of the 2 previous years, while only 11% span a longer timeframe.

[12] Which may be due to the biennial frequency and the proximity in time of the two conferences.

source copy	# self-reuse	# self-plagiarism	# reuse	# plagiarism	total	top 5
acl	15	45	0	0	60	4
acmtslp	0	1	0	0	1	
alta	1	3	0	0	4	
anlp	1	4	0	0	5	
cath	1	2	0	0	3	
cl	0	3	2	2	7	
coling	16	47	0	0	63	3
conll	4	5	0	0	9	
csal	0	7	0	0	7	
eacl	11	9	0	0	20	
emnlp	7	18	0	0	25	
hlt	5	8	0	0	13	
icassps	5	10	0	0	15	
ijcnlp	11	5	0	0	16	
inlg	0	5	0	0	5	
isca	38	79	1	0	118	1
lre	4	2	0	0	6	
lrec	34	32	0	0	66	2
ltc	12	21	0	0	33	5
mts	11	9	0	0	20	
muc	0	0	0	0	0	
naacl	1	1	0	0	2	
paclic	6	12	0	0	18	
ranlp	10	6	0	0	16	
sem	3	7	0	0	10	
speechc	1	1	0	0	2	
tacl	2	0	0	0	2	
taslp	0	2	0	0	2	
tipster	1	1	0	0	2	
trec	4	5	0	0	9	
total	204	350	3	2	559	

Table 4 LREC papers borrowing NLP4NLP papers

15. Results: LREC papers borrowed by NLP4NLP papers, i.e. forward study

The focus of the computation still being the LREC papers, all the LREC papers of a given year (let's say Y) will be compared with the other papers younger or equal to this year Y. The result is in Table 5.

The following table shows that the main conference "inspired" by LREC is LREC. Then come ISCA, ACL, COLING and, not surprisingly, LRE as papers published at the LREC conference may be invited or may take the initiative to submit in the LRE journal. It is also interesting to notice a regular flow from LREC papers to other journals (Computational Linguistics, Computer Speech and Language) and increasingly to IEEE ICASSP**.**

Seven cases of possible plagiarism were detected, but here also it appeared that they correspond to the content of a third paper where a method, a corpus, a platform they used, an evaluation campaign or a project they participated in is described, and that they both properly acknowledge. In some cases, the authors are different but belong to the same laboratory. A set of 445 documents

(about 10% of the LREC papers) have been reused by their authors with or without citing the source paper. Only 41% of the papers cite the source paper. The maximum degree of similarity is 75%. Fifty percent of the pairs (the same 224 papers as above) include papers published in the same year, 35% fall within a window of the 2 previous years, while only 15% span a longer time scale.

target paste	# self-reuse	# self-plagiarism	# reuse	# plagiarism	total	top 5
acl	26	28	0	0	54	3
acmtslp	1	2	0	0	3	
alta	0	0	0	0	0	
anlp	0	2	0	0	2	
cath	2	5	0	0	7	
cl	3	11	0	0	14	
coling	23	29	0	0	52	4
conll	1	3	1	0	5	
csal	0	13	0	1	14	
eacl	4	10	0	0	14	
emnlp	5	9	0	2	16	
hlt	5	5	0	0	10	
icassps	4	9	0	0	13	
ijcnlp	0	6	0	0	6	
inlg	0	2	0	0	2	
isca	27	33	0	1	61	2
lre	17	34	0	0	51	5
lrec	34	32	0	0	66	1
ltc	6	4	0	0	10	
mts	4	2	0	0	6	
muc	0	0	0	0	0	
naacl	1	1	0	0	2	
paclic	5	7	0	0	12	
ranlp	0	3	1	1	5	
sem	7	4	0	0	11	
speechc	1	4	0	1	6	
tacl	1	1	0	0	2	
taslp	2	3	0	0	5	
tipster	0	1	0	0	1	
trec	3	0	0	1	4	
total	182	263	2	7	454	

Table 5 NLP4NLP papers borrowing LREC papers

16. Discussion

The first obvious ascertainment is that self-reusing is much more important than reusing the content of others. With a comparable threshold of 0.04, when we consider the total of the two directions, there are 386 self-reuse and 613 self-plagiarism pairs, including the 224 duplicates, compared with 5 reuse and 9 plagiarism pairs. Within self-reuse and self-plagiarism, there are slightly more LREC papers borrowing (55%) than being borrowed, and, globally, the source papers are quoted only in 39% of the cases on average, a percentage which falls down from 49% to 25% if the papers are published on the same year. Let's recall that self-reuse and self-plagiarism concern a pair of authors with a given author in the intersection of the authors. Of course, a copy & paste operation is easy and frequent but there is another phenomena to take into account which is difficult to distinguish from copy & paste: this is the style of the author. Everybody has habits to formulate its ideas, and, even on a long period, most authors seem to keep the same chunks of prepared words.

