Speech-Assisted Radiology System for Retrieval, Reporting and Annotation

Tim Weninger, Daniel Greene, Jack Hart, William H. Hsu
Department of Computing and Information Sciences
Kansas State University
Manhattan, KS, 66506
{weninger, dgreene, jhart, bhsu}@ksu.edu

Surya Ramachandran
AIdentity Matrix Inc.
960 Industrial Drive Ste. 7
Elmhurst, IL 60126
surya@aidentitymatrix.com

Abstract

We present a system capable of interpreting speech commands given by a radiologist in order to accurately diagnose a set of findings and impressions for medical images, such as MRI, CT, PET, etc. The system is also extended to interpret search cues from speech in order to retrieve previously annotated images and previously diagnosed patients, enabling Computer Aided Differential Diagnosis (CADD). This system uses advanced radiology techniques such as structured reporting and PACS to help radiologists by providing a natural and configurable spoken English interface. Finally, we experimentally show that the system provides a significant improvement in accuracy over existing methods.

1 Introduction

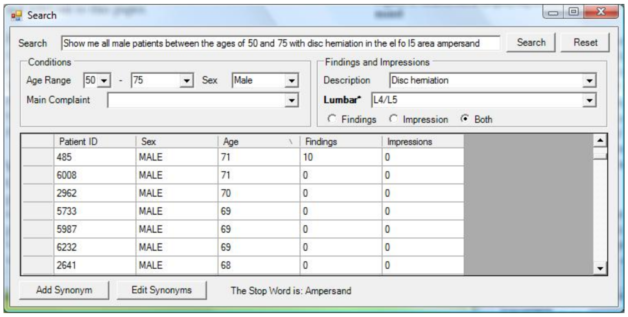
Medical imaging continues to play an important role in the advancement of medical science by non-invasively producing images of a body for use in clinical diagnosis. Once an image or a set of images is produced for a patient, radiologists use their domain knowledge to assert a set of findings about specific medical conditions they find. They also annotate images by specifically creating a note pointing out interesting features on one or more of the images. Currently, this is commonly done by dictating their findings into a voice recording device while drawing annotations directly onto the radiologic film. In these instances, a medical transcriptionist will then transcribe the voice recordings into text and then attach the text annotations to the set of images [1].

Several advances in voice recognition have allowed the creation of systems that eliminate the transcriptionist by converting the radiologist's dictations directly into a report, but these systems fail to extract any meaningful information from the dictated sentences. These voice recognition modules are often integrated into Picture Archiving and Communication Systems (PACS) and other reporting environments [5]. Recently, structured reporting systems take this a step further by asking the radiologist to choose certain attributes of the dictated sentences to enable the system to automatically classify the case [3].

In this paradigm, a *finding* is a medical diagnostic interpretation of a particular artifact as seen by the radiologist. An *annotation* is the expression (drawn arrow, circle, etc.) of a medical opinion related to a specific image. For example, consider a set of three MRI images taken of an elderly male patient with degenerative disc disease. A radiologist, seeing a positive indication of degeneration will dictate a finding such as - "Slight degenerative disc disease is identified in the L4/L5 and L5/S1 region". Further the radiologist might annotate two out of the three MRI images with arrows or circles highlighting the specific region.

The current process, described generally above, is deficient in many ways. (1) A radiologist is required to manually search and cross-reference patients' current images with their medical history and previous radiology scans. (2) Transcription errors may occur resulting in an inaccurate diagnosis or medication error. For instance, Berlin [2] found that 43% of all medication errors were caused by transcription errors. (3) Transcribed dictations cannot be easily attached to the specific image(s). Most commonly, the entire transcribed text is simply attached to the complete set of images without any mapping of the findings to their related images. (4) If the radiologist wishes to draw a marking on the image while dictating findings the radiologist may do so, however, these annotations are not exact and often serve only as a general visual cue to the area in question. (5) Image comparison between previous scans of the same patient or comparisons to similar cases becomes time consuming (finding other patients in a film archive) and inaccurate (impossible to narrow down the search criteria to return only relevant matches).

