Characterizing and Quantifying Diagnostic (Un)Certainty in Medical Reports Through Natural Language Processing

Kathleen Isenegger-Contact Author
Dept. of Computer Science
Amherst College
Amherst, Massachusetts
kisenegger20@amherst.edu

Yilan Dong
Dept. of Computer Science
Grinnell College
Grinnell, Iowa
dongyila@grinnell.edu

Mengyuan Shang
College of Education and Human Sciences
University of Nebraska-Lincoln
Lincoln, Nebraska
mshang2@unl.edu

Jacob Furst
School of Computing
DePaul University
Chicago, Illinois
jfurst@cdm.depaul.edu

Daniela Raicu
School of Computing
DePaul University
Chicago, Illinois
draicu@cdm.depaul.edu

***Abstract*—Miscommunication of diagnostic uncertainty can deeply affect the quality of treatment a patient receives. A standardized quantification based on the language used in medical reports is a solution for gaining clarity about the amount of uncertainty an author intended to convey. We use natural language processing techniques to create a dictionary of terms and phrases used in a corpus of radiology reports that are indications of uncertainty or certainty. Using this dictionary, we model reports by analyzing them as both a collection of sentences and a collection of words. We assign reports a rating on a scale of 0-5 to quantify how uncertain a particular report is. Our results suggest that by using a dictionary of both certainty and uncertainty descriptors, we can characterize and quantify diagnostic uncertainty of medical reports.**

***Index Terms*—medical uncertainty, diagnostic uncertainty, natural language processing, NLP, word2vec**

Full/Regular Research Paper for CSCI-ISHI

I. Introduction

Uncertainty in medicine can arise with the use of any medical instrument or test, as well as at the stage of diagnosis, as there is always a margin of error. Diagnostic uncertainty is an innate aspect of the medical field with major impacts: at least 1 in 20 U.S. adults experience diagnostic errors [1]. A misunderstanding between two parties about the intended amount of uncertainty to be conveyed can result in incorrect decisions by either over treating or under diagnosing [2]. These misunderstandings are prevalent as the manifestation of diagnostic uncertainty is a vague area without clear models or measurements. Medical reports, due to their direct link to patient care, are a modality with greater consequences in terms of miscommunicating a level of diagnostic uncertainty. A standardized model to quantify the uncertainty intended by an individual writing a medical report could have significant benefits in managing diagnostic uncertainty.

II. Background and Related Work

A. Medical Uncertainty

Past research has shown that manually identifying uncertain sentences in medical text is an achievable task [3], [4], [5]. Diagnostic uncertainty is defined through the language that is used in medical reports to express a patient's condition [6]. Specific words and phrases are shown to be indicators of (un)certainty. When annotating medical reports for uncertainty manually, experts look for this diction as an indication that a sentence should be marked.

In [4], guidelines are provided for manual annotation of speculation and scope in medical text in creation of the BioScope corpus, a collection of sentences annotated for speculation and negation. Similar work was done by [3] in which they attempted to classify sentences as high, low, and no speculation, but found that a significant difference between high and low speculation could not be achieved. This result provides support for a binary classification between no speculation and speculation, but does not consider expressed certainty.

While medical reports, among all medical text, have the most direct impact for individual patient diagnosis, the corpus used by [3] contains only abstracts and no medical reports, and the BioScope corpus constructed by [4] contains only 1,954 radiology reports. The limited amount of medical reports used in previous research prevents a complete understanding of the language used in medical reports indicating (un)certainty.

Research has also shown that it is possible to classify sentences as uncertain given a human-annotated ground truth [5], [7], [8]. As far as automated methods, the natural language text processor developed by [9] to identify clinical information in radiology reports strove to identify the following concepts: "no", "low certainty", "moderate certainty", "high certainty", and "cannot evaluate". It was not clear, however, how these concepts were assigned to the reports and what they represent quantitatively in terms of levels of (un)certainty.

What has not been shown is a completely automatic method for quantifying (un)certainty using a dictionary of terms and phrases that either indicate certainty or uncertainty. While this analysis has only been previously performed at a sentence-level analysis, our work includes both sentence and word-level analysis.

