Redundancy in French Electronic Health Records: A preliminary study

Eva d'Hondt, Xavier Tannier, Aurélie Névéol

Redundancy in French Electronic Health Records: A preliminary study

Eva D'hondt
LIMSI-CNRS UPR 3251
Rue John von Neuman
91403 Orsay, France
dhondt@limsi.fr

Xavier Tannier
LIMSI-CNRS UPR 3251
Univ. Paris-Sud
Rue John von Neuman
91403 Orsay, France
xtannier@limsi.fr

Aurélie Névéol
LIMSI-CNRS UPR 3251
Rue John von Neuman
91403 Orsay, France
neveol@limsi.fr

Abstract

The use of Electronic Health Records (EHRs) is becoming more prevalent in healthcare institutions world-wide. These digital records contain a wealth of information on patients' health in the form of Natural Language text. The electronic format of the clinical notes has evident advantages in terms of storage and shareability, but also makes it easy to duplicate information from one document to another through copy-pasting. Previous studies have shown that (copy-paste-induced) redundancy can reach high levels in American EHRs, and that these high levels of redundancy have a negative effect on the performance of Natural Language Processing (NLP) tools that are used to process EHRs automatically. In this paper, we present a preliminary study on the level of redundancy in French EHRs. We study the evolution of redundancy over time, and its occurrence in respect to different document types and sections in a small corpus comprising of three patient records (361 documents). We find that average redundancy levels in our subset are lower than those observed in U.S. corpora (respectively 33% vs. up to 78%), which may indicate different cultural practices between these two countries. Moreover, we find no evidence of the incremental increase (over time) of redundant text in clinical notes which has been found in American EHRs. These results suggest that redundancy mitigating strategies may not be needed when processing French EHRs.

1 Introduction

Electronic Health Records (EHRs) are becoming prevalent in most healthcare institutions and have been recognized to contain crucial information about patients' health in the form of Natural Language text. As a result, specialized Natural Language Processing (NLP) methods and tools are being developed to unlock this wealth of medical information from EHRs and use it in medical applications such as clinical decision support (Demner-Fushman et al., 2009). The electronic format of the clinical notes in the patient records makes it easy to duplicate information from one document to another through copy-pasting methods. Previous studies (Wrenn et al., 2010; Zhang et al., 2011; Cohen et al., 2013) have shown that the amount of redundancy introduced by copy-pasting could reach up to 78% in clinical notes from American hospitals. This situation makes it difficult to exploit the content of EHRs both for humans and NLP tools (Cohen et al., 2013).

One motivation for introducing content redundancy in clinical documents is a need for completeness: While a patient's history grows over time with each new hospital visit, the key information remains the same. By copy-pasting all available information into the most recent document, this becomes a stand-alone document which offers a complete and up-to-date overview of the patient's history and status. Other reasons for the observed redundancy are more pragmatic: logistical issues such as the time to load previous documents from the EHR or restricted access rights to documents created in a different hospital department can lead health professionals to ensure that all relevant information is present in the clinical note they are currently writing.

Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis (Louhi), pages 21–30,
Lisbon, Portugal, 17 September 2015.

These practices can make health professionals more efficient but they may also represent potential risks to patient care by creating confusion between what is relevant to the patient's current versus his or her past medical condition (Siegler and Adelman, 2009; Weis and Levy, 2014). Interestingly, in the process of copying information from an older document into a new one, small changes can be introduced into the narrative, such as typo corrections, acronym expansion, and information updates. For these reasons it is important to study the nature and extent of redundancy in clinical notes in more detail (Zhang et al., 2014). An important goal is to identify which portions of a clinical document are entirely new and which portions are redundant from previous documents in the records, and whether the redundant portions are identical to previous documents, or modified.

To our knowledge, the issue of redundancy in Electronic Records has been predominantly studied for English, and more specifically in documents produced in healthcare institutions in the United States. In this paper, we present a preliminary study that addresses redundancy in French clinical narratives from a group of healthcare institutions in France. We analyse a corpus of 361 documents from 3 patient records and examine to what extent and under which conditions redundancy is present.

2 Objective and research questions

The goal of this study is to characterize redundancy in French EHRs with a view to gauge its impact on NLP processing. This work is motivated by the findings of similar studies on US EHRs, which have shown that a certain level of redundancy (30% and more) affects word distribution in the documents and therefore has a (negative) impact on language models. Also, when annotating data, redundancy could lead to duplicate work, which we do want to avoid.

