Error Propagation in EHRs via Copy/Paste: An Analysis of Relative Dates

Kirk Roberts, PhD, Amos Cahan, MD, Dina Demner-Fushman, MD, PhD
U.S. National Library of Medicine, Bethesda, MD

Abstract

We present a method for identifying errors in EHRs caused by copying and pasting text containing relative dates. Our method utilizes a sequence alignment method for recognizing instances of copy/paste, and regular expressions for relative dates. We furthermore present an analysis of our method on the MIMIC-II dataset.

Introduction

Electronic health records (EHR) are a powerful enabling technology, but concerns exist about the dangers of copying text from one clinical record to another.[1,2] While strategies exist for mitigating copy/paste in text mining approaches, little work has been done to evaluate the types of errors introduced to the EHR through copy/paste. This work studies errors due to the copying of relative dates. When a note contains a relative date (e.g., "*3 days ago*", "*past 4 weeks*", "*43 y/o*"), performing temporal reasoning requires this date be grounded, typically to the date the report was written. However, if the text is copied to a later report, the ability to ground this date is lost. We propose a method called DupLink, which links pasted text back to its original source (Figure 1). We then utilize regular expressions to identify relative dates in pasted text, enabling such cases to be flagged or even automatically corrected.

Methods

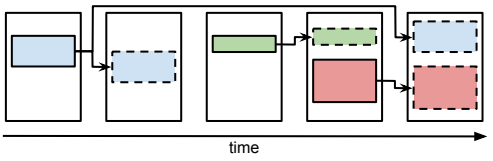

Figure 1: DupLink

DupLink recognizes copy/paste using Smith-Waterman local sequence alignment, a dynamic programming algorithm for detecting similar regions from two sequences. For each patient in MIMIC-II[3], reports are chronologically ordered, and Smith-Waterman is run pairwise to identify duplicate spans. A text span can act as the source (copy) to any number of targets (paste); a target can only have one source, creating the linking structure in Figure 1. DupLink iteratively finds the best local alignment on each report pair until no more high-scoring alignments are found. For each instance of copy/paste, regular expressions identify six classes of relative dates, best illustrated by example: (P1) *fourteen days ago*, (P2) *last year*, (P3) *on Saturday*, (P4) *yesterday*, (P5) *this July*, (P6) *45 years old*. To evaluate the severity of copying errors, we used five classes: (C1) *updated* in the pasted text, (C2) *valid*, as little time has passed, (C3) *resolvable* by a nearby absolute date, (C4) inconsistent, but *minor*, and (C5) inconsistent, and *serious*. Some analysis can be performed automatically (e.g., when reports are from the same day), otherwise human analysis is necessary. An experienced internist (AC) annotated a sample of each of the six pattern classes.

	matches	annotations	%C1	%C2	%C3	%C4	%C5
P1	156	142 (100)	-	53	26	20	2
P2	50	50 (7)	-	86	2	8	4
P3	47	47 (18)	11	85	4	-	-
P4	1278	731 (100)	6	62	17	12	1
P5	790	432 (100)	8	71	9	9	-
P6	388	199 (100)	-	100	-	-	-

Table 1: Results. Annotations in parenthesis are the manual annotations. Percentages are projections from manual and automatic annotations.

Results

See Table 1. Given the size of MIMIC, there were few instances of copy/paste. In other institutional cultures, however, copy/paste is more prevalent. Most copy/paste instances in MIMIC occur on the same day (C2). The resolvable class is the next most common, suggesting there are many relative dates that should be resolved to a local anchor instead of the report date. While few serious inconsistencies were found, they do exist along with many minor inconsistencies. This suggests the importance of integrating a copy/paste detection system like DupLink into the EHR. Further information as well as a software implementation of DupLink can be obtained by contacting the authors.

Acknowledgement This work was supported by the intramural research program at NLM/NIH.

References

1. Robert E. Hirschtick. Copy-and-Paste. *Journal of the American Medical Association*, 295(20):2335–2336.
2. Jesse O. Wrenn, Daniel M. Stein, Suzanne Bakken, and Peter D. Stetson. Quantifying clinical narrative redundancy in an electronic health record. *Journal of the American Medical Informatics Association*, 17:49–53, 2010.
3. DJ Scott, J Lee, I Silva, S Park, GB Moody, LA Celi, and RG Mark. Accessing the public MIMIC-II intensive care relational database for clinical research. *BMC Med Inform Decis Mak*, 13(9), 2013.