A Study of Reuse and Plagiarism in Speech and Natural Language Processing papers

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

joseph.mariani@limsi.fr, gil.francopoulo@wanadoo.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, including Speech Processing (NLP). The search space of the comparisons is a corpus labelled as NLP4NLP, which includes 34 different sources and gathers a large part of the publications in the NLP field over the past 50 years. This study considers the similarity between the papers of each individual source and the complete set of papers in the whole corpus, according to four different types of relationship (self-reuse, self-plagiarism, reuse and plagiarism) and in both directions: a source paper borrowing a fragment of text from another paper of the collection, or in the reverse direction, fragments of text from the source paper being borrowed and inserted in another paper of the collection.

Keywords: Plagiarism Detection, Text reuse, Natural Language Processing, Speech Processing, Scientometrics, Informetrics

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 in this area.

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]. On our side, we also conducted a specific study of reuse and plagiarism in the papers published at the Language Resources and Evaluation conference (LREC), from 1998 to 2014 [Francopoulo et al 2016].

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. 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 specific field.

4. The corpus: NLP4NLP

The corpus is a large content of our own research field, i.e. NLP, covering both written and spoken language processing 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 50 years from 1965 to 2015. Broadly

[1] http://pan.webis.de
[2] www.nlp4nlp.org

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 detail of NLP4NLP 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	67,937[6]					1965-2015	577
Total without duplicates	65,003					1965-2015	558

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

A phase of preprocessing has been applied to represent the various sources in a common format. This format follows the organization of the ACL Anthology with two parts in parallel for each document: the metadata and the content. 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".

For the metadata, we faced four different types of sources with different flavors and character encodings: BibTeX (e.g. ACL Anthology), custom XML (e.g. TALN), database downloads (e.g. IEEE) or HTML program of the conference (e.g. TREC). We wrote a series of small Java programs to transform these metadata into a common BibTeX format under UTF8. Each file comprises the author names and the title. The file is located in a directory which designates the year and the corpus.

Concerning the content, we faced different formats possibly for the same corpus, and the amount of documents being huge, we cannot designate the file type by hand individually. To deal with this, we wrote a program to self-detect the type and sub-type as follows:

- A small amount of texts are in raw text: we keep them in this format.
- The vast majority of the documents are in PDF format of different sub-types. First, we used PDFBox[7] to determine the sub-type of the PDF content: when the content is a textual content, we use PDFBox

[3] http://aclweb.org/anthology
[4] www.isca-speech.org/iscaweb/index.php/archive/online-archive
[5] https://www.ieee.org/index.html
[6] In the case of a joint conference, the papers are counted twice. This number reduces to 65,003, if we count only once duplicated papers. Similarly, the number of venues is 577 when all venues are counted, but this number reduces to 558 when the 19 joint conferences are counted only once.

again to extract the text, possibly with the use of the "Legion of the Bouncy Castle"[8] to extract the encrypted content. When the PDF is a text under the form of an image, we use PDFBox to extract the images and then Tesseract OCR[9] to transform the images into a textual content.

Then, and after some experiments, two filters are applied to avoid getting rubbish content:

- The content should be at least 900 characters.
- The content should be of good quality. In order to evaluate this quality, the content is analyzed by the morphological module of TagParser [Francopoulo 2007], a deep industrial parser based on a broad English lexicon and Global Atlas (a knowledge base containing more than one million words from 18 Wikipedias) [Francopoulo et al. 2013] to detect out-of-the-vocabulary (OOV) words. Based on the hypothesis that rubbish strings are OOV words, we retain a text when the ratio OOV / number of words is less than 9%.

We then apply a set of symbolic rules to split the abstract, body and reference section. The file is recorded in XML. It should be noted that we made some experiments with other strategies, given the fact that we are able to compare them with respect to a quantitative evaluation of the quality, as explained before. The first experiment was to use ParsCit[10] [Councill et al. 2008] but the evaluation of the quality was bad, specially when the content is not pure ASCII. The result on accentuated Latin strings, or Arabic and Russian contents was awful. We also tried Grobid[11] but we did not succeed to run it correctly on Windows.

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" versus "Jean Luc" or "Herve" versus "Hervé") were searched in a tedious process 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%), and 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 language processing, spoken language processing, including signal processing, information retrieval or machine translation.