B. Rating Uncertainty

Reiner's [6] study proposes a standardized method for quantifying and characterizing diagnostic uncertainty using a scale from 0 (definitive level of certainty) to 5 (highest degree of uncertainty). To our knowledge, no other research proposing a comparable scale or method for rating diagnostic uncertainty exists. This scale is reasonable to implement partially due to its simplicity. The goal of quantifying uncertainty in medical reports is so authors and readers can easily use an automated system to clarify the amount of uncertainty present. A 6-level scale can capture the complexities of (un)certain language used in medical reports and offer a quantification to lessen the opportunities for miscommuncation of diagnostic uncertainty.

This scale is not clearly defined, but the descriptions that are given use words like "high", "highest", and "intermediate" to explain what each level of the scale represents. As these words are comparative in nature, they are not necessarily in agreement with a goal of being objective, as there may be variance between corpora as to what the highest level of uncertainty is. There is also an assumption that the absence of uncertain language is equivalent to "no uncertainty". However, [6] characterizes medicine as a generally imprecise practice. Therefore, the "neutral" language of a medical report may be less uncertain than directly uncertain language, but more uncertain than directly certain language, an observation [6] fails to appreciate. [6] does not use data to implement the proposed model or specify how it should be applied to actual medical reports. Improving on this previous work, the implementation we propose is a data-driven approach based on NLP to quantify diagnostic uncertainty. As the scale is specific towards radiology reports, this is they type of data we will use. The purpose of our analysis is to provide a comprehensive pipeline of rating uncertainty in medical reports and ultimately show the actual distribution of uncertainty in a set of data that is likely generalizes for all radiology reports.

III. Methodology

We implement the scale of [6] on a dataset of 20,238 medical reports to show that it is a valid way of representing the uncertainty communicated through the language of a report. Our methodology proposes two ways to model data and assign ratings.

A. Data

We built a dataset to be used in this research that is composed of medical reports from three databases of radiology teaching files: Medical Image Resource Center (MIRC) [10] (2319 teaching files), MyPACS [11] (16195 teaching files), and Eurorad [12] (7307 teaching files). These teaching files are formatted into 10 sections. During preprocessing, we removed all but 5: Title, Findings, Diagnosis, Discussion, and DDX. DDX stands for differential diagnosis, a type of diagnosis in which a healthcare provider compares different possible pathologies.

Typical natural language preprocessing steps were carried out to remove punctuation, numeric values, excess white space, make text lowercase, and stem each term in the text, and remove files under 10 words as they did contain enough semantic information for analysis. After all pre-processing steps were complete, the database of medical reports contained 20,238 distinct teaching files.

B. Dictionary of (Un)certainty Descriptors

To begin creating a dictionary that will adequately encompass the (un)certain language used in this corpus of medical reports, one can turn to experts and the language they categorize as "certainty descriptors" in medical ontologies. These terms and phrases are meant to be used to describe a range of uncertainty, including opinions of complete certainty. The first ontology used is RadLex from RSNA Informatics [13] This ontology clearly organizes terms and phrases into a "certainty descriptors" list, but does not differentiate between terms and phrases that are certain and uncertain. For example, both the terms "definite" and "uncertain" are featured in this category. The second ontology, SNOMED Clinical Terms (SNOMED CT) [14], can provide similar terms and phrases under the heading "finding status values" which contains subheadings such as "qualifier for certainty of diagnosis".

We separated these terms and phrases into two distinct word lists, certainty descriptors and uncertainty descriptors, due to the relative ambiguity of phrases and terms taken from these ontologies. The presence of terms and phrases expressing complete certainty in the medical ontologies shows that they are important elements for quantifying the uncertainty of a medical report, but there must be a differentiation between these semantically contrasting terms and phrases.

Using only these expert-defined terms and phrases as an initial dictionary of (un)certainty descriptors, we found that less than 50% of our teaching files contained these terms and phrases. Given the ubiquity of uncertainty in medical diagnosis, this finding demonstrates that this was not an adequate depiction of the language used for (un)certainty in our corpus of teaching files.

