Detection of Text Reuse in French Medical Corpora

Eva D'hondt
LIMSI, CNRS
Université Paris-Saclay
F-91405 Orsay
dhondt@limsi.fr

Cyril Grouin
LIMSI, CNRS
Université Paris-Saclay
F-91405 Orsay
grouin@limsi.fr

Aurélie Névéol
LIMSI, CNRS
Université Paris-Saclay
F-91405 Orsay
neveol@limsi.fr

Efstathios Stamatatos
Dept. of Information and
Communication Systems Engineering,
University of the Aegean,
Samos 83200
stamatatos@aegean.gr

Pierre Zweigenbaum
LIMSI, CNRS
Université Paris-Saclay
F-91405 Orsay
pz@limsi.fr

Abstract

Electronic Health Records (EHRs) are increasingly available in modern health care institutions either through the direct creation of electronic documents in hospitals' health information systems, or through the digitization of historical paper records. Each EHR creation method yields the need for sophisticated text reuse detection tools in order to prepare the EHR collections for efficient secondary use relying on Natural Language Processing methods. Herein, we address the detection of two types of text reuse in French EHRs: 1) the detection of updated versions of the same document and 2) the detection of document duplicates that still bear surface differences due to OCR or de-identification processing. We present a robust text reuse detection method to automatically identify text reuse in document pairs in two French EHR corpora that achieves an overall macro F-measure of 0.68 and 0.60, respectively and correctly identifies all redundant document pairs of interest.

1 Introduction

Over the last decade a large number of hospitals and medical institutions have adopted the use of Electronic Health Records (EHRs) to store patient records and medical details. Simultaneously, the lowered cost of computational resources has given rise to digitization efforts of existing (paper) collections. While the presence of such large, digital corpora opens up exciting possibilities for medical data and text mining or modelization efforts, this is not without certain caveats. The resulting digital collections often are noisy, with several issues that can have an impact on the accuracy of subsequent text mining processes, such as encoding errors, missing files, OCR errors, etc. One interesting issue in cumulatively constructed text corpora is the problem of 'text reuse'. Text reuse is defined here as the intentional or unintentional reusing of existing text (fragments) to create a new text, for example, by copy-pasting text fragments from one document to fit into a new document; or by adapting a report and saving both the old and the new version as separate documents. Text reuse is a complex phenomenon which has been studied in multiple settings such as newspaper journalism (Clough et al., 2002), programming code (Ohno and Murao, 2009), the analysis of text reuse in blogs and web pages (Abdel Hamid et al., 2009), etc. It is quite prevalent in the medical domain (Wrenn et al., 2010) and often seen as a negative factor: Cohen et al. (2013) found that copy-pasting practices in US hospitals have a significant negative impact on the accuracy of the subsequent text mining systems on the clinical notes. However, when text reuse is considered as a diachronic phenomenon, it has some interesting aspects. By identifying which text (fragments) have been reused we can follow the flow of information over time in a patient's file. Moreover, adjustments that are made to copied text (fragments) can give an insight into the thought process of the acting clinicians and may help identify potential errors or adjustments during the treatment process (Hirschtick, 2006).

Text reuse has been studied extensively in the context of authorship attribution and plagiarism detection (Stamatatos, 2009). In general we can distinguish between two main types of text reuse: 'global text reuse' in which the task is to pair up (near-) duplicate documents that exists in different locations, or whose differences are linked to version control issues; and 'local text reuse' which occurs when people borrow or plagiarize smaller text fragments such as sentences or passages from various sources to incorporate in a new text. Both types are included in this study.

While the goal is the same, there are some key differences between plagiarism detection and text reuse detection in the medical domain. Medical professionals work under an enormous time pressure, so rather than rewriting an existing text (fragment), they will merely add new information or adjust existing information, and at best edit out some orthographic errors or write out acronyms that existed in the previous version. Consequently, our methods can focus 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). Another difference is the quality of the written text. Depending on the quality and the nature of the text formatting tools that are available, electronic health records may contain an astounding number of orthographic errors (Ruch et al., 2003) or in the case of a digitized corpus, a large variety of OCR errors. Another source of potential minor surface variation is the de-identification process in which personal health identifiers (PHI) such as patients names, phone numbers, record numbers are replaced by plausible synthetic surrogates (Sweeney, 1996; Meystre et al., 2010). Depending on how the process is implemented, for example with the inclusion of random substitution, numbers that were the same in the original documents can appear with slight variations in the de-identified documents. Text reuse metric tools in this domain therefore need to be robust to the noise of these sources of surface variation and correctly detect similar text segments even when the surface forms do not match 100%.