As a tentative to moderate these figures and to justify self-reuse and self-plagiarism of previously published material, it is worth quoting Pamela Samuelson [Samuelson 1994]:

- The previous work must be restated to lay the groundwork for a new contribution in the second work,
- Portions of the previous work must be repeated to deal with new evidence or arguments,
- The audience for each work is so different that publishing the same work in different places is necessary to get the message out,
- The author thinks they said it so well the first time that it makes no sense to say it differently a second time.

17. Further developments

A limitation of our approach is that it fails to identify copy & paste when the original text has been strongly altered. Our study of graphical variations of a common meaning is presently limited to geographical variants, technical abbreviations (e.g. HMM vs Hidden Markov Model) and resource names aliases from the LRE Map. We plan to deal with "rogeting" which is the practice of replacing words with supposedly synonymous alternatives in order to disguise plagiarism[13] by obfuscation, see [Potthast et al 2010][Chong et al 2011][Ceska et al 2009] for another presentation. Detecting paraphrases and transpositions of passive / active sentences, seem in contrast rather difficult to implement [Barron-Cedeno et al 2013]. A more tractable development is to artificially modify the n-gram to match as presented in [Nawab et al 2012]. Another track of development could be to simplify the input to retain only the plain words, a process labeled as "stopwords n-gram" by [Stamatatos 2011b]. Another direction of improvement is to isolate and ignore tables in order to reduce noise, but this is a complex task as documented in [Frey et al 2015]. Let's note that this is not a big problem in our approach, as we ignore sentences without any verb and as verbs are not very frequent within a table,. Finally, we also plan to extend the present comparison of a single source to the whole collection to the full comparison of all sources of the collection.

18. Conclusion

To our knowledge, this paper is the first which reports results on the study of copy & paste operations on corpora of NLP archives of this size. Based on a simple method of n-gram comparison, this method is rather easy to implement. Of course, this process makes a large number of pairwise comparisons, but this is not a practical limitation for a modern computer. Extending the study of one source to all sources may be more demanding in terms of computing power.

As our measures show, self-plagiarism is a common

[13] https://en.wikipedia.org/wiki/Rogeting

practice in our field. This is not specific to our field, at the present. This is certainly related to the current tendency which is called "salami-slicing" publication caused by the publish-and-perish demand. But we gladly notice that plagiarism is very uncommon.

19. Acknowledgements

We'd like to thank Wolfgang Hess for the ISCA archive, Douglas O'Shaughnessy, Denise Hurley, Rebecca Wollman and Casey Schwartz for the IEEE data, Nicoletta Calzolari, Helen van der Stelt and Jolanda Voogd for the LRE Journal articles, Olivier Hamon and Khalid Choukri for the LREC proceedings, Nicoletta Calzolari, Irene Russo, Riccardo Del Gratta, Khalid Choukri for the LRE Map, Min-Yen Kan for the ACL Anthology, Florian Boudin for the TALN proceedings and Ellen Voorhees for the TREC proceedings.

20. Bibliographical References

Barron-Cedeno Alberto, Potthast Martin, Rosso Paolo, Stein Benno, Eiselt Andreas (2010). Corpus and Evaluation Measures for Automatic Plagiarism Detection, Proceedings of LREC, Valletta, Malta.

Barron-Cedeno Alberto, Vila Marta, Marti Maria Antonia, Rosso Paolo (2013). Plagiarism Meets Paraphrasing Insights for the Next Generation in Automatic Plagiairism Detection, Computational Linguistics.

Bensalem Imene, Rosso Paolo, Chikhi Salim (2014). Intrinsic Plagiarism Detection using N-gram Classes, Proceedings of the Conference on Empirical Methods in Natural Language Processing, Doha, Qatar.

Bird Steven, Dale Robert, Dorr Bonnie J, Gibson Bryan, Joseph Mark T, Kan Min-Yen, Lee Dongwon, Powley Brett, Radev Dragomir R, Tan Yee Fan (2008). The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics, Proceedings of LREC, Marrakech, Morocco.

Calzolari Nicoletta, Del Gratta Riccardo, Francopoulo Gil, Mariani Joseph, Rubino Francesco, Russo Irene, Soria Claudia (2012). The LRE Map. Harmonising Community Descriptions of Resources, Proceedings of LREC, Istanbul, Turkey.

Ceska Zdenek, Fox Chris (2009). The Influence of Text Pre-processing on Plagiarism Detection, Proceedings of the Recent Advances in Natural Language Processing, Borovets, Bulgaria.

Chong Miranda, Specia Lucia (2011). Lexical Generalisation for Word-level Matching in Plagiarism Detection, Proceedings of Recent Advances in Natural Language Processing, Hissar, Bulgaria.