C. Expansion of Dictionary: Prior Works

Prior research [3], [4], [5] involving human linguistic annotation provided additional terms and phrases that were determined to denote uncertainty. Adding these terms and phrases to the initial dictionary generated from the medical ontologies yielded a dictionary of (un)certainty descriptors with 86 terms and phrases indicating uncertainty and 5 terms and phrases indicating certainty. Using this dictionary of (un)certainty descriptors, we found that 15.5% of teaching files contained no instances of any certainty descriptor. This dictionary contained words and phrases that clearly indicate (un)certainty, but did not yet consider the context of (un)certainty in medical reports.

D. Expansion of Dictionary: Word2Vec

A useful tool for considering the context of language is Word2Vec [15]. Word2Vec produces high-quality word vectors that models similar words in close proximity in a multidimensional space. Words are considered to have multiple degrees of similarity so that different qualities of a particular word can be analyzed for similarity in a subspace of the overall model. Word2Vec is also able to model n-grams accurately [16]

We used a Word2Vec model to learn new (un)certainty terms and phrases to add to the dictionary based on the cooccurrence of these terms and phrases in the corpus of teaching files with the terms and phrases already in our dictionary of (un)certainty descriptors. We expect the words garnered from this method would be more characteristically latent (un)certainty terms and phrases related to (un)certain concepts, ideas, and contexts. For this process, we built a Word2Vec model using all cleaned but un-stemmed data and generated lists of the top 10 similar words for each of our (un)certainty dictionary terms and phrases which included 2 through 6 word n-grams. We then used the Python library TextBlob to tag parts of speech and only added adjectives, verbs, and adverbs to our dictionary. This was done to combat the noisiness of medical terms and diagnoses that are prevalent in our dataset, since there is a greater chance of these commonly used medical terms and diagnoses to co-occur with (un)certainty terms and phrases. It is the (un)certainty descriptors, however, not the diagnoses themselves, which indicate (un)certainty. Therefore, we considered the medical terms and diagnoses noise when they appeared in the Word2Vec similarity lists and did not add them to the (un)certainty dictionary.

After this expansion, our dictionary contained 254 (un)certainty descriptors: 230 uncertain terms and phrases, and 24 certain terms and phrases. While the initial division of certain/uncertain terms and phrases was performed manually, all words added to the dictionary from the Word2Vec model were automatically assigned to whichever class their parent word was in. A final manual check was performed to assure words were in the correct class before proceeding. With the expanded dictionary, only 11% of teaching files contained no instances of any (un)certainty descriptors. Although there were many more terms and phrases in the uncertainty class of the dictionary, we chose not to perform normalization because most teaching files do not contain most of the terms and phrases in the dictionary.

We continued to get more similar terms and phrases from our Word2Vec model to see if results could be improved, but, as a considerable decrease in teaching files without (un)certainty descriptors was not seen, the dictionary was left at 254 descriptors. Finally, we stemmed each term in the dictionary. For the n-grams, we stemmed each term in the ngram separately and rejoined the stemmed terms with a white space.

E. Modelling and Rating Uncertainty

In order to give a more complete characterization of (un)certainty in medical reports, we chose to analyze teaching files on not just the number of times (un)certainty descriptors appeared throughout a document, explained in the next section, but also by how many sentences contain (un)certainty descriptors.

1) Sentence Level: for sentence level analysis, we calculated a relative count of the uncertain sentences in a teaching file. If a sentence contains any term from the uncertain class, it was labelled as uncertain, whereas sentences with any term from the certain class were labelled as certain. It is possible for sentences to be labelled as both certain and uncertain in our algorithm. We then subtracted the number of certain sentences from the number of uncertain sentences, so that any sentence labelled as both certain and uncertain would be considered the same as sentence that was not labelled. The final result was then found by dividing the previous result by the total number of sentences in that teaching file, in order to account for variable document length.