The current paper presents a simple but effective tool for text reuse detection in the medical domain, both for global and local text reuse detection, which proves robust to surface variation prevalent in medical texts by allowing for character gaps when calculating the blocks of reused texts. The tool is meant to figure as a module in a larger framework, i.e. a pipeline which normalizes and extracts information from documents in a patient file in order to model the patient's treatment over time. This adds a practical component to the evaluation of the proposed tool. Missing a case of text reuse is a more grievous error than (mis)labeling a false positive. A mislabeled case, i.e either not correctly determining between different degrees of reuse, or erroneously spotting text reuse, can be spotted by the information extraction module later on in the pipeline. When a case of text reuse is not identified, however, no subsequent processing will occur for that document pair and the information is effectively lost for the information extraction process. In this paper we present and evaluate the text reuse detection tool in isolation and discuss its strengths and weaknesses.

2 Background

A traditional approach for the detection of verbatim copying[1] is to compute the similarity between the source and target text as the proportion of substring sequences that the two texts have in common. These substring sequences can either be defined as character n-grams (Cohen et al., 2013), words (Wrenn et al., 2010), or word n-grams (Adeel Nawab et al., 2012). These methods are mainly based on fingerprinting and hashing techniques, i.e. the documents are represented as sets of unique digital signatures, and are highly precise but are not robust to much surface variation. Some methods, however, are adapted to deal with insertions and deletion of words or characters. For example, as an extension of the 'longest common substring' algorithm (Gusfield, 1997), which calculated text similarity as the length of the longest continuous sequence of characters normalized by the sum of the document lengths, Wise et al. (1996) developed the 'Greedy String Tiling' which allows for insertions and deletions. It determines the maximum set of contiguous substrings that two documents have in common, wherein each substring has the largest possible length. However by eliminating word order through the construction of the

[1] As opposed to semantic reuse where the same idea of message is rewritten in a different manner.

set, valuable information on the ordering of the subsequences is lost. The method proposed in this paper aims to address this problem by allowing for a 'mismatch gap' (see section 3) while still keeping information on the original subsequence order when calculating the similarity score. Another form of surface variation that needs to be caught—especially in OCRed corpora—is due to differences in formatting: Lopresti (2000) developed a string matching algorithm that distinguishes between differences in content and differences in formatting within a document pair.

The study of text reuse detection in the medical domain has either focused on plagiarism detection in medical articles in PubMed (Errami et al., 2010; Sun et al., 2010) or for dedicated journals (Baždarić et al., 2012) or on text reuse during the creation of medical corpora and its consequences for database integrity or subsequent text mining applications (Wrenn et al., 2010). Zhang et al. (2011) found that redundant information contained in US clinical notes increases over time and that a text reuse detection tool with domain-specific knowledge is a necessary step in the detection of novel information within clinical files (Zhang et al., 2012).

3 Text reuse detection tool

The text reuse detection tool presented in this article consists of three main modules and is inspired by the best practices from recent research in plagiarism detection (Potthast et al., 2014). In a first step, the text is split into character n-grams of a user-defined length. Each substring unit is indexed with information on its position in the source document (character offsets). The document is thus transformed into a bag of overlapping character n-grams. We then apply a global alignment algorithm[2] to find the alignment of sequences with the largest global overlap between the two documents. In a second step, we then resolve gaps in the alignment, i.e. disjoint blocks, and construct larger blocks of aligned text. Where a large number of consecutive in-common substring sequences are detected that are interspersed by spurious non-matching blocks, the substring sequences are merged into a larger (quasi-)matching block by a user-defined 'gap parameter'. This parameter was heuristically set to 3 characters for the experiments described in this paper. For the OCRed corpus we also experimented with a variant in which larger character gaps were allowed if the non-matching blocks of the two documents contained 'confusion pairs'[3] of common OCR errors that were extracted from a training corpus. This 'gap parameter' catches small differences in formatting or character variations, i.e. a misspellings or OCR errors, between the two documents.