It is important to note that our study focuses exclusively on *surface redundancy*, which we define as text sections of a document that are copied verbatim (or with marginal edits) to another document. While surface redundancy, i.e. literal copying, entails redundant information, the reverse is not necessarily true: In a patient's records, the same information may be repeated in various documents but using different wording. Although this type of paraphrasing may be considered as conveying redundant information, it does not have a negative impact on language models, because the utterances are part of the natural language diversity of expression.

In this paper we aim to answer the following research questions:

- Does surface redundancy grow over time in a patient's records?
- Which parts of the documents contain the most redundancy?
- Are certain types of documents in EHRs more likely to contain redundant information than others?
- Is there more redundancy within patient's records versus between patient's records?

We expect the results of this study to provide some insight on the suitable natural language processing methods to apply to French EHRs. In particular, in light of the research conducted on US clinical notes, we need to understand the nature of redundancy in French data in order to decide whether redundancy mitigation strategies are needed.

3 Background

Redundancy detection is closely related to the research topic of plagiarism detection, but there are some key differences between the two fields. Since redundancy is introduced by (indiscriminate) copying of text without much human intervention, redundancy detection is mainly focused on literal string matching, rather than employing semantic similarity measures (other than detecting spelled-out variants of acronyms) or paraphrase detection. Furthermore, redundancy detection is usually performed within a closed reference collection (as opposed to plagiarism detection systems that use the entire internet as a reference base).

It is important to note that near-duplicate blocks of texts that are copied from a source text can occur in different positions in the new document. As a result, similarity measures that treat the whole document as one string i.e., global alignment, are not optimal for redundancy detection (see also discussion in Zhang et al. (2011)).

Table 1 presents an overview of the tools and methods whose suitability for redundancy detection in our corpus we reviewed in the course of

this study. The Baldr[1] and Sherlock[2] (Mozgovoy et al., 2005) software packages have been developed for plagiarism detection exclusively. Baldr is a source-code plagiarism-detecting software that uses 'information distance' (Vitányi et al., 2009) to measure similarity between two documents. The intuition underlying this distance is that two objects (in this case text documents) are similar if the transformation function to transform one document into the other is simple to describe. If, however, all such functions are complex, the objects are deemed dissimilar. Baldr uses real-world compression software to calculate the transformation metrics (Chen et al., 2004). The Sherlock software package uses the more common method of fingerprinting, i.e. hashing substrings of the text into unique digital signatures. Redundancy is then calculated as the proportion of common signatures between an incoming document and documents in the comparison set. Unlike the method described below for Cohen et al. (2013), the Sherlock program operates on word level, i.e. uses words as its basic units, instead of characters.

The other methods in Table 1 were developed specifically for the task of redundancy assessment in EHRs. They were all applied on corpora from healthcare institutions in the United States. Cohen et al. (2013) developed a character-based fingerprinting method similar to the BLAST sequence similarity method (Altschul et al., 1990) which is popular in bioinformatics. When applied to a subset of a large corpus of 22,500 patient notes, they observed an average level of redundancy of around 30% within patient records, but a much lower amount of redundancy (on average 2.9%) between patient records. They also found that redundancy in a large corpus has a significant negative effect on the performance of language modelling applications.

Wrenn et al. (2010) developed a token-based Levenshtein edit distance measure to perform sequence alignment between two documents. The reported redundancy score is the proportion of aligned tokens over the total number of tokens in the base document. In their study they looked at the occurrence of redundancy over time in a corpus of 100 EHRs (admissions) and found levels of redundancy between 54% and 78% depending on the clinical note (i.e. document) type. Furthermore they noted that the level of redundancy consistently increased over time in the corpus.

Zhang et al. (2011) used vector-based semantic similarity measures to measure redundancy in outpatient notes. They analysed a corpus of notes from 178 patients and found that these notes contain a large amount of redundancy. Like Wrenn et al. (2010) they also studied time progression, and observed that note redundancy increased over time.

4 Material and Methods

4.1 Corpus

For this study, we used a set of French clinical notes where personally identifying information (PII) had been marked and replaced by surrogates (Grouin and Névéol, 2014). The documents were also marked with four types of content sections: letterhead, patient header, content and footer (Deléger et al., 2014). One of our goals is to assess whether there is more redundancy in notes belonging to one patient, compared to redundancy in notes across patient records. To this end we selected three complete patient records for our corpus. These records contain a total of 361 documents. Each record comprised of at least 100 documents and tracks the treatment of a patient over the course of several years. To allow for a fair comparison of redundancy within and between patient records, we selected three patient records with similar profiles, i.e. patients that were admitted for renal transplantation and follow-up care.