2) Word Level: : for a word level analysis, multiple ways of calculating uncertainty values were considered. Performing an absolute count analysis, in which each instance of a term is given equal weight regardless of the length of a teaching file or what the rarity of the term in question, is too simple of an approach. Relative count provides for more sound results as it does take the length of a document into account. Still, there is reason to believe that more common language should not so heavily influence the outcomes. For example, the uncertainty term "or" is very frequently used in our data (a total of 52769 occurrences). Therefore, we chose to use Term Frequency Inverse Document Frequency (TF-IDF) in this model.

TF-IDF is a common statistic used in information retrieval and NLP. [17] introduces the idea of weighing a term based on its occurrences throughout the corpus in document retrieval. This study states that some frequent terms and phrases can be less effective as a means of retrieval due to their non-discriminating nature and proposes that less frequent words be more valuable. More specifically, this study suggests correlating a terms matching value with its collection frequency (how many times it appears throughout all documents). This is the basic idea of inverse document frequency (IDF).

[18] later introduces the concept of TF-IDF by combining Jones' [17] idea of IDF and Luhn's [19] theory of term frequency (TF). The latter assumes frequently-used terms and phrases to be often essential to the content of the document. Salton's [18] approach accounts for both specification of the content of a single document as well as the relationship of such a document within a larger corpus by giving more weight to not only terms and phrases that are more frequent, but also terms and phrases that appear less across all documents. This study finds TF-IDF to be a more effective approach for indexing purposes and information retrieval compared to conventional boolean retrieval [20].

We generated the TF-IDF scores for the teaching files based on the following formulas:

$$w_{i,j} = tf_{i,j} \times \log(\frac{N}{df_i}) \tag{1}$$

$$\textit{TF-IDF}_{d_i} = \sum w_{(U)i,j} - \sum w_{(C)i,j} \tag{2}$$

where in (1), $w_{i,j}$ is the TD-IDF values of the jth (un)certainty descriptor in the ith document, $tf_{i,j}$ is the term frequency of jth (un)certainty descriptor in the ith document, or the sum of occurences of the jth (un)certainty descriptor in the ith document divided by the length of the ith document, N is the total number of documents in the corpus, and df_i is the number of documents that contain the jth (un)certainty descriptor; in (2), TF-IDF$_{d_i}$ is the TF-IDF values of the ith document, $\sum w_{(U)i,j}$ is the TF-IDF value of the jth uncertainty descriptor in the ith document, while $\sum w_{(C)i,j}$ is the TF-IDF value of the jth certainty descriptor in the ith document. To summarize, we first calculated the TF-IDF value of each descriptor in the dictionary, summed up the TF-IDF values of all uncertain and certain descriptors and phrases separately, and then used the difference between the two sums as the TF-IDF values of the teaching files.

F. Hierarchical Clustering

To show the validity of ratings generated from the TF-IDF values, we performed hierarchical clustering as a proof of concept on the entire corpus. With clusters of teaching files generated from hierarchical clustering, we then assigned a second set of (un)certainty ratings to files in each cluster, based on the most frequent TF-IDF value-generated rating in the corresponding cluster. We chose hierarchical clustering over k-means clustering, due to the repeatability and lack of randomness of the former, regardless of the number of clusters. We define accuracy of clustering-assigned ratings to be the number of teaching files whose two ratings match, over the total number of teaching files.

We first generated feature vectors for every teaching file. Each vector had 258 features: TF-IDF values of each term and phrases in the (un)certainty descriptor dictionary (254 features), the sum of TF-IDF values of all certainty descriptors, the sum of TF-IDF values of all uncertainty descriptors, the overall TF-IDF value of the teaching file, and the median of TF-IDF values of all teaching files. This last feature was included as a constant to avoid the existence of zero-vectors, as they would prevent the usage of cosine similarity as a distance measure. We used both cosine similarity and euclidean distance to cluster the files.

To determine the optimal number of clusters, we calculated the accuracy of clustering-assigned ratings on 6 to 1000 clusters. Less than 6 clusters did not allow for potential 100% accuracy, as there would not be enough clusters for every rating. Accuracy greatly leveled off with 1000 clusters, so this number was deemed large enough for our analysis.

