Linking Uncertainty in Physicians' Narratives to Diagnostic Correctness

Wilson McCoy
Department of Interactive Games and Media
wgm4143@rit.edu

Cecilia Ovesdotter Alm
Department of English
coagla@rit.edu

Cara Calvelli
College of Health Sciences and Technology
cfcscl@rit.edu

Jeff B. Pelz
Center for Imaging Science
pelz@cis.rit.edu

Pengcheng Shi
Computing and Information Sciences
pengcheng.shi@rit.edu

Anne Haake
Computing and Information Sciences
anne.haake@rit.edu

Rochester Institute of Technology

Abstract

In the medical domain, misdiagnoses and diagnostic uncertainty put lives at risk and incur substantial financial costs. Clearly, medical reasoning and decision-making need to be better understood. We explore a possible link between linguistic expression and diagnostic correctness. We report on an unusual data set of spoken diagnostic narratives used to computationally model and predict diagnostic correctness based on automatically extracted and linguistically motivated features that capture physicians' uncertainty. A multimodal data set was collected as dermatologists viewed images of skin conditions and explained their diagnostic process and observations aloud. We discuss experimentation and analysis in initial and secondary pilot studies. In both cases, we experimented with computational modeling using features from the acoustic-prosodic and lexical-structural linguistic modalities.

1 Introduction

Up to 20% of post-mortem diagnoses in the United States are inconsistent with the diagnosis before death (Graber, 2005). These misdiagnoses cost both human lives and estimated millions of dollars every year. To find where and why misdiagnoses occur, it is necessary to improve our understanding of doctors' diagnostic reasoning and how it is linked to diagnostic uncertainty and correctness. Our contribution begins to explore the computational modeling of this phenomenon in diagnostic narratives. From a cognitive science perspective, we are contributing to the research on medical reasoning and how it is linguistically expressed. In the long term, this area of work could be a useful decision-making component for flagging diagnoses that need further review.

The study used an unusual multimodal data set collected in a modified Master-Apprentice interaction scenario. It comprises both gaze and linguistic data. The present study focuses on the linguistic data which in turn can be conceptualized as consisting of both acoustic-prosodic and lexical-structural modalities. This data set can further be used to link vision and language research to understand human cognition in expert decision-making scenarios.

We report on a study conducted in two phases. First, an initial pilot study involved a preliminary annotation of a small subset of the collected diagnostic narratives and also investigated the prediction of diagnostic correctness using a set of linguistic features from speech recordings and their verbal transcriptions. This provided initial features relevant to classification, helped us identify annotation issues, and gave us insight on how to improve the annotation scheme used for annotating ground truth data. Next, a second pilot study was performed, building on what was learned in the initial pilot study. The second pilot study involved a larger data set with a revised and improved annotation scheme that considered gradient correctness at different steps of the diagnostic reasoning process: (1) medical lesion morphology (e.g. recognizing the lesion type as a scaly erythematous plaque), (2) differential diagnosis (i.e. providing a set of possible final diagnoses), and (3) final diagnosis (e.g. identifying the disease condition as psoriasis). We also experiment with

Proceedings of the ACL-2012 Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics (ExProM-2012), pages 19–27, Jeju, Republic of Korea, 13 July 2012.

classification using an expanded feature set motivated by the initial pilot study and by previously published research. We report on results that consider different algorithms, feature set modalities, diagnostic reasoning steps, and coarse vs. fine grained classes as explained below in Section 4.3.

2 Previous Work

Much work has been done in the area of medical decision-making. Pelaccia et al. (2011) have viewed clinical reasoning through the lens of dual-process theory. They posit that two systems are at work in the mind of a clinician: the *intuitive* system which quickly produces a response based on experience and a holistic view of the situation, versus the *analytic* system which slowly and logically steps through the problem with conscious use of knowledge. Croskerry (2009) stated that "[i]f the presentation is not recognized, or if it is unduly ambiguous or there is uncertainty, [analytic] processes engage instead" (p. 1022); for instance, if a clinician is unfamiliar with a disease or unsure of their intuitive answer. We assume that different reasoning systems may cause changes in linguistic behaviors. For example, when engaging the slower analytic system, it seems reasonable that frequent pausing could appear as an indication of, e.g., uncertainty or thoughtfulness.

Several studies have explored the task of detecting uncertainty through language. Uncertainty detection necessitates inference of extra-propositional meaning and is arguably a subjective natural language problem, i.e. part of a family of problems that are increasingly receiving attention in computational linguistics. These problems involve more dynamic classification targets and different performance expectations (Alm, 2011). Pon-Barry and Shieber (2009) have shown encouraging results in finding uncertainty using acoustic-prosodic features at the word, word's local context, and whole utterance levels. Henriksson and Velupillai (2010) used "speculative words" (e.g., *could, generally, should, may, sort of*, etc.) as well as "certainty amplifiers" (e.g., *definitely, positively, must*, etc.) to determine uncertainty in text. Velupillai (2010) also applied the same approach to medical texts and noted that acoustic-prosodic features should be considered alongside salient lexical-structural features as indicators of uncertainty. In this work, we draw on the insight of such previous work, but we also extend the types of linguistic evidence considered for identifying possible links to diagnostic correctness.