4.2 Measuring redundancy

For comparability with previous work, we measured corpus redundancy using the fingerprinting method (Cohen et al., 2013). We also developed our own fingerprinting method, which is an extension of the Cohen method: Like the Cohen measure our implementation calculates a similarity score based on the proportion of n-character fingerprints which a target document has in common with a base document or collection, over the total number of fingerprints in the target document. In other words, it shows what proportion of the text (expressed in fingerprints) is redundant, i.e. has also appeared in the base document. Unlike Cohen's method our implementation allows for the extraction of overlapping fingerprints, which im-

[1] http://wassner.blogspot.fr/2014/05/baldr-loutil-anti-fraude-anti-plagiat.html

[2] http://www2.warwick.ac.uk/fac/sci/dcs/research/ias/software/sherlock/

name	method	score range	time-ordering	comparison
Cohen et al.	non-overlapping fingerprints	0-1	no	document-pairs
adaptedCohen	overlapping fingerprints	0-1	yes	document-pairs, corpus
Wren et al.	Levenstein distance	0-1	yes	corpus
Zhang et al.	semantic similarity	0-1	yes	corpus
Baldr	compression range	0-1	no	document-pairs
Sherlock	overlapping fingerprints	0-100	no	document-pairs

Table 1: Overview of redundancy measuring tools reviewed; the tools specifically used in this work appear in *italic* font.

proves both coverage of the original text and allows for a more precise calculation of the number of fingerprints that are in common. It is also robust against differences in lower/uppercase, insertion of spaces and newlines. For the analyses reported in this paper, we converted the whole document to a single string and extracted overlapping fingerprints of 30 characters with 10-character intervals. (In the original Cohen implementation non-overlapping 30-character fingerprints are extracted line per line.) Since we are interested in the temporal aspects of patient records, our script takes timestamps of documents into account which allows for chronological sorting and comparison between individual documents as well as that of a document to the concatenation of all older documents in the corpus.

4.3 Vizualization of redundancy

In section 6 we describe a prototype system for the visualization of patient records and how it can be used for annotation purposes. The code underlying the prototype is a Python wrapper script that takes temporally-ordered document-pair redundancy scores from the adapted Cohen script and uses this information to dynamically generate a graph (using the GraphViz software package) which depicts the flow of information in the patient's records over time.

5 Results

5.1 Incremental redundancy

The three subfigures in Figure 1 show the progression of redundancy over time in each of the three patient records in our corpus, measured with the Cohen script, and our own adapted Cohen script, respectively. Each data point shows the proportion of redundant text in a given document (ticks on the x-axis), compared to the concatenation of text from all older documents in the corpus. The documents on the x-axis are ordered chronologically. Since the original Cohen script does not allow for sequential comparison implementation, we performed manual data selection of the older documents to ensure that the two implementations were tested on the same corpus subsets.

While similarity measures should not be directly compared, they both show similar evolutions in the patient's records. We see that there is no clear incremental growth of redundancy over time such as has been reported for American EHRs, in any of the patient records. It should be noted that patient records 3 shows an increase in redundancy scores for the 20 most recent (i.e. right-hand) documents. Closer analysis shows that this is likely due to the type of documents, namely discharge notes (*Compte Rendu de Séjour*). The level of redundancy in different document types is discussed below.

We find that although both measures show very similar progressions, the adapted Cohen script allows for a more precise measuring than the original Cohen script (as evidenced by the higher average redundancy scores in all three subfigures). We will therefore be using this implementation for the other analyses reported in the rest of the paper.

5.2 Comparison between sections

While the previous analysis showed that there is no incremental growth of redundancy in the corpus, redundancy is still present: On average[3], 33% of the text in a document in the corpus is redundant. To estimate the impact on text mining and determine whether it will be beneficial or harmful, it is important to characterize which parts of the document are more likely to contain redundant information. To this end we created a second version of the corpus in which the header and footer information for all documents had been removed,