IV. Results

A. Modelling and Rating Uncertainty: Assignment of Uncertainty Level Ratings

1) Sentence Level: in the sentence level model, the range of relative sentence count values was standard going from -1 to 1. We defined ratings of the 0-5 scale by an equal division of the model into 6 ranges. In this scale, 0 represents a definitive level of certainty and 5 represents the highest level of uncertainty. The distribution of these ratings is shown in the top row of Table I.

2) Word Level: in the word level model, we obtained TFIDF values ranging from -0.0767 to 0.2055. The findings from this analysis show that the model takes the shape of a relatively normal curve. This observation was confirmed by the proximity between the mean (0.0181) and the median (0.0184) of TF-IDF values of all teaching files. Therefore, ratings are assigned using the typical distribution of a bell curve with respect to standard deviation. Rating distribution is shown in the second row of Table I.

B. Hierarchical Clustering: Accuracy of Clustering-Assigned Uncertainty Level Rating

We graphed accuracy of uncertainty level rating assignment by cluster number by both euclidean distance and cosine similarity, and manually picked out points of interest, such as points that indicate significant increase in accuracy, beginning and end points, as well as the cluster number at which accuracy reaches 80%. Table II shows the selected number of clusters and their corresponding accuracy.

V. Discussion

What our research has shown is that it is possible to quantify uncertainty based on (un)certain language in more than one way with normalized results. However, without a gold standard truth, it is not possible to determine what the best way to assign ratings really is. Still, we are able to show the natural distribution of uncertainty in medical reports in two different ways, and this is a characterization of uncertainty in medical reports that has not been previously presented.

Our method tries to adequately capture the complexities of language by not limiting our dictionary to the terms and phrases that can be manually identified as indicators of

TABLE I
THIS TABLE SHOWS THE DISTRIBUTION OF RATINGS AMONG EACH MODEL AS WELL AS THE AGREEMENT OF RATINGS BETWEEN BOTH MODELS.

Teaching Files with:	Rating 0	Rating 1	Rating 2	Rating 3	Rating 4	Rating 5	Total Number of Files
Sentence Level (S)	19	24	2937	4537	10990	1731	20238
Word Level (W)	216	3038	6633	8103	1842	406	20238
Agreement (A)	9	11	481	931	1159	114	2705
$\frac{A}{S \cup W}$	4.0%	0.4%	5.3%	8.0%	10.9%	5.6%	15.4%

TABLE II
CLUSTER LABEL ACCURACY

Cluster Number	Euclidean Accuracy	Cosine Accuracy
6	74.45%	55.78%
10	**80.48%**	55.79%
21	81.06%	56.78%
22	85.92% **(+4.88%)**	56.79%
127	89.68%	61.77%
128	89.68%	75.05% **(+13.28%)**
711	93.51%	**80.00%**
1000	93.91%	81.78%
Total Increase	19.46%	26.00%

(un)certainty. Our dictionary of descriptors includes the terms and phrases discovered from the Word2Vec model because they are likely to be associated with (un)certainty, even in a latent way, considering that they are used in similar ways to the (un)certainty descriptors defined by experts.

Our analysis on medical reports at the word level demonstrates that uncertainty follows a normal curve. Therefore, we use ranges based on the model of the data to dictate what rating a teaching file received. The results of the sentence level analysis do not have the same normalized shape, but rather a skewed distribution with very few teaching files at the left side of the model, but also very few at the other extreme. As these values range from -1 to 1, they are best fit to a quantification by equally dividing the range of values into 6 levels of uncertainty without any outlier calculations.

The low levels of agreement between the two analyses, shown in Table I, indicate that modelling uncertainty on a collection of words and a collection of sentences are very different processes conceptually. The reasoning behind doing a word level analysis is the available characterization of uncertainty from radiologists through RadLex and SNOMED. These (un)certainty descriptors are not organized as complete sentences, but rather individual terms and phrases that could be used in conjunction with one another. Modelling on individual words using TF-IDF presumes that descriptors are not equally (un)certain, and that a greater term frequency, regardless of distribution across sentences, or a lower document frequency, indicates greater uncertainty. Conversely, the sentence level analysis is primarily supported by previous research from linguists in the creation of corpora of sentences that are said to indicate uncertainty. These corpora consider a binary classification of sentences being either uncertain or not with the phrases indicating uncertainty being considered equally uncertain. Our analysis is much the same, but with the addition of a "certain" classification.