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 as authors should be free to use their own wordings, we use this term because it is the usual habit within the community of 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 "source papers". As a convention, we call "focus" the corpus corresponding to the source which is studied. The whole NL4NLP collection is the "search space". We examine

[7] https://pdfbox.apache.org

[8] http://www.bouncycastle.org/

[9] https://code.google.com/p/tesseract-ocr

[10] https://github.com/knmnyn/ParsCit

[11] https://github.com/kermitt2/grobid

the copy & paste operations in both directions: we study the configuration with a source paper borrowing fragments of text from other papers of the NLP4NLP collection, in other words, a backward study, and we also study in the reverse direction the fragments of the source paper being borrowed by papers of 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. In our case, 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 as follows:

- For each document of the focus (the source corpus), all the sliding windows[12] of lemmas (typically 5 to 7, 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 the source corpus), all the sliding windows are built and **only the windows** contained in the focus index are recorded in an index locally to this 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 in order to retain only significant similarity 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" versus "normalize"), spellchecking (particularly useful when the PDF is an image and when the extraction is of low quality) and abbreviation ("NP" versus "Noun Phrase" or "HMM" versus "Hidden Markov Model"). Some rubbish sequence of characters (e.g. a series of hyphens) were also detected and cleaned.
Given that a parser takes all these variations and cleanings into account, we decided to apply a full linguistic parsing, as a second strategy. 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 the "Global Atlas" Knowledge Base [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 those two strategies concerning all types of copy & paste situations above the threshold, choosing the LREC source as the focus. The results are presented in Table 2, with the last column adding the two other columns without the duplicates produced by the couples of the same year.

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

Table 2. Comparison of the two strategies on the LREC corpus

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[13]. The other differences (20%) are

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

[13] 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.

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[14], 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 was 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. All experiments were conducted with LREC as the focus and NLP4NLP as the search space.

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 show 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 tokens, in agreement with [Citron and Ginsparg 2014].

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 (3 to 4%) 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.

After running the first trials, we discovered that using the Jaccard distance resulted in considering as similar a set of two papers, one of them being of small content. This may be the case for invited talks, for example, when the author only provide a short abstract. In this case, a simple acknowledgement to the same institution may produce a similarity score higher than the threshold. The same happens for some eldest papers when the OCR produced a truncated document. In order to solve this problem, we added a second threshold on the minimum number of shared windows that we set at 50 after considering the corresponding erroneous cases.

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 versus self-plagiarism and reuse versus 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 precise date of submission. This is why we also consider the same year within the comparisons. In this case, it is difficult to determine which are the borrowing and borrowed papers, and in some cases they may even have been written simultaneously. However, if one paper cites a second one, while it is not cited by the second one, it may serve as a sign to consider it as being the borrowing paper.

12. Resulting files

The program computes a detailed result for each individual source as an HTML page where all similar pairs of documents are listed with their similarity score, with the common fragments displayed as red highlighted snippets and HTML links back to the original 67,937 documents[15]. For each of the 4 categories (Self-reuse, Self-Plagiarism, Reuse and Plagiarism), the program produces the list of couples of “similar” papers according to our criteria, with their similarity score, and the global results in the form of matrices displaying the number of papers

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

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

that are similar in each couple of the 34 sources, in the forward and backward directions (the using sources are on the X axis, while the used sources are on the Y axis). The total of used and using papers, and the difference between those totals, are presented, while the 7 (Table 3) or 5 (Table 4) top using or used sources are indicated in green.

We conducted a manual checking of the couples of papers showing a very high similarity: the 14 couples that showed a similarity of 1 were the duplication of a paper due to an error in editing the proceedings of a conference. We also found after those first trials erroneous results of the OCR for some eldest papers which resulted in files containing several papers, in full or in fragments, or where blanks were inserted after each individual character. We excluded those 86 documents from the corpus being considered.

Checking those results, we also mentioned several cases where the author was the same, but with a different spelling, or where references were properly quoted, but with a different wording, a different spelling (American English versus British English, for example) or an improper reference to the source. We had to manually correct those cases, and move the corresponding couples of papers in the correct category (from reuse or plagiarism to self-reuse or self-plagiarism in the case of authors names, from plagiarism to reuse, in the case of references).

13. Self-reuse and Self-Plagiarism

Table 3 provides the results of merging self-reuse (authors reusing their own text while quoting the source paper) and self-plagiarism (authors reusing their own text without quoting the source paper). As we see, it is a rather frequent phenomenon, with a total of 12,493 documents (i.e. 18% of the 67,937 documents!). In 61% of the cases (7,650 self-plagiarisms over 12,493), the authors do not quote the source paper. We found that 205 papers have exactly the same title, and that 130 papers have both the same title and the same list of authors! Also 3,560 papers have exactly the same list of authors. Given the large number of documents, it is impossible to conduct a manual checking of all the couples.