Finally, in the third step, the tool outputs the constructed larger 'matching blocks' with offset information for local text reuse detection, and calculates the proportion of matching text over the length of its source document to give an estimation of the global text reuse between the two documents. At this point the tool does not yet filter out text blocks that are below a certain length threshold (thus eliminating spurious matches). The use of short character n-grams (n=3) ensures that the similarity score will not be largely affected by small differences in detected fragments caused by OCR errors or differences in formatting. Figure 1 illustrates step 2 and 3 of the process.

4 Corpora

We show the performance of the tool on two separate and distinct corpora of French clinical notes, which exemplify the problem of local and global similarity, respectively.

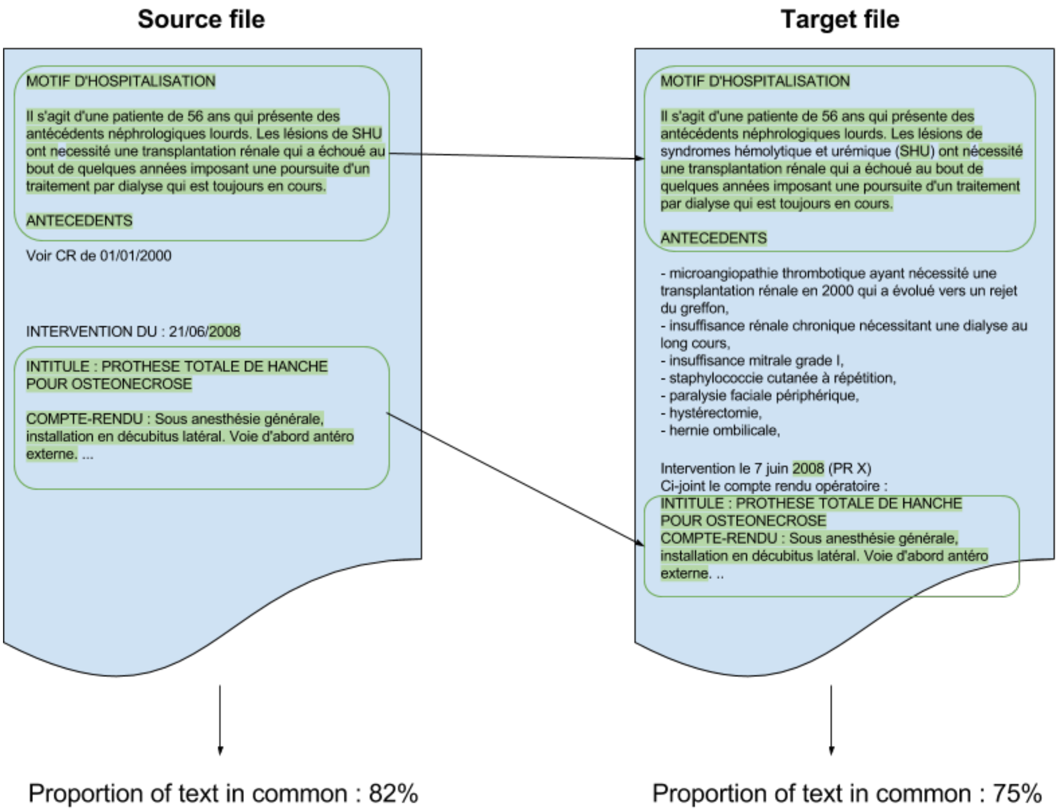
4.1 Corpus with local text reuse (LTR)

The first corpus consists of 107 documents that describe a patient's illness, renal transplantation and follow-up case over time through various lab results, consultation reports, etc. The corpus is originally an EHR corpus, that is, the original text was edited in Word documents which were later on automatically transformed into text files using the `AntiWord`[4] tool which converts MS Word documents into plain

[2] Implemented in the Python difflib library

[3] Confusion pairs are systematic OCR errors in which a character or a sequence of characters in the source document is consistently replaced by another character or sequence of characters during the OCR process. For example, characters like 'i' and 'l' are visually similar and thus often confused.