On the word level, teaching files containing either no (un)certainty descriptors or an equal number of uncertain/certain descriptors will have a quantification of 0.0, similarly with teaching files containing either no (un)certainty sentences or an equal number of uncertain/certain sentences on the sentence level. It is reasonable to expect that these teaching files would receive a rating of 2 or 3, but this expectation is not confirmed by a word-level analysis where they have a rating of 1 (minimal uncertainty). Additionally, it is worth noting that only 512 teaching files have a net negative score in this analysis, meaning there are the only 512 teaching files that are more certain than uncertain. In sentence level analysis, the normality of the range of values rates 0.0 in the middle at rating 2 as expected, but there are still only 171 teaching files with a net negative score. These results challenge [6] as the descriptions provided for ratings 0 and 1 show that teaching files in these categories should be more certain than uncertain. Instead, data shows that the limited amount and model range of teaching files with a net negative score means that a scale of uncertainty in medical reports should likely not be a symmetric scale of certainty and uncertainty but rather a scale in which most ratings are indicative of more uncertainty than certainty.

We observe from Table II that euclidean distance greatly outperforms cosine similarity. More specifically, given 6 clusters, an accuracy of 55.78% is achieved by cosine similarity, compared to 74.45% by euclidean distance. This observation is also evident when we set an accuracy threshold of 80%. Under euclidean distance, accuracy reaches 80% at as early as 10 clusters, while it takes 711 clusters under cosine similarity to achieve a similar accuracy. However, we see a greater overall increase in accuracy under cosine similarity (26.00% compared to 19.46% of euclidean distance), as well as likely a faster increase (biggest increase of accuracy between two adjacent cluster number of cosine similarity is 13.28% compared to 4.88% of euclidean distance). This is likely due to the fact that, in this model, the overall inaccurate nature of cosine similarity allowed for more room for improvement.

Although this result is seemingly contradictory to the general expectation, as cosine similarity is commonly used in document comparison [21], it can be explained by the use of TF-IDF values of (un)certainty descriptors as features in document vectors. It is typical of document comparison for every word in a corpus to be a feature which results in very large, sparse vectors. Since our vectors only included the 254 (un)certainty descriptors, they were in comparison much

smaller and less sparse. The variability of document length is a main reason to use cosine similarity for document comparison, but TF-IDF values already normalize this factor. As a result, the magnitude of the vector is no longer an indication of the length of the file, but conveys more general information about the uncertainty of the teaching file. Euclidean distance better captures such information than cosine similarity.

A limitation of our research is our lack of ground-truth uncertainty ratings for medical reports to allow for a measure of accuracy of our uncertainty ratings, but it may be difficult to obtain valid manual ratings considering the minimal descriptions given in [6] explaining the 0-5 scale. Additionally, a further limitation could be that our dataset consists only of teaching files. These reports are written with a different intention than the average medical report, as they strive to teach or demonstrate a diagnosis. Therefore, there is reason to believe that they overall contain less uncertainty than the average medical report.

VI. Conclusion and Future Works

Miscommunication of diagnostic uncertainty in medical reports can lead to delayed or incorrect diagnosis, and the quantification of this uncertainty can lessen, or possibly eliminate, such miscommunication. This study has implemented a scale of uncertainty in medical reports from 0 (definitive level of certainty) to 5 (highest degree of uncertainty) on a large corpus of data to show the shape and distributions of uncertainty in medical reports. We found that using NLP to detect instances of (un)certainty terms and phrases in medical reports can lead to the modelling of diagnostic uncertainty quantities in more than one way depending on if a medical report is conceptually viewed as a collection of words or of sentences. Additionally, a dictionary of these (un)certainty terms and phrases to be detected can be robust and include the language of (un)certainty found in most medical reports of a certain corpus by using medical ontologies, prior work on diagnostic uncertainty in medical texts, and Word2Vec to create the dictionary.