Clough Paul, Gaizauskas Robert, Piao Scott S L, Wilks Yorick (2002a). Measuring Text Reuse. Proceedings of ACL'02, Philadelphia, USA.

Clough Paul, Gaizauskas Robert, Piao Scott S L, (2002b). Building and annotating a corpus for the study of journalistic text reuse, Proceedings of LREC, Las Palmas, Spain.

Clough Paul, Stevenson Mark (2009). Developing a Corpus of Plagiarised Short Answers, Language Resources and Evaluation, Springer.

Francopoulo Gil (2007). TagParser: well on the way to ISO-TC37 conformance. Proceedings of ICGL (International Conference on Global Interoperability for Language Resources), Hong Kong.

Francopoulo Gil, Marcoul Frédéric, Causse David, Piparo Grégory (2013). Global Atlas: Proper Nouns, from Wikipedia to LMF, in LMF Lexical Markup Framework (Francopoulo, ed), ISTE Wiley.

Francopoulo Gil, Mariani Joseph, Paroubek Patrick (2015). NLP4NLP: the cobbler's children won't go unshod, in D-Lib Magazine: The magazine of Digital Library Research[14].

Frey Matthias, Kern Roman (2015). Efficient Table Annotation for Digital Articles, in D-Lib Magazine: The magazine of Digital Library Research[15].

Gaizauskas Robert, Foster Jonathan, Wilks Yorick, Arundel John, Clough Paul, Piao Scott S L (2001). The METER Corpus: A Corpus for Analysing Journalistic Text Reuse. Proceedings of the Corpus Linguistics Conference, Lancaster, UK.

Guo Yuhang, Che Wanxiang, Liu Ting, Li Sheng (2011). A Graph-based Method for Entity Linking, International Joint Conference on NLP, Chiang Mai, Thailand.

Gupta Parth, Rosso Paolo (2012). Text Reuse with ACL: (Upward) Trends, Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries, Jeju, Republic of Korea.

Hoad Timothy C, Zobel Justin (2003). Methods for identifying Versioned and Plagiarised Documents, Journal of the American Society for Information Science and Technology.

HaCohen-Kerner Yaakov, Tayeb Aharon, Ben-Dror Natan (2010). Detection of Simple Plagiarism in Computer Science Papers, in Proceedings of the 23rd International Conference on Computational Linguistics (COLING), Beijing, PRC.

Kasprzak Jan, Brandejs Michal (2010). Improving the Reliability of the Plagiarism Detection System Lab, in Proceedings of the Uncovering Plagiarism, Authorship and Social Software Misuse (PAN), Padua, Italy.

Lyon Caroline, Malcolm James, Dickerson Bob (2001). Detecting Short Passages of Similar Text in large

[14] www.dlib.org/dlib/november15/francopoulo/11francopoulo.html
[15] www.dlib.org/dlib/november15/frey/11frey.html

document collections, Proceedings of the Empirical Methods in Natural Language Processing Conference, Pittsburgh, PA USA.

Moro Andrea, Raganato Alessandro, Navigli Roberto (2014). Entity Linking meets Word Sense Disambiguation : a Unified Approach, Transactions of the Association for Computational Linguistics.

Nawab Rao Muhammad Adeel, Stevenson Mark, Clough Paul (2012). Detecting Text Reuse with Modified and Weighted N-grams, First Joint Conference on Lexical and Computational Semantics, Montréal, Canada.

Potthast Martin, Stein Benno, Barron-Cedeno Alberto, Rosso Paolo (2010). An Evaluation Framework for Plagiarism Detection, in Proceedings of the 23rd International Conference on Computational Linguistics (COLING), Beijing, PRC.

Radev Dragomir R, Muthukrishnan Pradeep, Qazvinian Vahed, Abu-Jbara, Amjad (2013). The ACL Anthology Network Corpus, Language Resources and Evaluation 47: 919–944, Springer.

Samuelson Pamela (1994). Self-plagiarism or fair use? Communications of the ACM 37 (8):21-5.

Stamatatos Efstathios, Koppel Moshe (2011a). Plagiarism and authorship analysis: introduction to the special issue, Language Resources and Evaluation, Springer.

Stamatatos Efstathios (2011b). Plagiarism detection using stopword n-grams. Journal of the American Society for Information Science and Technology.

Stein Benno, Lipka Nedim, Prettenhofer Peter (2011). Intrinsic plagiarism analysis, Language Resources and Evaluation, Springer.

Vilnat Anne, Paroubek Patrick, Villemonte de la Clergerie Eric, Francopoulo Gil, Guénot Marie-Laure (2010). PASSAGE Syntactic Representation: a Minimal Common Ground for Evaluation. Proceedings of LREC 2010, Valletta, Malta.