[3]Calculated with adapted Cohen script

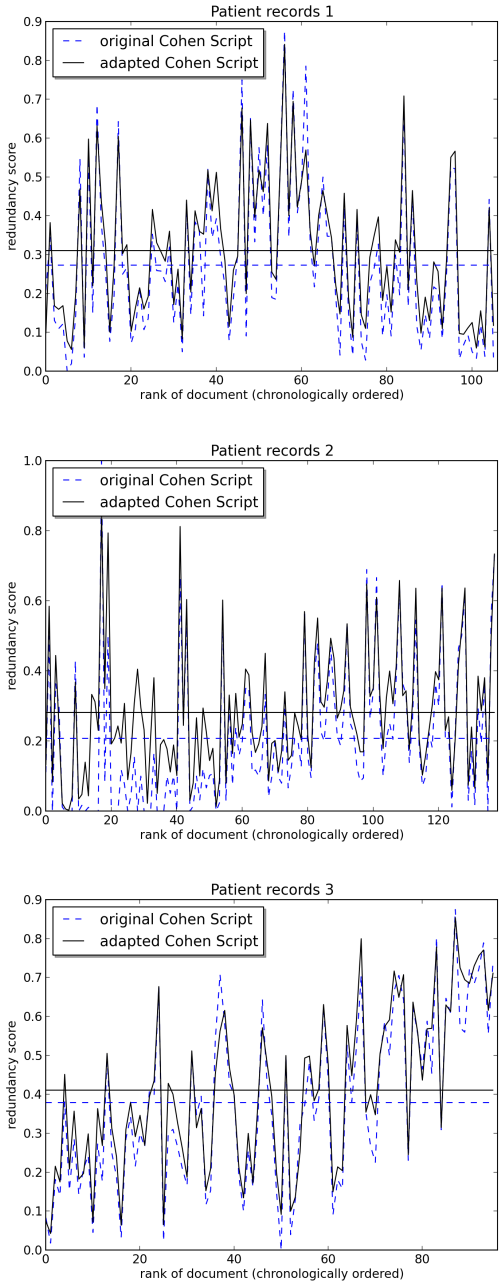

Figure 1: Redundancy over time of original text in three patient records, calculated by the Cohen and adaptedCohen script. The flat lines show the average over the whole patient records.

Category	Records 1	Records 2	Records 3[4]
CR d'Acte	10.3 (14.1)	16.3 (21.2)	15.9 (23.8)
CR de Séjour	25.7 (25.9)	N/A[5]	38.0 (30.5)
TA de Séjour	7.3 (9.9)	9 (15.9)	9.8 (12.3)

Table 2: Average redundancy for different document types. All numbers are percentages. Between brackets is the standard deviation. 'CR' stands for 'Compte Rendu' (*report*), 'TA' stands for 'Text Associé' (*associated text*).

leaving only 'topical content'.

Figure 2 shows the same progression of redundancy over time as Figure 1, but calculated over the version of the corpus without header or footer information. We see that the overall trends of the respective graphs remain similar for the different patient records but that the average redundancy level has decreased drastically. In patient records 1, 2 and 3, the average redundancy level decreases from 31% to 17.8%, 28% to 15.7% and 41% to 23.3%, respectively.

It is clear that most of the redundant text appears in the header and footer sections of the document. This text is not very informative by nature: Headers and footers contain contact information such as names, addresses, de-identified patient names, which will only add noise for text mining purposes that want to exploit the Natural Language in the EHRs. In the following analyses, we therefore only use the NoHeaders versions of the patient records, that is, only the free text that makes up the body of the notes in the patient records.

5.3 Comparison between document categories

Patient records contain a wealth of information in a variety of document types, such as test results and surgery notes (*Compte Rendu d'Acte*), discharge summaries (*Compte Rendu de Séjour*), and correspondence between doctors from various hospital departments (*Texte associé de Séjour*). As each document type describes a different aspect of the patient's stay in a hospital, they are likely to differ in writing style but also in their purpose in the hospital. Following Wrenn et al. (2010) we studied the differences in redundancies between different document types.

Table 2 shows the differences in average redundancy levels of documents from the three main document categories in the three patient records.