We see that the most used sources are the large conferences: ISCA, IEEE-ICASSP, ACL, COLING, HLT, EMNLP and LREC. The most using sources are not only those large conferences, but also the journals: IEEE-Transactions on Acoustics, Speech and Language Processing (and its various avatars) (TASLP), Computer Speech and Language (CSAL), Computational Linguistics (CL) and Speech Com. If we consider the balance between the using and the used sources, we clearly see that the flow of papers goes from conferences to journals. The largest flows of self-reuse and self-plagiarism concern ISCA and ICASSP, in both directions, but especially from ISCA to ICASSP, ICASSP and ISCA to TASLP (also in the reverse direction) and to CSAL, ISCA to Speech Com, ACL to Computational Linguistics, ISCA to LREC and EMNLP to ACL.

If we want to study the influence a given conference (or journal) has on another one, we must however 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, LREC is a conference with more than 4,500 documents compared to LRE which is a journal with only 308 documents. If we relate the number of published papers that reuse another paper to the total number of published papers, we may see that 17% of the LRE papers (52 over 308) use content coming from the LREC conferences, without quoting them in 66% of the cases. Also the frequency of the conferences (annual or biennial) and the calendar (date of the conference and of the submission deadline) may influence the flow of papers between the sources.

The similarity scores range from 4% to 97% (Fig. 1). We see that about 4,500 couples of papers have a similarity score equal or superior to 10%; about 900 (1.3% of the total number of papers) have a score superior or equal to 30%. Looking at the ones with the largest similarity score, we found a few examples of important variants in the spelling of the same authors' names, and cases of republishing the corrigendum of a previously published paper or of republishing a paper with a small difference in the title and one missing author in the authors' list. In one case, the same research center is described by the same author in two different conferences with an overlapping of 90%. In another case, the difference of the two papers is primarily in the name of the systems being presented, funded by the same project agency in two different contracts, while the description has a 45% overlap!

