Journal of Biomedical Informatics 42 (2009) 53–58

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Improving language models for radiology speech recognition

John M. Paulett [a], Curtis P. Langlotz [b,*]

[a] School of Engineering and Applied Science, University of Pennsylvania, USA
[b] Department of Radiology, University of Pennsylvania, Radiology Administration, Penn Tower Lobby Level, 399 South 34th Street, Suite 100, Philadelphia, PA 19104, USA

ARTICLE INFO

Article history:
Received 12 December 2007
Available online 12 August 2008

Keywords:
Radiology
Speech recognition
n-Gram
Trigram model
Word frequency
Radiology reports

ABSTRACT

Speech recognition systems have become increasingly popular as a means to produce radiology reports, for reasons both of efficiency and of cost. However, the suboptimal recognition accuracy of these systems can affect the productivity of the radiologists creating the text reports. We analyzed a database of over two million de-identified radiology reports to determine the strongest determinants of word frequency. Our results showed that body site and imaging modality had a similar influence on the frequency of words and of three-word phrases as did the identity of the speaker. These findings suggest that the accuracy of speech recognition systems could be significantly enhanced by further tailoring their language models to body site and imaging modality, which are readily available at the time of report creation.

1. Background

As hospitals have sought to tighten their budgets, and referring providers demand rapid turn around time for radiology reports, radiology departments have shifted from using transcription services to the implementation of speech recognition systems. Speech recognition systems replace expensive transcription services and enable much quicker report delivery. However, lower accuracy rates of these systems compared to transcription services have affected the productivity of radiologists by causing them to spend a larger portion of their time correcting the inaccuracies of the computer generated report [1,2]. Over the past several years, speech recognition technology has greatly improved, with some vendors now advertising accuracy rates of up to 99 percent [3]. But the few errors that do occur in radiology reports can have profound effects when clinicians rely on the reports to make life-altering decisions. For example, a speech recognition engine can interpret a radiologist as saying, "There is *now* evidence of tuberculosis," when the radiologist actually said, "There is *no* evidence of tuberculosis."

Overall speech recognition error rates can be reduced by several means. For example, most speech recognition engines can be "trained" or tailored to an individual's voice and speech pattern. Such training can dramatically improve the accuracy rate of the engine. Additionally, many radiologists utilize macros, which serve as templates for reports that are commonly used. Typically, only a few blanks, called fields, in the macro need to be dictated. Macros thereby reduce the error rate of dictation by limiting the amount of text that is generated by the speech recognition engine. Furthermore, radiologists proofread their reports before signing them. However, even a single error can cause an incorrect diagnosis. Therefore, every additional effort should be made to systematically improve the accuracy of speech recognition.

Most modern speech recognition engines rely upon probabilistic measures of word combinations, represented in a language model to translate audio input into words. Typically, statistical language models of speech recognition engines use trigrams, or three-word phrases, to determine which words the speaker used. While products such as Dragon NaturallySpeaking Medical (Nuance, Burlington, MA) include a medical dictionary and a radiology language model, no effort has been made to use the unique properties of radiological examinations to further refine the language model [4,5]. The goal of this study was to demonstrate that certain properties of imaging examinations, such as the body site, modality, and subspecialty, which are known prior to dictation, have an effect on word and trigram frequency that is comparable to the identity of the speaker. Because these former attributes of the report are known in advance, they could be used to dynamically refine the language model used by speech recognition engines, thereby improving accuracy.

2. Methods

2.1. Setting

The Radiology department of the Hospital of the University of Pennsylvania has stored reports electronically in a Radiology Information System (GE Healthcare, Waukesha, WI) for about two dec-

* Corresponding author. Fax: +1 215 349 5925.
E-mail address: langlotc@uphs.upenn.edu (C.P. Langlotz).

1532-0464/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2008.08.001

ades, resulting in over two million digital radiological reports [6]. An automated speech recognition system with macro capabilities [7] (TalkStation, Agfa-Gevaert, Mortsel, Belgium), which incorporated an earlier version of the Dragon speech engine (Nuance, Burlington, MA), was implemented in late 1999.

2.2. Research database

The data was loaded into an Oracle 10 g Standard Edition database server (Oracle, Redwood Shores, CA), installed on a AMD Athlon desktop running Windows XP SP2. The data set for this study was the same as was used for the study conducted by Lakhani et al. [6]. The study sample consisted of 2,169,967 completed radiology reports from the Hospital of the University of Pennsylvania between January 1, 1998 and November 11, 2005.

The authors removed patient identifiers from the database, including name, date of birth, and medical record number. Of the remaining reports, 29,736 contained only patient information with no report text. The authors deleted all 29,736 of these records. There remained 72 reports that contained the patient's name within the report text. The authors manually edited these to remove the patient information.