In a time when medical professionals most value accuracy and clarity of medical information, especially with the soaring cost of malpractice litigation seen in radiology and other medical specialties [6], a system is needed to solve these deficiencies while maintaining a high degree of accuracy. With these deficiencies in mind, we developed a system for the automatic retrieval, dictation of structured diagnoses, and annotation of medical images. This paper will describe the system with particular attention paid to its novel user interface characteristics. First, the software system is described briefly. Next, the voice-directed search, structured diagnosis, and image annotation mechanisms are described with screen captures of a working prototype shown to illustrate the interface. Afterwards, image-annotation embedding is described as a solution to the shortcomings of current annotation systems. Finally, experiments are desribed and results are shown that demonstrate the improvement this system makes over current methods.



Search: Show me all male patients between the ages of 50 and 75 with disc hemiation in the el fo l5 area ampersand

Conditions

Age Range: 50 - 75 Sex: Male

Main Complaint:

Findings and Impressions

Description: Disc hemiation

Lumbar*: L4/L5

○ Findings ○ Impression ● Both

	Patient ID	Sex	Age	Findings	Impressions
	485	MALE	71	10	0
	6008	MALE	71	0	0
	2962	MALE	70	0	0
	5733	MALE	69	0	0
	5987	MALE	69	0	0
	6232	MALE	69	0	0
	2641	MALE	68	0	0

The Stop Word is: Ampersand

Figure 1. Search form displaying results after interpreting a voice command

2 Voice-Directed Search

Currently, if a radiologist wishes to search for a past or present patient then the radiologist must either manually enter patient identification data into a computer form or retrieve the paper records and film manually. If, for example, while reviewing a current case the radiologist wishes to look at past instances of the same complaint, symptoms, etc. then the radiologist would be required to manually enter each symptom name and all of the appropriate search terms. Most PACS systems currently allow only a very limited scope of search related terms, such as "Main Complaint", "Social Security Number", etc. Using these limited keywords the radiologist must deal with several patients that are not directly pertinent to the current case being retrieved. Therefore, we argue that an English-based, medically oriented voice-directed search would be more practical, easier to use and would return more-relevant, patient cases.

Consider the completed search form shown in Figure 1 generated by a English query from a radiologist. Although the query information shown in the figure could have been manually entered, it was instead spoken and interpreted by the system. The form shown in Figure 1 was automatically populated by dictating, "Show me all male patients between the ages of 50 and 75 with disc herniation in the el for l5 area ampersand" [sic]. Note that the utterance does not need to be structured nor does it need to even contain the exact matches of the labels of the input menus. Moreover, the correct description "L4 *slash* L5" does not necessarily need to be so exactly phrased. Instead, the case-client parsing engine (CPE) was able to make the correct selection automatically. Radiologists may chose to include long pauses in a single utterance, therefore the trailing term 'ampersand' is necessary to signify the end of an utterance. The term 'ampersand' itself is not significant, rather this uncommon term within the medical dictionary was arbitrarily selected and can be substituted for any arbitrary word.

To populate the search form, the case-client first uses a commercially available speech recognition engine to internally create a text string representing the spoken query string. Although the speech recognition system claims to be 97% accurate without any training, our experience shows that small errors in deciphering utterances, especially in the medical domain, occur in some of the speech recognition system's text interpretations. Therefore a new approach is needed in order to parse and correct faulty speech interpretations.

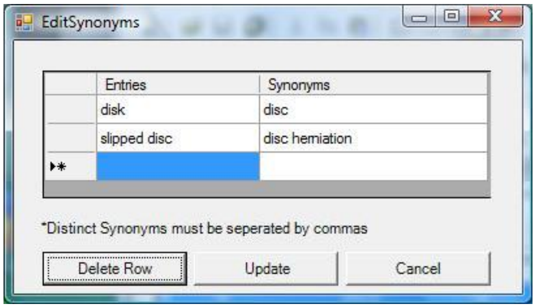
The idea of restricting user input to relevant menu items has been in use since at least the early 1980's (cf. [7]). The major problem with the original approach is that spoken words do not often contain exact matches to the items in the menu. For example, although the radiologist might utter "...disc buldge..." the corresponding menu item which is labeled "Buldging disc" will not be selected. Arguably, the most common complaint among users of speech-assisted medical systems is that the system severely restricts the word choice and verbage of the user. Our goal is to allow radiologists to speak to the system as if they were dictating for a transcriptionist in their most comfortable style.