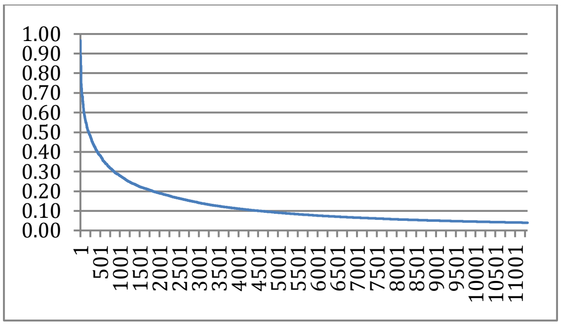

Fig. 1 Similarity scores of the couples detected as self-reuse / self-plagiarism

Used Using	acl	acmtslp	alta	anlp	cath	cl	coling	conll	csal	eacl	emnlp	hlt	icassps	ijcnlp	inlg	isca	jep	lre	lrec	ltc	modulad	mts	muc	naacl	paclic	ranlp	sem	speechc	tacl	tal	taln	taslp	tipster	trec	Total used	Total using	Difference	
acl	22	8	1	4	8	136	78	25	31	22	83	85	29	31	7	48	0	20	71	4	0	19	1	51	8	5	26	1	2	0	0	24	4	9	863	625	238	acl
acmtslp	1	0	0	0	0	0	0	0	2	0	0	2	3	2	0	6	0	1	1	0	0	0	0	2	0	0	1	0	1	0	0	2	0	0	24	93	-69	acmtslp
alta	3	0	2	0	0	1	5	0	1	2	5	0	0	1	0	4	0	0	4	0	0	0	0	0	0	1	0	0	0	0	0	0	0	4	33	14	19	alta
anlp	7	0	0	1	3	5	8	1	1	2	1	4	0	0	0	1	0	0	5	0	0	1	0	2	1	0	0	0	0	0	0	0	2	5	50	50	0	anlp
cath	1	0	0	1	7	2	0	0	0	1	0	1	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	18	50	-32	cath
cl	9	0	0	4	3	0	4	0	2	4	3	1	0	0	0	0	0	2	5	0	0	0	0	0	1	0	4	0	0	0	0	0	0	0	42	433	-391	cl
coling	74	10	3	8	7	62	19	24	17	15	43	49	8	24	7	42	0	14	90	4	0	9	2	33	12	5	25	3	0	0	0	12	6	5	632	500	132	coling
conll	26	1	1	1	1	20	18	8	5	6	16	11	2	14	2	2	0	2	10	1	0	3	0	7	0	5	13	0	1	0	0	3	0	0	179	151	28	conll
csal	3	0	0	0	0	4	4	2	7	0	3	2	20	1	0	35	0	2	7	0	0	0	0	0	0	0	2	6	0	0	0	13	0	0	111	643	-532	csal
eacl	16	2	0	2	5	31	12	6	3	1	8	13	3	1	2	9	0	0	21	1	0	1	0	13	1	1	4	0	0	0	0	5	0	1	162	130	32	eacl
emnlp	103	2	2	1	2	44	52	26	18	9	16	30	14	47	1	27	0	5	29	0	0	7	0	22	2	1	19	0	3	0	0	20	1	5	508	355	153	emnlp
hlt	83	12	0	5	3	48	48	11	42	14	33	22	29	30	2	104	0	4	26	1	0	13	2	6	1	0	9	8	0	0	0	25	7	19	607	476	131	hlt
icassps	16	5	0	0	0	3	4	1	130	4	7	21	262	2	0	1005	0	0	19	0	0	2	0	14	2	0	0	65	0	0	0	746	0	3	2311	2160	151	icassps
ijcnlp	27	6	1	0	0	3	29	10	7	2	34	18	2	4	3	7	0	5	19	3	0	9	0	13	4	8	3	0	0	0	0	4	0	1	222	237	-15	ijcnlp
inlg	7	0	0	1	1	6	5	2	0	3	1	3	0	1	2	4	0	1	6	0	0	1	0	4	0	0	0	0	0	0	0	1	0	0	49	35	14	inlg
isca	56	23	0	2	0	13	45	0	317	10	25	116	1531	10	4	879	0	10	133	19	0	12	0	38	6	0	1	233	0	0	0	669	0	5	4157	2460	1697	isca
jep	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	16	18	-2	jep
lre	2	1	0	0	0	2	3	0	0	0	0	1	0	0	0	2	0	2	6	0	0	0	0	1	1	0	0	0	0	0	0	1	0	0	22	146	-124	lre
lrec	58	3	0	2	6	16	80	6	13	15	16	17	16	10	2	72	0	52	67	12	0	6	0	11	11	4	12	5	2	0	0	6	1	3	524	660	-136	lrec
ltc	4	0	0	0	0	0	0	0	0	0	0	0	0	2	0	15	0	1	35	10	0	2	0	0	6	6	1	4	0	0	0	0	0	0	86	71	15	ltc
modulad	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	modulad
mts	13	0	0	0	0	2	9	2	0	2	9	10	3	9	0	9	0	2	20	2	0	8	0	8	5	2	1	1	0	0	0	2	0	0	119	109	10	mts
muc	2	0	0	2	0	2	3	0	0	1	0	7	0	0	0	0	0	0	0	0	0	0	10	1	0	0	0	0	0	0	0	0	18	1	47	28	19	muc
naacl	46	10	0	2	1	24	30	7	12	11	22	5	15	22	3	30	0	3	16	1	0	9	0	3	0	0	9	1	0	0	0	8	0	3	293	251	42	naacl
paclic	4	0	0	0	1	0	12	1	1	1	1	0	2	8	0	3	0	5	18	7	0	3	0	0	21	7	1	0	0	0	0	1	0	0	97	85	12	paclic
ranlp	3	2	0	0	0	0	2	4	4	2	2	1	0	7	0	0	0	2	19	5	0	2	0	1	2	4	2	1	0	0	0	0	0	1	66	54	12	ranlp
sem	25	2	0	0	0	7	16	14	4	1	12	12	0	8	0	0	0	13	12	1	0	1	0	8	1	4	53	0	0	0	0	0	0	1	195	188	7	sem
speechc	0	0	0	0	0	1	1	0	11	0	0	4	17	0	0	48	0	0	2	0	0	0	0	0	0	0	0	1	0	0	0	17	0	0	102	344	-242	speechc
tacl	1	1	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	7	9	-2	tacl
tal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	13	0	0	0	18	59	-41	tal
taln	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	53	9	0	0	0	65	22	43	taln
taslp	0	5	0	0	0	0	1	1	13	0	1	4	197	0	0	103	0	0	2	0	0	1	0	2	0	0	0	15	0	0	0	49	0	0	394	1610	-1216	taslp
tipster	3	0	0	3	0	0	6	0	0	0	1	5	0	0	0	0	0	0	2	0	0	0	13	1	0	0	0	0	0	0	0	0	2	7	43	65	-22	tipster
trec	10	0	4	11	2	1	6	0	2	2	11	32	7	3	0	5	0	0	10	0	0	0	0	10	0	1	1	0	0	0	0	2	24	287	431	362	69	trec
Total using	625	93	14	50	50	433	500	151	643	130	355	476	2160	237	35	2460	18	146	660	71	0	109	28	251	85	54	188	344	9	59	22	1610	65	362	12493	12493	0	

Table 3. Self-reuse and Self-Plagiarism Matrix, with indication of the 7 most using and used sources.