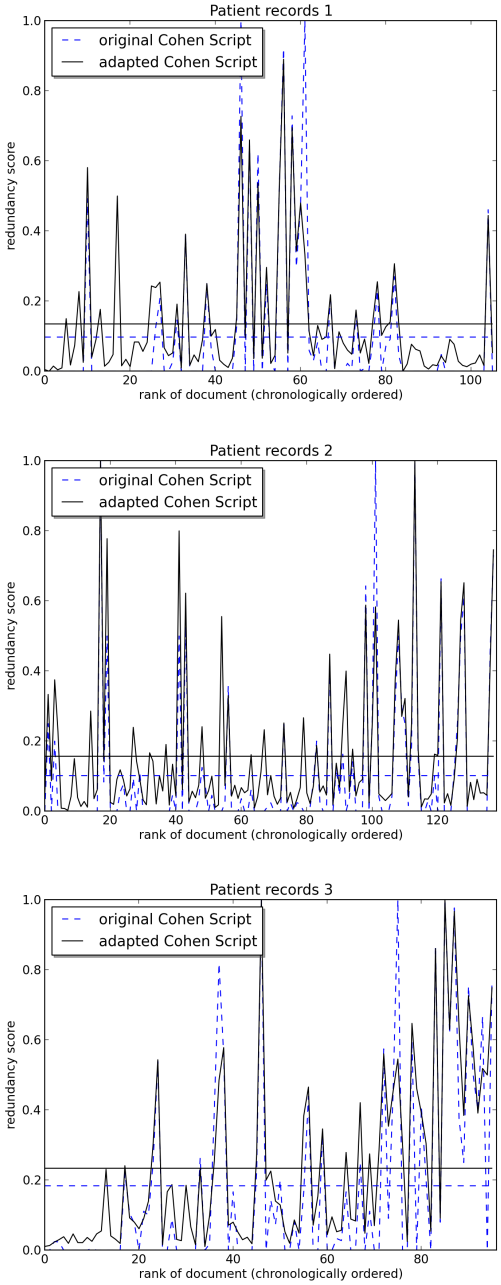

Figure 2: Redundancy over time of text in three patient records with header and footer information removed, calculated by the Cohen and adaptedCohen script. The flat lines show the average over the whole patient records.

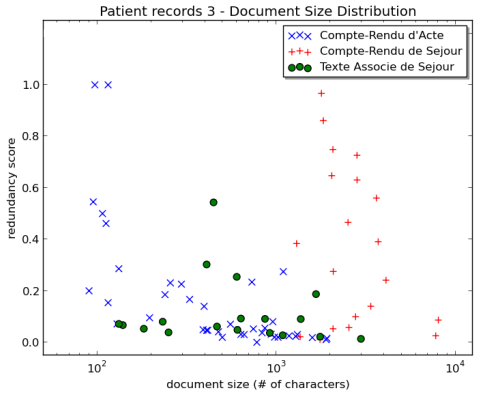

Figure 3: Document size (in number of characters) versus redundancy levels of documents from the main three categories of the patient records of patient 3. X-axis is set to logarithmic scale.

We can see that the average level of redundancy differs substantially between the different document types. Discharge summaries (*Compte Rendu de Séjour*) contain by far the most redundancy, while the associated correspondence between doctors (*Texte Associé de Séjour*) have a fairly low amount of redundancy. This can be explained by their structure and use: Discharge summaries generally have a fixed structure and aim to give a full overview of the patient's stay in the hospital, as well as a short overview of the patient's history. The associated correspondence, however, is optional, and is typically in the form of a letter that contains free text and no fixed structure.

To ensure that the measured redundancy levels are not an artefact of document length, we performed an additional analysis of document size versus redundancy ratio. Figure 3 shows the size distribution (in number of characters) of the documents from the three largest categories in the patient records of patient 3. Patient records 1 and 2 show similar distributions.

We find no direct correlation between document length and redundancy level, but rather a U-shaped distribution. The longest documents documents, i.e. discharge summaries (*Compte Rendu de Séjour*) have the highest average redundancy.

[4] As the metadata provided for these patient records contained errors, the documents have been manually reclassified into the different document types.

[5] The patient records did not contain enough documents of this type to calculate a reliable average

S \ T	Records 1	Records 2	Records 3
Records 1	*17.8 (13.5)*	14.4 (18.8)	12.4 (15.1)
Records 2	12.8 (17.9)	*15.7 (20.4)*	15.0 (18.0)
Records 3	9.5 (13.7)	12.1 (18.4)	*23.3 (24.9)*

Table 3: Redundancy scores between different patient records. All numbers are percentages. Between brackets is the standard deviation. S stands for *source* corpus. T stands for *target* corpus. The number in italics are the average redundancy levels that correspond to the black flat lines in Figure 2.

5.4 Inter-/Intra-patient record comparison

We saw that redundancy within patient record is fairly low (compared to the scores reported for the American EHRs), but given the fixed structure of certain document types (*Compte Rendu de Séjour* and *Compte Rendu d'Acte*) phrases or formulations may be shared between different records. Identifying these would be helpful for categorizing similar interventions and tests (as described over different documents) in different patient records.

Table 3 shows that the average redundancy between patient records is slightly lower than those within the EHRs. We can conclude that even though the patient illnesses and treatments (and the associated forms to record these) are fairly similar, the text in the patient records differs substantially between patients.