2.1 Parsing Spoken Text

We operate under the assumption that what the radiologist dictates and what the speech recognition system should interpret is available to the parsing algorithm as background knowledge. Since the radiologist is attempting to select items from a menu, albeit a large menu, the parsing algorithm would easily be able to try and match uttered words with menu items rather than by attempting to parse the sentence with only grammar rules, etc. The voice-directed search algorithm takes, as input, the speech recognition system's interpreted word-string and uses a sliding window approach on that string to match the interpreted words with menu items. Here the 'sliding window' refers to a variable-size window that considers groups of one or more adjacent words when matching an utterance to a menu item. The window ‘slides’ when a menu item is selected. In this way the algorithm accounts for menu items that contain multiple words and the more difficult case of when separate menu items contain one or more of the same words. For example, ‘disc’, ‘disc herniation’, and ‘degenerative disc disease’ all are description menu items that all contain ‘disc’. The sliding window algorithm chooses the menu item with the most contiguous words in common with the words in the algorithm’s ‘window’.

Speech recognition systems often interpret medical terms incorrectly. For example, the utterance “show me *male patients*” is commonly misinterpreted as “show me *mail patience*”. To resolve these discrepancies we assume that hom-onyms do not exist in the same menu. With this in mind our algorithm chooses the menu item closest to the errant text when no textual match is available. The ‘closest’ item is determined by an edit-distance ranking algorithm [4].

Radiologists sometimes use non-standard medical terminology to describe findings (for example, “there is a slight disc herniation” and “there is a minor disc herniation”). Our system strives to standardize reporting in radiology. As a solution the case-client offers the ability to directly enter synonyms which automatically substitute one word for the other into the speech recognition system’s interpreted text. A powerful extension of this method provides the ability for radiologists to speak in the manner in which they are most comfortable. For example, if a radiologist chooses to synonimize the common finding ‘slipped disc’ with the standard medical terminology ‘disc herniation’ then they will thereafter be able to dictate ‘slipped disc’ to the case-client and the appropriate medical description will be selected. This extension also provides a way to standardize medical findings without inconveniencing the radiologist. Figure 2 shows the substitution list where the colloquial term ‘slipped disc’ is replaced with the formal, medical term ‘disc herniation.’



	Entries	Synonyms
	disk	disc
	slipped disc	disc herniation
▸*		

*Distinct Synonyms must be seperated by commas

Figure 2. List of words to by automatically substituted into the speech engine’s text interpretation

In the event that all attempts at correction fail, the errant menu item can be changed through manual selection.

3 Voice-Directed Findings

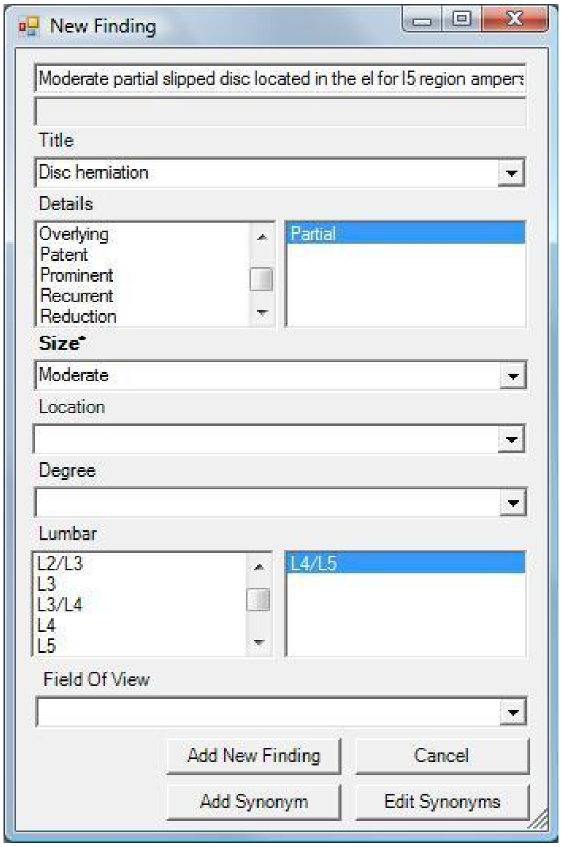
Radiologists can verbally or manually select appropriate patient-cases from the results list in the search results form or choose a new patient for diagnosis. The case-client interface contains a digital film stop, patient electronic medical records, and past history. In addition to these standard features, medical findings and their corresponding annotations are listed and can be edited, sorted and reviewed in detail. Command button also allow radiologists to quickly find only the images that indicate a particular finding, or omit images that do not indicate a finding.