Used Using	acl	acmtslp	alta	anlp	cath	cl	coling	conll	csal	eacl	emnlp	hlt	icassps	ijcnlp	inlg	isca	jep	lre	lrec	ltc	modulad	mts	muc	naacl	paclic	ranlp	sem	speechc	tacl	tal	taln	taslp	tipster	trec	Total used	Total using	Difference	
acl	1	0	0	0	1	1	2	2	0	0	4	3	0	3	0	2	0	0	1	1	0	0	1	1	1	1	3	0	0	0	0	0	0	0	28	7	21	acl
acmtslp	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	acmtslp
alta	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	alta
anlp	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	anlp
cath	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	-2	cath
cl	0	0	0	0	0	0	1	0	0	0	1	1	0	1	0	0	0	0	4	0	0	1	0	1	2	0	0	0	0	0	0	0	0	0	12	5	7	cl
coling	0	0	0	0	1	0	0	0	0	0	0	2	1	1	0	2	0	0	2	0	0	0	1	1	1	0	2	0	0	0	0	1	0	0	15	7	8	coling
conll	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	3	5	-2	conll
csal	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	3	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	7	6	1	csal
eacl	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	0	eacl
emnlp	0	0	0	0	0	2	0	2	0	1	1	2	0	0	0	0	0	0	0	0	0	1	0	2	0	0	0	0	0	0	0	0	0	2	13	15	-2	emnlp
hlt	2	0	0	0	0	1	0	1	1	0	2	1	1	1	0	2	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0	2	17	17	0	hlt
icassps	0	0	0	0	0	0	0	0	1	0	1	2	3	0	0	32	0	0	0	0	0	0	0	2	0	0	0	2	0	0	0	5	0	0	48	37	11	icassps
ijcnlp	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	2	9	-7	ijcnlp
inlg	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	inlg
isca	0	0	0	0	0	1	1	0	1	0	0	1	18	1	0	7	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	3	0	0	36	70	-34	isca
jep	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	jep
lre	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	-1	lre
lrec	0	0	0	0	0	0	0	0	1	0	2	0	1	0	0	1	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1	8	8	0	lrec
ltc	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	-4	ltc
modulad	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	modulad
mts	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	4	3	1	mts
muc	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	3	3	0	muc
naacl	1	0	0	0	0	0	0	0	1	0	1	1	1	0	0	1	0	0	0	0	0	0	0	1	2	0	0	0	0	0	0	0	0	0	9	10	-1	naacl
paclic	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	10	-8	paclic
ranlp	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	-3	ranlp
sem	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	3	7	-4	sem
speechc	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	4	5	-1	speechc
tacl	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	tacl
tal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	tal
taln	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	taln
taslp	0	0	0	0	0	0	0	0	1	1	0	0	10	0	0	16	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	30	10	20	taslp
tipster	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	2	2	0	tipster
trec	1	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	13	13	0	trec
Total using	7	0	0	0	2	5	7	5	6	2	15	17	37	9	0	70	0	1	8	4	0	3	3	10	10	3	7	5	0	0	0	10	2	13	261	261	0	

Table 4. Reuse and Plagiarism Matrix, with indication of the 5 most using and used sources

14. Reuse and Plagiarism

Table 4 provides the results of merging reuse (authors reusing fragments of the texts of other authors while quoting the source paper) and plagiarism (authors reusing fragments of the texts of other authors without quoting the source paper). As we see, there are very few cases altogether. Only 261 papers (i.e. less than 0.4% of the 67,937 documents) reuse a fragment of papers written by other authors that they quote. In 60% of the cases (156 plagiarisms over 261), the authors do not quote the source paper, but these possible cases of plagiarism only represent 0.23% of the total number of papers. Given those small numbers, we were able to conduct a manual checking of those couples.