[4] http://www.winfield.demon.nl/


Source file
Target file
Proportion of text in common : 82%
Proportion of text in common : 75%

MOTIF D'HOSPITALISATION

Il s'agit d'une patiente de 56 ans qui présente des antécédents néphrologiques lourds. Les lésions de SHU ont necessité une transplantation rénale qui a échoué au bout de quelques années imposant une poursuite d'un traitement par dialyse qui est toujours en cours.

ANTECEDENTS

Voir CR de 01/01/2000

INTERVENTION DU : 21/06/2008

INTITULE : PROTHESE TOTALE DE HANCHE POUR OSTEONECROSE

COMPTE-RENDU : Sous anesthésie générale, installation en décubitus latéral. Voie d'abord antéro externe. ...

MOTIF D'HOSPITALISATION

Il s'agit d'une patiente de 56 ans qui présente des antécédents néphrologiques lourds. Les lésions de syndromes hémolytique et urémique (SHU) ont nécessité une transplantation rénale qui a échoué au bout de quelques années imposant une poursuite d'un traitement par dialyse qui est toujours en cours.

ANTECEDENTS

- microangiopathie thrombotique ayant nécessité une transplantation rénale en 2000 qui a évolué vers un rejet du greffon,
- insuffisance rénale chronique nécessitant une dialyse au long cours,
- insuffisance mitrale grade I,
- staphylococcie cutanée à répétition,
- paralysie faciale périphérique,
- hystérectomie,
- hernie ombilicale,

Intervention le 7 juin 2008 (PR X)
Ci-joint le compte rendu opératoire :
INTITULE : PROTHESE TOTALE DE HANCHE POUR OSTEONECROSE
COMPTE-RENDU : Sous anesthésie générale, installation en décubitus latéral. Voie d'abord antéro externe. ..

Figure 1: Report samples of the local text reuse corpus with fake data and realistic text reuse examples. The source file constitutes an earlier (i.e. older) version of the the target file on the right within the same patient records. Green highlighting indicates initial matching characters and the green blocks show the constructed 'matching blocks' with variation gaps included. For reasons of legibility we do not show the overlapping character sequences that were created in step 1. Please note that inconsistencies with regards to dates are caused by the de-identification method. The proportion of text in common that is used to calculate the similarity between the document pair is based on the entire documents, here we only show an excerpt.

text. This corpus is a subset of larger corpus which was used in a previous study on text reuse (D'hondt et al., 2015). While it does not contain a large amount of text reuse, the corpus exhibits an important temporal dimension, i.e. medical cases that span multiple years. For this reason, the corpus contains a series of documents reporting on regular check-ups that each build on the previous one, by retelling the medical history of the patient and completing it with the most recent exam results. Another interesting series of documents are follow-up exams that are conducted several times as part of the patient care pathway, and may yield similar results on each instance.

4.2 Corpus with global text reuse (GTR)

The second corpus consists of 1,007 documents from French foetopathology[5] reports, with data from 25 different patients. This corpus was assembled and digitized within the context of the Accordys project. The digitization effort consisted of OCRing the original typed-out pages, which was then followed by a de-identification step. There is a substantial amount of redundancy in this corpus: For some documents, several (nearly-identical[6]) copies of the same original document were added to the patient's folder. However, the de-identification process has deleted parts of the text for some copies, but not for others. While

[5] The medical domain which specializes in the treatment and diagnosis of illnesses in unborn children.

[6] While the original paper documents might be identical, the process of OCR and de-identification has introduced enough noise that very few identical files remain.

the patient files in the corpus do not span a long time individually, there are multiple cases in which different (intermediate) versions of a document were contained in the file. It is therefore a considerable challenge to distinguish between near-duplicate files that originate from the same original document, or those that came from two different versions of that document.

5 Corpus analysis

For both corpora we generated all possible document pairs for each patient. These document pairs were then labeled by two independent annotators[7] with regards to the similarity between the two documents. The annotators took care to distinguish between (near-)duplicate text[8] (category '2') and documents that are either different versions of the same report, e.g. an intermediate version versus the final version with more information, or similar reports on two different events (category '1'). Table 1 shows the labeling scheme, and the cut-off scores that were used to classify the output of the duplication detection tool. The number of document pairs for each of the three categories can be found in Tables 2 and 3 for the two different categories.