6 Discussion

Cross-culture differences in generating and managing medical information in text have been studied previously for breast cancer forums in Germany vs. United Kingdom (Weissenberger et al., 2004) and EHRs structure and narrative style for China vs. United States (Wu et al., 2013), and Sweden vs. Finland (Allvin et al., 2011). In this study we offer a preliminary comparison of the occurrence of redundancy in French EHRs, compared to numbers reported on redundancy in EHRs from the United States.

Although our corpus of the three patient records is too small to give conclusive results, it does offer some interesting insights: We find that the level of surface redundancy present in (and between) French medical records is fairly low. Moreover, we do not see clear indications of an incremental increase of redundancy over time, as has been reported for American EHRs. Most of the redundancy that is present in the French records in our study comes from document headers and footers. This text does not offer information on the course of the patient illness and should be discarded so as not to harm the performance of Text Mining or NLP applications that use the EHRs as training material. The absence of redundancy in the body of texts is beneficial for text mining purposes, and we can conclude that the mitigating strategies that have been developed for American EHRs are not probably not needed when processing French EHRs.

Subjective review of some of the French clinical notes confirms our findings and suggests that the copy-paste practices observed in American hospitals, which are meant to give health professionals access to comprehensive information about a patient within a single document, are not used in France. We find that rather than copy-pasting content from previous documents, references to previous documents are inserted in new documents (such as *cf. CR précédent*, see previous report, *examen biologique: voir feuilles ci-joint*, lab results: see attached).

However, as Table 2 indicates, this does differ between category types: In discharge summaries (*Compte-Rendus de Séjour*), which are generally the longest documents in the corpus, the measured redundancy levels are higher than in other types of documents, which indicates that entire text portions are copied from older documents. Discharge summaries are meant to be stand-alone documents integrating information about an entire patient stay which is otherwise described minutely in several other documents. We notice that these copied portions of text are often not strictly copy-pasted as they integrate small differences corresponding to re-writing of the text for clarity, addition of details previously unavailable and correction of erroneous information. So rather than copy-pasting content indiscriminately, French health professionals seem to do it strategically. In this way, redundancy should not be seen as a source of noise in a corpus but rather as an indication of information flow between documents.

These observations suggest that documents containing highly redundant sections are key documents in the patient records. While many of these documents are identified as *Compte-Rendu de Séjour* in the metadata, this is not always the

case. Therefore, it would be important to automatically identify these documents in a given EHR, so as to provide a new doctor with the most complete overview of a patient's history. Such information could later also serve for the purposes of automatic summarization.

As an exploration of our hypothesis and to gather more insight into the structure of patient records, we developed the prototype of a visualization tool that would allow us to track how information is transferred in a patient record over time. Figure 4 shows a screen capture of (part of) the graph generated for one of the patient records used in this study. Each block in the figure corresponds to one document in the patient's records. The documents are ordered chronologically along the Y-axis from earliest (top) to most recent (bottom). Documents that were created at the same moment, i.e. during the same hospital stay, are thus positioned next to one another. The shape of the blocks refers to the document type (*Compte-Rendu de Séjour* are square, *Compte-Rendu d'Acte* are circles, ...), and their size is relative to the document size. The number in the block refers to the document identifier in the patient's records. The interlinking lines refer to the proportion of redundant information in the more recent document that comes from the older document. A user-defined cut-off parameter allows for interactive exploration.

In this study we have focused exclusively on surface redundancy, i.e. the (almost) literal repetition of a piece of text from an older document. While string-based similarity measures are useful to detect blatant copy-pasting that throws off word distributions in language models, they prove too crude when we want to detect the flow of information, i.e. strategically used copy-pasting. More specifically, the current method cannot deal with highly similar text that is used to described two different events. For example, blood test results tend to be communicated using the same standard form. If two different blood tests yield the same results, this will result in two highly similar documents, the most recent of which would be judged as highly redundant by our current method. A more precise method is needed that incorporates semantic components such as identification of temporal expressions, or even event detection, into the string similarity method.

7 Future Work

The work presented in this paper is a preliminary study on a small-scale corpus and was meant to gain insight into copy-paste-induced redundancy in French EHRs. We find that rather than focusing on mitigating methods (as needed for American EHRs), we should look toward developing high-precision measures that capitalize on the existing redundancy in French EHRs. A first step for future work will be to replicate this study on a larger and more diverse corpus of patient records with different disease profiles, so as to confirm our findings and see to what extent text is shared between patient records from different hospital departments.

As a follow-up of this study we also plan to address two new main lines of research. First, we intend to develop a more precise surface redundancy measure which takes temporal expressions and terminological variation into account and which is more robust to small changes within large highly similar context. We will use the WiCoPaCo corpus (Max and Wisniewski, 2010) to train models that can automatically identify reformulations, and distinguish those from (error) corrections and updates.