Among the couple papers placed in the "Reuse" category, it appeared that 12 have a least one author in common, but with a somehow different spelling and should therefore be placed in the "Self-reuse" category. Among the couples of papers placed in the "Plagiarism" category, 25 have a least one author in common, but with a somehow different spelling and should therefore be placed in the "Self-plagiarism" category and 14 correctly quote the source paper, but with variants in the spelling of the authors' names, of the paper's title or of the conference or journal source or forgetting to place the source paper in the references and should therefore be placed in the "Reuse" category. It therefore resulted in 107 cases of "reuse" and 117 possible cases of plagiarism (0.17% of the papers) that we studied more closely. We found the following explanations:

- The paper cites another reference from the same authors of the source paper (typically a previous reference, or a paper published in a Journal) (46 cases)
- Both papers use extracts of a third paper that they both cite (31 cases)
- The authors of the two papers are different, but from the same laboratory (typically in industrial laboratories or funding agencies) (11 cases)
- The authors previously co-authored papers (typically as supervisor and PhD student or postdoc) but are now in a different laboratory (11 cases)
- The authors of the papers are different, but collaborated in the same project which is presented in the two papers (2 cases)
- The two papers present the same short example, result or definition coming from another source (13 cases)

If we exclude those cases, only 3 cases of possible plagiarism remain that correspond to the same paper which appears as a patchwork of 3 other papers, while sharing several references with them.

The similarity scores range from 4% to 42% (Fig. 2). Only 34 couples of papers have a similarity score equal or higher than 10%. For example, the couple showing the highest similarity score comprises a paper published in 1998 and a paper published in 2000 which both describe *Chart parsing* using the words of the initial paper published 20 years earlier in 1980, that they both properly quote. Among the three remaining possible cases of plagiarism, the highest similarity score is 10%, with a shared window of 200 tokens.

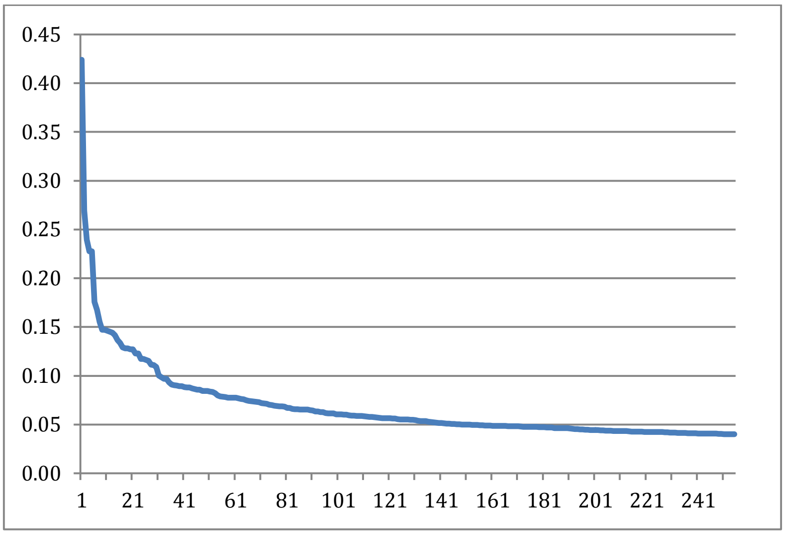

Fig. 2 Similarity scores of the couples detected as reuse / plagiarism

15. Time delay between publication and reuse

We now consider the duration between the publication of a paper and its reuse (in all 4 categories) in another publication. It appears that 38% of the similar papers were published on the same year, 71% within the next year, 83% over 2 years and 93% over 3 years (Figure 3 and 4). Only 7% reuse material from an earlier period. The average duration is 1.22 years. 30% of the similar papers published on the same year concern the couple of conferences ISCA-ICASSP.

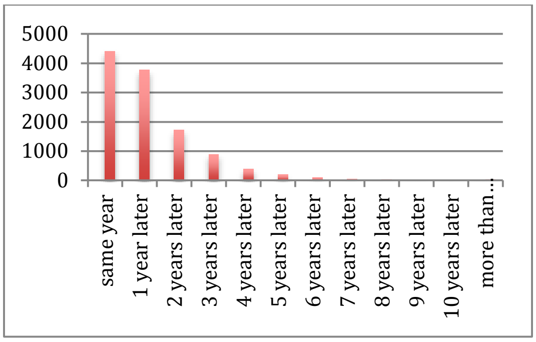

Fig. 3 Time delay between publication and reuse

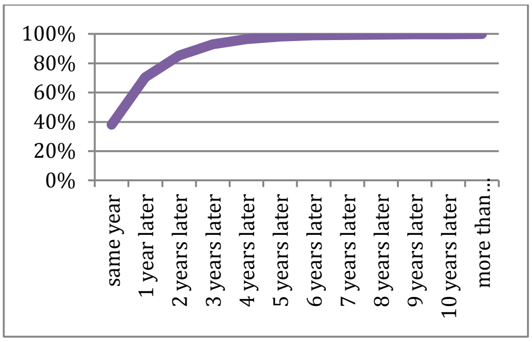

Fig. 4 Time delay between publication and reuse (in %)