Category label	Category description	Score cut-off
2	near-duplicates	x >= 0.9
1	different version of same base document or different events	0.5 >= x < 0.9
0	documents are unrelated	x < 0.5

Table 1: Explication of labels used in study.

5.1 Corpus with local text reuse (LTR)

Category label	# document pairs in reference set	Precision	Recall	F1-score
2	2	0.20	1.00	0.33
1	6	0.60	0.86	0.71
0	99	1.00	1.00	1.00
macro-average	-	0.60	0.95	0.68

Table 2: Precision and Recall scores for EHR corpus (LTR)

The local text reuse corpus only has a small number of positive examples of text reuse but the tool still categorizes the majority of the document pairs correctly. The low Precision score for category '2' is caused by the distinctive structure in the yearly follow-up reports that are included in the corpus. While the documents contain different information, i.e. one follow-up report describes the state of patient one year after the transplant, a second document describes his/her state after 5 years, they follow a similar structure and formatting and contain little free text. To correctly identify that such documents pertain to different medical events, additional information such as identifying the documents time stamps is needed. Copy-pasting of results from smaller, non-structured report into the medical overview was successfully detected however. From a medical perspective of effectively reviewing the patient record, all document pairs with some form of reuse have been successfully identified, so that the bulk of the manual review work can be lightened using this tool.

5.2 Corpus with global text reuse (GTR)

The second corpus contains more examples of near-duplicates. Interestingly, we find that our tool has severe problems with the detection of intermediary versions of reports, and often categorizes them as category 2 (identical pairs). A deeper analysis of the errors shows that the current method does not take

[7] We did not calculate IAA but few conflicting annotations occurred. Conflicts in annotations were resolved after discussion.

[8] In the case of local text reuse this can signify parts of the document, in case of global text reuse it refers to the entire document.

Category label	# document pairs in reference set	Precision	Recall	F1-score
2	218	0.96	0.66	0.78
1	55	0.04	0.05	0.05
0	1451	0.95	0.99	0.97
macro-average	-	0.65	0.57	0.60

Table 3: Precision and Recall scores for OCRed corpus (GTR)

document length into account. Some reports are highly similar in all sections but for the 'conclusion section'. In other document pairs, the intermediary versions are missing only one section which is present in the final version. One way of dealing with such differences between text versions would be to add a boosting factor for longer text insertion, i.e. a long block of inserted text should have a stronger (negative) impact on the similarity score than the same number of inserted characters spread out over various, shorter blocks of inserted text. This approach would certainly improve classification accuracy between the '1' and '2' categories, but the booster factor would be hard to determine with regards to the document length. A more accurate approach would be to equip the tool with additional data, either on document structure, e.g. perform the comparison on section level rather than document level, or on the time stamps of the generated documents. While many documents of category '1' are mislabeled as near-duplicates, analysis of the correct pairings in category 2 shows that the tool exhibits a high precision in extracting real near-duplicates, even in the face of a high OCR error rate (D'hondt et al., 2016).

6 Conclusion

In this paper we present a character-based tool for the detection of text reuse, and evaluate its usability on two different French EHR corpora. We find that our tool is robust to the surface variation in the two corpora which were introduced by OCR and orthographic errors as well as variations introduced by the de-identification process. As such we believe it is well-suited to be included in a NLP pipeline that will process a large variety of medical corpora. While the tool generally achieves a high recall score which is important for the subsequent pipeline, it lacks in precision. The tool is not able to distinguish more 'semantic' differences such as the differences between intermediate and final versions, or when reports describe highly similar yet different events. To capture such information the tool needs to be coupled with additional information in the NLP pipeline such as information on the time stamp of the document, or information on document structure, i.e. so that the tool will only be run on parts of the document that contain free text. One limitation of this study is the size of the EHR corpus used for testing the method. While the preliminary results obtained here are encouraging, they would need to be confirmed on a larger data set. We plan to address this in future work in collaboration with physicians who will also provide qualitative feedback on the usability of the tool in a clinical setting.

Acknowledgements

This work was supported by the French National Agency for Research under the Accordys[9] ANR-12-CORD-0007, and CABeRneT[10] ANR-13-JS02-0009-01 grants.

References

Ossama Abdel Hamid, Behshad Behzadi, Stefan Christoph, and Monika Henzinger. 2009. Detecting the origin of text segments efficiently. In *Proc of the 18th international conference on World wide web*, pages 61–70. ACM.