Second, we will study redundancy on the level of the patient's records as a whole, not just on the document level. We intend to develop a measure that uses information on redundancy levels, the number of documents copied, the (temporal) distance of information that has been copied, ... to identify key documents within a patient's record. To this end we will need a reference set of correctly identified key documents in a set of patient records. This will be carried out by a group of health professionals, who will manually classify the documents in a sizable set of EHRs in terms of their importance (and information content). Since such annotations are very expensive, these professionals will be provided with an improved version of the visualization tool described in section 6 to select potential documents and speed up the annotation process.

8 Conclusion

In this paper, we present a preliminary study on the presence of copy-paste-induced redundancy in French EHRs. We find that the high levels of redundancy and incremental increase of redundant text over time which have been observed in American EHRs, does not feature in our subset. As a re-

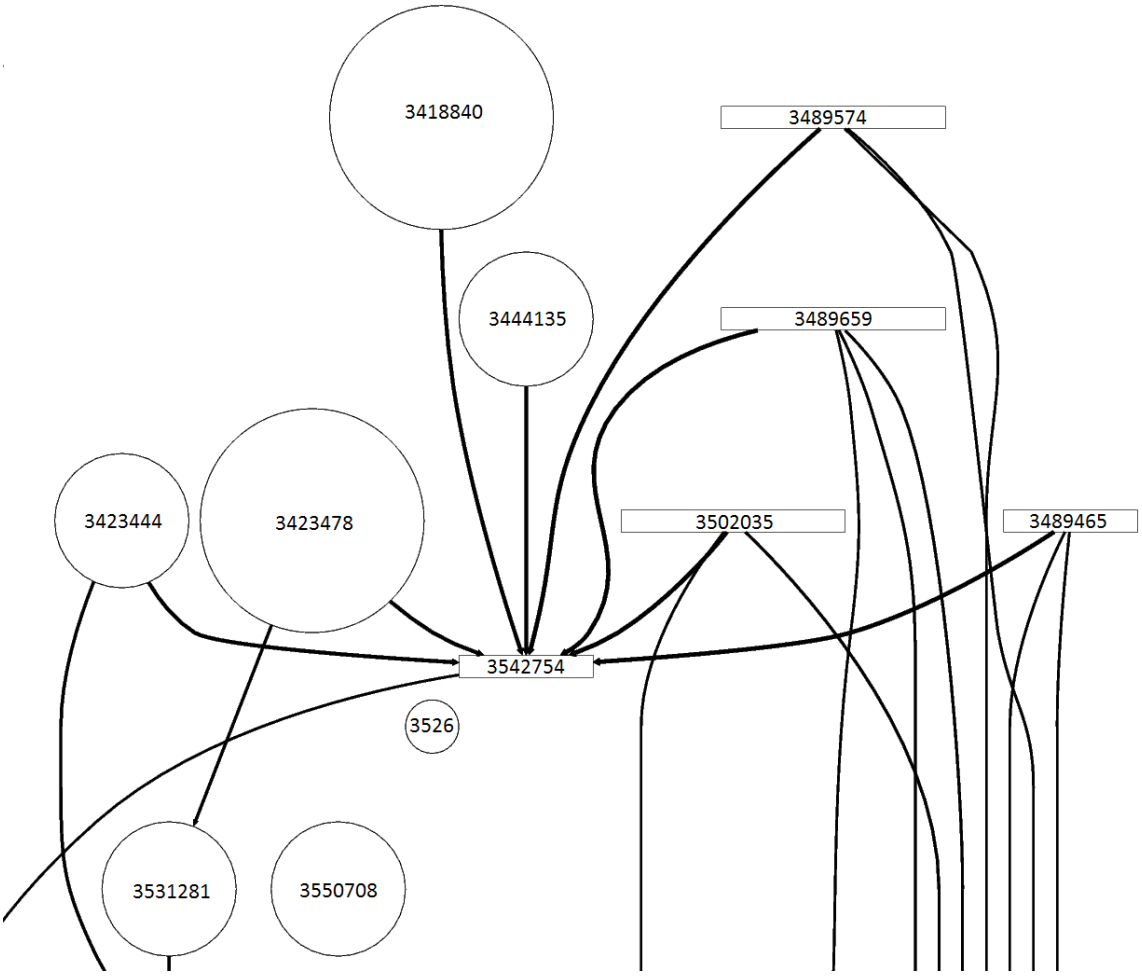

Figure 4: Screen shot of zoom-in from the visualization tool. Documents (represented by blocks) are ordered chronologically (Y-axis) with the oldest at the top, and the most recent at the bottom. The number within a block refer to the document identification number within the respective patient's records.

sult, there is no expected impact from redundancy on language models or other natural language processing methods applied to French EHRs. Rather, the limited redundancy that is present in the corpus may be strategically exploited to yield important information from the records.