We now consider the reuse of conference papers in journal papers (Figures 5 and 6). We observe here a similar time schedule, with a delay of one year: 12% of the reused papers were published on the same year, 41% within the next year, 68% over 2 years, 85% over 3 years and 93% over 4 years. Only 7% reuse material from an earlier period. The average duration is 2.07 years.

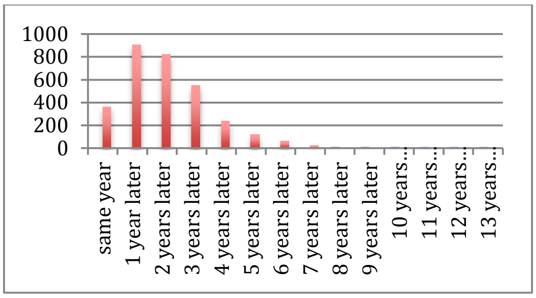

Fig. 5 Time delay between publication in conferences and reuse in journals

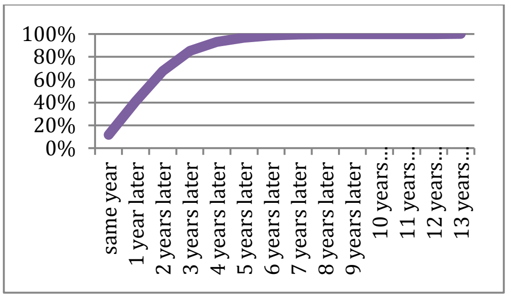

Fig. 6 Time delay between publication in conferences and reuse in journals (in %)

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 4843 self-reuse and 7650 self-plagiarism detected pairs, compared with 105 reuse and 156 plagiarism detected pairs. Globally, the source papers are quoted only in 39% of the cases on average, a percentage which falls down from 39% to 23% if the papers are published on the same year.

Plagiarism may raise legal issues if it violates copyright, but the *right to quote*[16] exists in certain conditions: "National legislations usually embody the Berne convention limits in one or more of the following requirements:

- the cited paragraphs are within a reasonable limit,
- clearly marked as quotations and fully referenced,
- the resulting new work is not just a collection of quotations, but constitutes a fully original work in itself",
- we could also add that the cited paragraph must have a function in the goal of the citing paper.

Obviously, most of the cases reported in this paper comply with the right to quote. The *limits of the cited paragraph* vary from country to country. In France and Canada, for example, a limit of 10% of both the copying and copied texts seems to be acceptable. As we've seen, we stay within those limits in all cases in NLP4NLP.

Self-reuse and self-plagiarism are of a different nature. Let's recall that they concern papers that have at least one author in common. 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 we've seen, almost 40% of the cases concern papers that are published on the same year: authors submit two similar papers at two different conferences on the same year, and publish the two papers in

[16] https://en.wikipedia.org/wiki/Right_to_quote

both conferences if both are accepted. It is very difficult to prevent those cases as none of the papers are published when the other is submitted. Another frequent case is the publication of a paper in a journal after its publication in a conference. Here also, it is a natural and usual process, sometimes even encouraged by the journal editors after a pre-selection of the best papers in a conference.

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 authors think they said it so well the first time that it makes no sense to say it differently a second time.

She considers that 30% is an upper limit in the reuse of parts of a previously published paper.

We believe that following these two sets of principles regarding (self) reuse and plagiarism will help maintaining an ethical behavior in our community.

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 versus 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[17] 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, seems 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.

More generally, we could also study the position and rhetorical structure of the copy & paste in order to identify and justify their function.

We may finally explore whether copy & paste is more common for non native English speakers, given that it is frequent that they publish first in their native language at a national conference and then in English in an international conference or an international journal, in order to broaden their audience.

18. Conclusions

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 after text processing using NLP, this method is easy to implement. Of course, this process makes a large number of pairwise comparisons (65,000*65,000), which still represents a practical computing limitation.