Rao Muhammad Adeel Nawab, Mark Stevenson, and Paul Clough. 2012. Detecting text reuse with modified and weighted n-grams. In *Proc of the First Joint Conference on Lexical and Computational Semantic (*SEM)*, pages 54–58, Montréal, QC. Association for Computational Linguistics.

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

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

Ksenija Baždarić, Lidija Bilić-Zulle, Gordana Brumini, and Mladen Petrovečki. 2012. Prevalence of plagiarism in recent submissions to the Croatian Medical Journal. *Science and engineering ethics*, 18(2):223–239.

Paul Clough, Robert Gaizauskas, Scott SL Piao, and Yorick Wilks. 2002. Meter: Measuring text reuse. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 152–159. Association for Computational Linguistics.

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

Eva D'hondt, Xavier Tannier, and Aurélie Névéol. 2015. Redundancy in French electronic health records: A preliminary study. In *Proc of the 6th Health Text Mining and Information Analysis Work (LOUHI)*, pages 21–30, Lisbon, Portugal.

Eva D'hondt, Cyril Grouin, and Brigitte Grau. 2016. Redundancy in French electronic health records: A preliminary study. In *Proc of the 7th Health Text Mining and Information Analysis Work (LOUHI)*, pages 61–68, Austin, TX.

Mounir Errami, Zhaohui Sun, Angela C George, Tara C Long, Michael A Skinner, Jonathan D Wren, and Harold R Garner. 2010. Identifying duplicate content using statistically improbable phrases. *Bioinformatics*, 26(11):1453–1457.

Dan Gusfield. 1997. *Algorithms on strings, trees and sequences: computer science and computational biology*. Cambridge university press.

Robert E Hirschtick. 2006. Copy-and-paste. *Jama*, 295(20):2335–2336.

Daniel P Lopresti. 2000. String techniques for detecting duplicates in document databases. *International Journal on Document Analysis and Recognition*, 2(4):186–199.

Stephane M Meystre, F Jeffrey Friedlin, Brett R South, Shuying Shen, and Matthew H Samore. 2010. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC Med Res Methodol*, 10(70).

Asako Ohno and Hajime Murao. 2009. A new similarity measure for in-class source code plagiarism detection. *International Journal of Innovative Computing, Information and Control*, 5(11):4237–4247.

Martin Potthast, Matthias Hagen, Anna Beyer, Matthias Busse, Martin Tippmann, Paolo Rosso, and Benno Stein. 2014. Overview of the 6th international competition on plagiarism detection. In *Working Notes for CLEF 2014 Conference*, pages 845–876, Sheffield, UK.

Patrick Ruch, Robert Baud, and Antoine Geissbühler. 2003. Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record. *Artificial intelligence in medicine*, 29(1):169–184.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556.

Zhaohui Sun, Mounir Errami, Tara Long, Chris Renard, Nishant Choradia, and Harold Garner. 2010. Systematic characterizations of text similarity in full text biomedical publications. *PLoS One*, 5(9):e12704.

Latanya Sweeney. 1996. Replacing personally-identifying information in medical records, the Scrub system. In *Proceedings of the AMIA annual fall symposium*, pages 333–337. American Medical Informatics Association.

Michael J Wise. 1996. YAP3: Improved detection of similarities in computer program and other texts. *ACM SIGCSE Bulletin*, 28(1):130–134.

Jesse O Wrenn, Daniel M Stein, Suzanne Bakken, and Peter D Stetson. 2010. Quantifying clinical narrative redundancy in an electronic health record. *Journal of the American Medical Informatics Association*, 17(1):4953.

Rui Zhang, Serguei Pakhomov, Bridget T McInnes, and Genevieve B Melton. 2011. Evaluating measures of redundancy in clinical texts. In *AMIA Annual Symposium Proceedings*, volume 2011, page 1612. American Medical Informatics Association.

Rui Zhang, Serguei Pakhomov, and Genevieve B. Melton. 2012. Automated identification of relevant new information in clinical narrative. In *Proc of the 2nd SIGHIT International Health Informatics Symposium (IHI)*, pages 837–842, Miami, Florida, USA. ACM.