Acknowledgments

This work was supported by the French National Agency for Research under grants CABeRneT[6] ANR-13-JS02-0009-01 and Accordys[7] ANR-12-CORD-0007.

The authors thank the Biomedical Informatics Department at the Rouen University Hospital for providing access to the LERUDI corpus for this work.

[6]CABeRneT: Compréhension Automatique de Textes Biomédicaux pour la Recherche Translationnelle

[7]Agrégation de Contenus et de COnnaissances pour Raisonner à partir de cas dans la DYSmorphologie foetale

References

H Allvin, E Carlsson, H Dalianis, R Danielsson-Ojala, V Daudaravičius, M Hassel, D Kokkinakis, H Lundgrén-Laine, GH Nilsson, O Nytrø, S Salanterä, M Skeppstedt, H Suominen, and S Velupillai. 2011. Characteristics of finnish and swedish intensive care nursing narratives: a comparative analysis to support the development of clinical language technologies. *J Biomed Semantics*, Suppl 3:S1.

Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. 1990. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410.

Xin Chen, Brent Francia, Ming Li, Brian Mckinnon, and Amit Seker. 2004. Shared information and program plagiarism detection. *Information Theory, IEEE Transactions on*, 50(7):1545–1551.

R Cohen, M Elhadad, and N Elhadad. 2013. Redundancy in electronic health record corpora: analysis, impact on text mining performance and mitigation strategies. *BMC Bioinformatics*, 14:10.

L Deléger, C Grouin, and A Névéol. 2014. Automatic content extraction for designing a french clinical corpus. In *Proc AMIA Annu Symp*.

Dina Demner-Fushman, Wendy W Chapman, and Clement J McDonald. 2009. What can natural language processing do for clinical decision support? *J Biomed Inform*, 42:760–772.

Cyril Grouin and Aurélie Névéol. 2014. De-identification of clinical notes in french: towards a protocol for reference corpus developpement. In *J Biomed Inform*, Aug.

Aurélien Max and Guillaume Wisniewski. 2010. Mining naturally-occurring corrections and paraphrases from wikipedia's revision history. In *LREC*.

Maxim Mozgovoy, Kimmo Fredriksson, Daniel White, Mike Joy, and Erkki Sutinen. 2005. Fast plagiarism detection system. In *String Processing and Information Retrieval*, pages 267–270. Springer.

EL Siegler and R Adelman. 2009. Copy and paste: a remediable hazard of electronic health records. *Am J Med*, 122:495–496.

Paul MB Vitányi, Frank J Balbach, Rudi L Cilibrasi, and Ming Li. 2009. Normalized information distance. In *Information theory and statistical learning*, pages 45–82. Springer.

JM Weis and PC Levy. 2014. Copy, paste, and cloned notes in electronic health records: prevalence, benefits, risks, and best practice recommendations. *Chest*, 145(3):632–8, Mar 1.

C Weissenberger, S Jonassen, J Beranek-Chiu, M Neumann, D Müller, S Bartelt, S Schulz, JS Mönting, K Henne, G Gitsch, and G Witucki. 2004. Breast cancer: patient information needs reflected in english and german web sites. *Br J Cancer*, 91(8):1482–7, Oct 18.

JO Wrenn, DM Stein, S Bakken, and PD Stetson. 2010. Quantifying clinical narrative redundancy in an electronic health record. *Journal of the American Medical Informatics Association*, 17(1):49–53.

Y Wu, J Lei, WQ Wei, B Tang, JC Denny, ST Rosenbloom, RA Miller, DA Giuse, K Zheng, and H Xu. 2013. Analyzing differences between chinese and english clinical text: a cross-institution comparison of discharge summaries in two languages. In *Stud Health Technol Inform*, volume 192, pages 662–6.

R Zhang, S Pakhomov, BT McInnes, and GB Melton. 2011. Evaluating measures of redundancy in clinical texts. In *Proc. AMIA Annual Symposium*, page 1612–1620.

R Zhang, S Pakhomov, and GB Melton. 2014. Longitudinal analysis of new information types in clinical notes. In *Proc. AMIA Summits on Translational Science*, page 232–237.