As our measures show, self-reuse and self-plagiarism are common practices. This is not specific to our field and is certainly related to the current tendency which is called "salami-slicing" publication caused by the publish-and-perish demand[18]. But we gladly notice that plagiarism is very uncommon in our community.

19. Bibliographical references

1. 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.
2. Barron-Cedeno Alberto, Vila Marta, Marti Maria Antonia, Rosso Paolo (2013). Plagiarism Meets Paraphrasing Insights for the Next Generation in Automatic Plagiarism Detection, Computational Linguistics.
3. 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.
4. 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.
5. Calzolari Nicoletta, Del Gratta Riccardo, Francopoulo Gil, Mariani Joseph, Rubino Francesco, Russo Irene, Soria

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

[18] To this regard, we must ourselves admit that the reader will find a certain degree of overlapping between this paper and the one we published at LREC 2016 also on reuse and plagiarism, but specifically related to the LREC papers, at least on the description of the NLP4NLP corpus.

Claudia (2012). The LRE Map. Harmonising Community Descriptions of Resources, Proceedings of LREC, Istanbul, Turkey.

6. 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.
7. Chong Miranda, Specia Lucia (2011). Lexical Generalisation for Word-level Matching in Plagiarism Detection, Proceedings of Recent Advances in Natural Language Processing, Hissar, Bulgaria.
8. Citron Daniel T., Ginsparg Paul (2014). Patterns of text reuse in a scientific corpus, PNAS 2015 112 (1) 25-30; published ahead of print December 8, 2014, doi:10.1073/pnas.1415135111
9. Clough Paul, Gaizauskas Robert, Piao Scott S L, Wilks Yorick (2002a). Measuring Text Reuse. Proceedings of ACL'02, Philadelphia, USA.
10. 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.
11. Clough Paul, Stevenson Mark (2009). Developing a Corpus of Plagiarised Short Answers, Language Resources and Evaluation, Springer.
12. Councill, Isaac G., Giles, C. Lee and Kan, Min-Yen (2008), ParsCit: An open-source CRF reference string parsing package. In Proceedings of the Language Resources and Evaluation Conference (LREC 2008), Marrakesh, Morocco, May 2008
13. 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.
14. 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.
15. 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[19].
16. Francopoulo Gil, Mariani Joseph, Paroubek Patrick (2016). A Study of Reuse and Plagiarism in LREC papers. Proceedings of LREC 2016, Portorož, Slovenia.
17. Frey Matthias, Kern Roman (2015). Efficient Table Annotation for Digital Articles, in D-Lib Magazine: The magazine of Digital Library Research[20].
18. 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.
19. Guo Yuhang, Che Wanxiang, Liu Ting, Li Sheng (2011). A Graph-based Method for Entity Linking, International Joint Conference on NLP, Chiang Mai, Thailand.
20. 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.
21. Hoad Timothy C, Zobel Justin (2003). Methods for identifying Versioned and Plagiarised Documents, Journal of the American Society for Information Science and Technology.
22. 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.
23. 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.
24. Lyon Caroline, Malcolm James, Dickerson Bob (2001). Detecting Short Passages of Similar Text in large document collections, Proc. of the Empirical Methods in Natural Language Processing Conference, Pittsburgh, PA USA.
25. Moro Andrea, Raganato Alessandro, Navigli Roberto (2014). Entity Linking meets Word Sense Disambiguation: a Unified Approach, Transactions of the Association for Computational Linguistics.
26. 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.
27. 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.
28. Radev Dragomir R, Muthukrishnan Pradeep, Qazvinian Vahed, Abu-Jbara, Amjad (203). The ACL Anthology Network Corpus, Language Resources and Evaluation 47: 919–944, Springer.
29. Samuelson Pamela (1994). Self-plagiarism or fair use? Communications of the ACM 37 (8):21-5.
30. Stamatatos Efstathios, Koppel Moshe (2011a). Plagiarism and authorship analysis: introduction to the special issue, Language Resources and Evaluation, Springer.
31. Stamatatos Efstathios (2011b). Plagiarism detection using stopword n-grams. Journal of the American Society for Information Science and Technology.
32. Stein Benno, Lipka Nedim, Prettenhofer Peter (2011). Intrinsic plagiarism analysis, Language Resources and Evaluation, Springer.
33. 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.

[19] www.dlib.org/dlib/november15/francopoulo/11francopoulo.html

[20] www.dlib.org/dlib/november15/frey/11frey.html