During the review of a patient’s MRI images the radiologist dictates a finding by speaking free-form medical English text as described in the previous section. Figure 4 shows an example finding that has been automatically interpreted from the utterance: “New finding moderate partial slipped disc located in the el for l5 region ampersand” [sic]. The precise utterance ‘new finding’ cues the system that the proceeding utterance contains finding information. This practice of opening and closing free-text dictations is necessary to avoid miscellaneous speech to be incorrectly interpreted as a medical finding. As before, the utterance is parsed and the correct input boxes are populated; ‘slipped disc’ is again interpreted as ‘disc herniation’. Similar to the search process, the radiologist is able to manually enter and edit the interpreted finding before confirming its creation.

The radiologist may chose to further describe a finding by creating an annotation by selecting a shape or arrow and pointing to a specific part of the image using a stylus or a mouse. This input method is similar to the widely used practice of manually annotating an image. Annotations are linked to their respective findings and can be displayed, edited or deleted selectively when the radiologist chooses a particular finding. These capabilities are advantageous as the next section shall discuss.

4 Image-Annotation Embedding

Radiologists make annotations describing their findings directly on the film to identify specific locations for the benefit of other radiologists, referring physicians or interventional surgeons. This is commonly done by drawing shapes onto physical films with a grease pencil. While this free form annotation is useful, the resulting markings are separate from the indicated dictations and findings. This makes medical collaboration and the portability of medical records difficult because the other radiologist would not be able to definitively relate image annotations with findings



Moderate partial slipped disc located in the el for l5 region ampers

Title: Disc hemiation

Details: Overlying, Patent, Prominent, Recurrent, Reduction — Partial

Size*: Moderate

Location:

Degree:

Lumbar: L2/L3, L3, L3/L4, L4, L5 — L4/L5

Field Of View:

Figure 3. Completed new finding input form after a ‘disc herniation’ is described.

especially when multiple findings are indicated or there are multiple annotations on the same image.

For example, if a radiologist has 4 X-Ray images of a patient’s leg where two images show a fracture of the tibia and the two remaining images show a fracture in the femur then the radiologist would indicate two findings: (1) “Fracture in right tibia”, and (2) “Fracture in right femur”. In this example images 1 and 2 would show the lower leg (the location of the tibia) and images 3 and 4 would show the upper leg (the location of the femur). With image annotation embedding if a radiologist would like to see all of the images regarding the tibia fracture then the radiologist would select that finding and only images 1 and 2 would be displayed in the case-client. Maintaining this relatively simple relationship between findings, images and annotations has received mostly positive reviews among interested radiologists.

5 Experiments

An experimental study was performed in order to determine the accuracy of the case-client parsing engine (CPE). First, we obtained de-identified patient health information radiology transcriptions. These transcriptions are from radiologists who did not originally know that their dictations would be used for speech recognition, therefore we can safely assume that this text was freely-spoken without regard to its computer interpretation. We then separated the transcriptions into 75 individual annotation-phrases; this step simply involved separating sentences and the actual transcription text was not altered. Transcribed radiological queries are not available due to the novelty of this system, therefore we created 25 additional query-phrases based on the information contained in the 75 annotation-phrases. An example of annotation-phrases and the corresponding query-phrases are shown in Table 1.

Table 1. Example phrases used for testing. Annotation-phrases are taken directly from de-identified medical transcriptions. Query-phrases are based on the corresponding annotation-phrases.

Annotation-phrase	Query-phrase
There is moderate disc bulging at L5/S1	Show me all patient’s with moderate disc bulging at the L5/S1 location
There is moderate to marked narrowing of the L5/S1 foramina bilaterally	Show patient’s with narrowing of the L5/S1 foramina.

These 100 phrases were printed to paper and read verbatim by a radiologist to the case-client via a standard, non-noise reducing computer microphone. The system was configured to record the raw, recognized text and an output of the CPE’s rendering. Afterwards, these results were compared to the text of the original 100 phrases.

5.1 Metrics

After the tests were performed, 4 important pieces of information are compared: (1) the original text read by the radiologist, (2) the text output by the speech recognition engine, (3) the menus filled by the CPE, (4) the correct state of the menus (*i.e.* ground truth).