This work can be improved upon with the incorporation of a ground-truth for rating diagnostic uncertainty in medical reports. With this addition, it would be possible to compare the two different analyses presented here in terms of accuracy and determine if one can be considered better. The proposed distributions shown in our results could provide context to a manual rater to encourage more accurate and consistent ratings. Another advancement for this research could be to use medical reports that are not teaching files, as well as modelling (un)certainty in other medical texts such as journals. As the research shown in medical journals may later be incorporated into diagnosis, it is likely necessary that levels of uncertainty are clear to readers of these texts as much as readers of medical reports. [10]

References

[1] H. Singh, A. N. Meyer, and E. J. Thomas, "The frequency of diagnostic errors in outpatient care: estimations from three large observational studies involving us adult populations," *BMJ Qual Saf*, vol. 23, no. 9, pp. 727–731, 2014.

[2] V. Bhise, S. S. Rajan, D. F. Sittig, R. O. Morgan, P. Chaudhary, and H. Singh, "Defining and measuring diagnostic uncertainty in medicine: a systematic review," *Journal of general internal medicine*, vol. 33, no. 1, pp. 103–115, 2018.

[3] M. Light, X. Y. Qiu, and P. Srinivasan, "The language of bioscience: Facts, speculations, and statements in between," in *HLT-NAACL 2004 Workshop: Linking Biological Literature, Ontologies and Databases*, 2004, pp. 17–24.

[4] V. Vincze, G. Szarvas, R. Farkas, G. Móra, and J. Csirik, "The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes," *BMC bioinformatics*, vol. 9, no. 11, p. S9, 2008.

[5] C. Zerva, R. Batista-Navarro, P. Day, and S. Ananiadou, "Using uncertainty to link and rank evidence from biomedical literature for model curation," *Bioinformatics*, vol. 33, no. 23, pp. 3784–3792, 2017.

[6] B. I. Reiner, "Quantitative analysis of uncertainty in medical reporting: creating a standardized and objective methodology," *Journal of digital imaging*, vol. 31, no. 2, pp. 145–149, 2018.

[7] M.-H. Laves, S. Ihler, and T. Ortmaier, "Uncertainty quantification in computer-aided diagnosis: Make your model say "i don't know "for ambiguous cases," 2019.

[8] P.-A. Jean, S. Harispe, S. Ranwez, P. Bellot, and J. Montmain, "Uncertainty detection in natural language: A probabilistic model," in *Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics*. ACM, 2016, p. 10.

[9] C. Friedman, P. O. Alderson, J. H. Austin, J. J. Cimino, and S. B. Johnson, "A general natural-language text processor for clinical radiology," *Journal of the American Medical Informatics Association*, vol. 1, no. 2, pp. 161–174, 1994.

[10] Radiological Society of North America, "Rsna mirc site," http://mirc.rsna.org/query, Last accessed on 2019-06-08.

[11] Change Healthcare, "Mypacs.net," https://www.mypacs.net/, Last accessed on 2019-06-08.

[12] European Society of Radiology, "Eurorad beta," https://www.eurorad.org/, Last accessed on 2019-06-08.

[13] Radiological Society of North America, "Rsna informatics radlex," http://www.radlex.org/, Last accessed on 2019-06-08.

[14] SNOMED International, "Snomed ct," https://www.snomed.org/, Last accessed on 2019-06-08.

[15] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[16] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[17] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, 2004.

[18] G. Salton and C. T. Yu, "On the construction of effective vocabularies for information retrieval," in *ACM SIGIR Forum*, vol. 9, no. 3. ACM, 1973, pp. 48–60.

[19] H. P. Luhn, "A statistical approach to mechanized encoding and searching of literary information," *IBM Journal of research and development*, vol. 1, no. 4, pp. 309–317, 1957.

[20] G. Salton, E. A. Fox, and H. Wu, "Extended boolean information retrieval," Cornell University, Tech. Rep., 1982.

[21] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.