2.3. Comparing sets of words and trigrams

We opted to develop our own simple tokenizer because to our knowledge no existing tokenizers could be executed directly by the database, which was necessary to reduce processing time. Thus, each report was parsed into a set of tokenized words. Our tokenizer split words based upon a defined list of text delimiters, such as spaces, periods, commas, colons, semicolons, and letter-number interfaces. An exception list was created to list common delimiter characters that do not act as a delimiters but rather add meaning to the word, such as numbers with decimal points, times with colons, and 359 predefined text strings such as "H20", "2ND", "GM/CM", "T9/10", and "K-WIRE".

We compared the effects of modality, body site, and subspecialty to the effects of radiologist. We conducted four experiments comparing word frequency and one experiment comparing trigram frequency. For each comparison, we randomly divided our database into two subsets of reports of equal number. This enabled comparison of word frequency for both concordant and discordant reports. For the first word comparison experiment, we further subdivided each report subset by modality. We then generated a modality comparison matrix between modality pairs. Each comparison pair had an equal number of reports from each of the two database subsets. For example, we compared 10 modalities, which generated a 100 cell comparison matrix, in which we compared word frequencies between the two halves of the database.

The second experiment subdivided each half of the database of reports by both modality and subspecialty and compared the word frequencies in the subgroups. The third experiment subdivided each half of the dictionary by modality and body site and compared word frequency between the halves. The fourth experiment compared word frequencies in sets of reports by radiologist in similar fashion.

In addition to the experiments comparing word frequency, we also conducted experiments in which reports and their corresponding trigrams were divided and compared by modality as above. To reduce the huge computational burden, we limited the analysis of trigrams to only those occurring more than once in the database.

2.4. Metrics for comparing word and trigram frequencies

The metric used for comparison of word and trigram frequencies was based on the log-likelihood score (G^2). The log-likelihood score assists in finding words or trigrams that are particularly characteristic of a report. This statistic gives an accurate measure of the "surprising" nature of an event and gives a sense of the "distinctive" nature of a corpus [8].

Table 1
Table for determining the log-likelihood score [9]

	X	Y	
w	a	b	$a+b$
not w	c	d	$c+d$
	$a+c$	$b+d$	$a+b+c+d=N$

Kilgarriff [9] defined a framework for using the log-likelihood test to obtain the G^2 statistic for a given word, w, given two texts, X and Y. In our setting, X and Y were the two halves of our dataset. We tabulated a (the number of occurrences of w in X), c (the number of words in X that were not w), b (the number of occurrences of w in Y), and d (the number of words in Y that were not w). Table 1 lays out these calculations in a table. To then find G^2, Eq. (1) is used [9]. This calculation was performed only for words that existed in both halves of the data set.

Computation of the G^2 statistic, which is based on the log-likelihood score.

$$G^2 = 2(a\log(a) + b\log(b) + c\log(c) + d\log(d) - (a+b)\log(a+b) - (a+c)\log(a+c) - (b+d)\log(b+d) - (c+d)\log(c+d) + (a+b+c+d)\log(a+b+c+d)) \quad (1)$$

From the G^2 values, the significant log-likelihood (SLL), Eq. (2), expresses the percent of distinct words or trigrams that had significant G^2 values ($p < 0.05$). P-values were calculated from G^2 using a reference standard [10].

The significant log-likelihood (SSL) metric measures the percent of significant ($p > 0.05$) n-grams.

$$\mathrm{SLL} = \frac{n\ \mathrm{grams}_{G^2>3.84}}{n\ \mathrm{grams}_{\mathrm{mod\ el}} \cup n\ \mathrm{grams}_{\mathrm{test}}} \quad (2)$$

For each comparison of corpora, we used the G^2 statistic to compare frequencies of a word or trigram between subsets of reports then used the SSL log-likelihood statistic to determine significance of difference in word frequencies between corpora.

3. Results

3.1. Univariate analysis

After de-identifying the data and tokenizing the reports, there were 2,169,967 reports, consisting of 338,435,512 words (tokens). On average, there were 155.9 words per report with a standard deviation of 119.5. The distribution had a positive skew, with a mode of 61 (Fig. 1). The longest report contained 2620 words.

Table 2 shows the distribution of reports among the 12-modality types defined in this report database. Table 3 shows the distribution of reports across body site. Because we randomly split the data, an equivalent number of reports reside in each subgroup. An analysis of the properties of report length distribution and distributions for modality, body site, and subspecialty for the model set and test set and found no significant differences, indicating that the two sets were properly randomized.

3.2. Comparison of word frequencies by modality

To test for distinctive words in each corpus, the log-likelihood test was run for each modality. Table 4, shows an example of the most significant word frequency differences between corpora.