This comparison results in 3 paradigms that ultimately define the accuracy of the system. First, we find describe the the control in terms of speech recognition acuracy resulting in the speech recognition metric (SRM). The SRM is defined as the edit distance between the original text (1) and the text output by the speech recognition engine (2). Specifically, the levenstein word distance (LWD) method is used to compute the distance. The LWD differs from the traditional levenstein edit distance [4] in that LWD regards words as atomic and therefore calculates the distance in terms of the total number of inserts and deletions of entire words rather than characters. For example, the LWD between “There is moderate disc bulging at L5/S1” and “moderate disc bulging at L5/S1” is 2 because the insertion of 2 words (’There’ and ’is) is needed to trasform the first sentence into the second. The metric used to describe the system is called the parsing engine metric (PEM). The PEM is defined as the difference between the menus filled by the system (3), and the correct state of the menus (ground truth) (4). For example, if the title menu shows “degenerative disc disease” when it should have recognized “Disc herniation” then an error of 1 will be added to the PEM.

Error percentages can be computed from the SRM and PEM by taking the distance (*i.e.* error occurence) over the number of possible errors. For example, in the LWD example above the there are 7 words in the correct sentence and the LWD is 2, therefore the percentage correct is $(7-2)/7 \rightarrow 5/7$ or about 71%.

Alternatively, some may argue that certain key words should be regarded to be more important than others. To that end, we developed a weighting scheme that counts the title menu as 60%, the lumbar location menu as 20%, all other menus as 20%, and inconsequential words (*i.e.* stop words) with 0%. Consider the following example, the correct menus for the sentence, “There is moderate disc bulging at L5/S1” should be title=“Bulging Disc”, lumbar=“L5/S1”, and size=“Moderate”. Notice that the sentence contains “disc bulging” and not “Bulging disc”; in this instance the sentence would be found to mention an incorrect title and therefore the have a SRM of 60%. However, in this instance the CPE did determine the correct title menu therefore the PEM would be 0%.

Finally, in speech-assisted systems users may be more interested in a draconian all-or-not (*i.e.* pass/fail) metric, that is if the SRM and/or the PEM is 0 (*i.e.* no error) then the result is correct otherwise it is not.

These 3 metrics essentially compute the accuracy of the speech regonition system (in terms of SRM) and the accuracy of the CPE (in terms of PEM) given imperfect speech recognition as input. In order to find the improvement that our system provides with these metrics, we plot the SRM and PEM percentages over the 100 cases, and then use regression to find line that best fits the data (*i.e.* results in the lowest R^2 value). The difference of definate integrals from 0 to 100 for the two line plots, shown in Equation 1 describes an area of improvement. This area of improvement is then calculated as a improvement percentage.

$$\int_0^{100} (m_p x + b_p)\, dx - \int_0^{100} (m_s x + b_s)\, dx \tag{1}$$

6 Results

Results of the experiements are all described by the metrics presented in the preceeding subsection. We begin this section with the line equations plotted by regression on the SRM and the PEM. In all but 1 case, linear regression resulted in a line with the lowest R^2 value. In the single odd case exponential regression resulted in only a slight advantage, therefore, for consistency-sake, linear regression is used to plot lines in all cases.

Table 2 shows the results for all metrics in all of the paradigms. The rows in the distance paradigm show the word counts where Max Sum is the maximum amount (*i.e.* perfect score) and the Result represents the speech recognition metric (SRM) and/or the parsing engine metric (PEM). The rows in the weighted paradigm have a maximum sum equal to the number of tests because the weights at most add up to 1. The rows in the All-or-Not paradigm again have a maximum sum equal to the number of tests because the perfect score would return 25 and 75 correct tests for Query and Annotation respectively. The percentage correct can be found by dividing Result over Max Sum.

Table 2. Results for query and annotation speech tasks recognition metric (SRM) (the control) and parsing engine metric (PEM) from each judgement paradigm.

Paradigm	Task	Test	Max Sum	Result
Distance	Query	SRM	375	321
		PEM	101	91
	Annotation	SRM	797	716
		PEM	252	222
Weighted	Query	SRM	25	21.13
		PEM	25	20.90
	Annotation	SRM	75	44.43
		PEM	75	63.28
All-or-Not	Query	SRM	25	5
		PEM	25	17
	Annotation	SRM	75	30
		PEM	75	52

Table 3 shows the linear equations as well as the R^2 and corresponding areas for the SRM and PEM metrics for each of the 3 analysis paradigms (Distance, Weighted, All-or-Not).

The graph in Figure 4 shows plots of the best fit lines as well as individual data points in grey. A close viewing of Figure 4 shows that the PEM lines are always above their corresponding SRM lines for each paradigm. This demonstrates a clear performance improvement that is quantified in Table Table 4.

Table 3. Results of lines plotted by linear regression on speech recognition metric (SRM) and parsing engine metric (PEM) in all 3 paradigms.

Paradigm	Metric	Line	Area
Distance	SRM	$-.0007x + .9115$	87.65
	PEM	$.00008x + .9604$	96.44
Weighted	SRM	$.0041x + .4643$	66.93
	PEM	$.0009x + .9047$	94.97
All-or-Not	SRM	$-.0025x + .4831$	35.81
	PEM	$-.0005x + .7296$	70.46

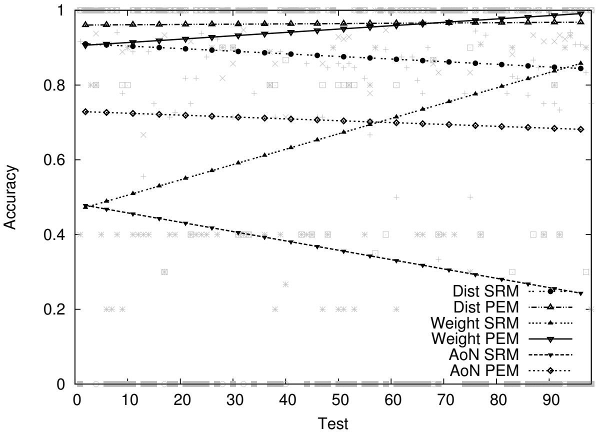


Figure 4. Graph of data points and linear regression lines.

The areas for each metric in each paradigm can be compared to render a percentage increase in accuracy that our system provides over the speech recognition system. Table 4 shows these increases.

7 Conclusions and Future Work

The results show that our system provides a significant improvement in speech recognition accuracy. Arguably the most straightforward interpretation of these results is that radiologists who use this system for speech-assisted image annotation and retrieval will not have to make corrections about 73% of the time, which is 96.76% better than the naively picking menu terms.

In conclusion, this paper describes a system capable of interpreting speech commands given by a radiologist in order to accurately diagnose findings from a set of medical images. Furthermore, we demonstrate a speech parsing algorithm which leverages the background knowledge of the medical paradigm in order to provide a natural and configurable, spoken English interface. Finally, we hypothesize that the integration of speech-assisted retrieval and annotation systems will help radiologists and hospital staff provide better healthcare. We plan to extend this system to incorporate more medical specialties. The system in its current form mostly handles images of and relating to the lumbar spine. We are currently working to extend the basic algorithm to emcompass all of MRI and the broader scope of radiology. We are also looking at ways this system can augment the input and retrieval of electronic medical records within a complete electronic healthcare solution.

Table 4. Percentage increase in accuracy for each paradigm ($(SRM - PEM)/SRM$).

Paradigm	Increase (Accuracy)
Distance	10.0285%
Weighted	41.8945%
All-or-Not	96.7607%

8 Acknowledgements

The authors thank Michael Schekall, MD and Deborah Templeton, BS, CNMT, RT(R), LRT of the Hutchinson Clinic PA, Hutchinson, KS, for their medical input. We also thank Jeff Barber and Andrew Walters for their help.

References

[1] Benitez Y., Forrester L., Hurst C. and Turpin D. Hospital Reduces Medication Errors Using DMAIC and QFD. *Quality Progress*, 2007.

[2] Berlin L. Malpractice Issues in Radiology. *American Journal of Roentgenology*, 6(170):1417–1422, 1998.

[3] Cox C., Phalen J., Dworak T. Voice Recognition Dictation: An Adjunct to Medical Student Radiology Education. *Academic Radiology*, 14(2):221–228, 2007.

[4] Levenshtein V. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Soviet Physics-Doklady*, 10(10):707–710, 1966.

[5] Pilling J. R. Picture archiving and communication systems: the users view. *British Journal of Radiology*, 76:519–524, 2003.

[6] Studdert D. M., et al. Claims, Errors, and Compensation Payments in Medical Malpractice Litigation. *New England Journal of Medicine*, 354(19):2024–2033, 2006.

[7] H. R. Tennant, K. M. Ross, R. M. Saenz, C. W. Thompson, and J. R. Miller. Menu-based natural language understanding. In *Proceedings of the 21st annual meeting on Association for Computational Linguistics*, pages 151–158, Morristown, NJ, USA, 1983. Association for Computational Linguistics.