As another type of linguistic evidence, disfluencies make up potentially important linguistic evidence. Zwarts and Johnson (2011) found that the occurrence of disfluencies that had been removed could be predicted to a satisfactory degree. Pakhomov (1999) observed that such disfluencies are just as common in monologues as in dialogues even though there is no need for the speakers to indicate that they wish to continue speaking. This finding is important for the work presented here because our modified use of the Master-Apprentice scenario results in a particular dialogic interaction with the listener remaining silent. Perhaps most importantly, Clark and Fox Tree (2002) postulated that filled pauses (e.g., *um, uh, er*, etc.) play a meaningful role in speech. For example, they may signal that the speaker is yet to finish speaking or searching for a word. There is some controversy about this claim, however, as explained by Corley and Stewart (2008). The scholarly controversy about the role of disfluencies indicates that more research is needed to understand the disfluency phenomenon, including how it relates to extra-propositional meaning.

3 Data Set

The original elicitation experiment included 16 physicians with dermatological expertise. Of these, 12 were attending physicians and 4 were residents (i.e. dermatologists in training). The observers were shown a series of 50 images of dermatological conditions. The summary of this collected data is shown in Table 1, with reference to the pilot studies.

The physicians were instructed to narrate, in English, their thoughts and observations about each image to a student, who remained silent, as they arrived at a differential diagnosis or a possible final diagnosis. This data elicitation approach is a modified version of the Master-Apprentice interaction scenario (Beyer and Holtzblatt, 1997). This elicitation setup is shown in Figure 1. It allows us to extract information about the Master's (i.e. in this case, the physician's) cognitive process by coaxing them to

Data parameters	Quantity
# of participating doctors	16
# of images for which narratives were collected	50
# of time-aligned narratives in the initial pilot study	160
# of time-aligned narratives in the second pilot study	707

Table 1: This table summarizes the data. Of the collected narratives, 707 are included in this work; audio is unavailable for some narratives.

vocalize their thoughts in rich detail. This teaching-oriented scenario really is a monologue, yet induces a feeling of dialogic interaction in the Master.

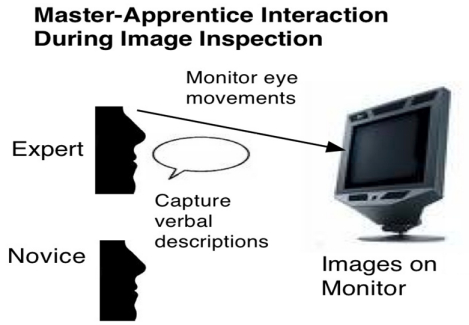

Figure 1: The Master-Apprentice interaction scenario allows us to extract information about the Master's (here: doctor's) cognitive processes.

The form of narratives collected can be analyzed in many ways. Figure 2 shows two narratives, recently elicited and similar to the ones in the study's data set, that are used here with permission as examples. In terms of diagnostic reasoning styles, referring to Pelaccia et al. (2011), we can propose that observer A may be using the *intuitive* system and that observer B may be using the *analytical* system. Observer A does not provide a differential diagnosis and jumps straight to his/her final diagnosis, which in this case is correct. We can postulate that observer A looks at the general area of the lesion and uses previous experience or heuristic knowledge to come to the correct diagnosis. This presumed use of the *intuitive* system could potentially relate to the depth of previous experience with a disease, for example. Observer B, on the other hand, might be using the

A. This patient has a pinkish papule with surrounding hypopigmentation in a field of other cherry hemagiomas and nevoid type lesions. The only diagnosis that comes to mind to me is Sutton's nevus.

B. I think I'm looking at an abdomen, possibly. I see a hypopigmented oval-shaped patch in the center of the image. I see that there are two brown macules as well. In the center of the hypopigmented oval patch there appears to be an area that may be a pink macule. Differential diagnosis includes halo nevus, melanoma, post-inflammatory hypopigmentation. I favor a diagnosis of maybe post-inflammatory hypopigmentation.

Figure 2: Two narratives collected in a recent elicitation setup and used here with permission. Narratives A and B are not part of the studied data set, but exemplify data set narratives which could not be distributed. Observers A and B are both looking at an image of a halo or Sutton's nevus as seen in Figure 3. Disfluencies are considered in the experimental work but have been removed for readability in these examples.

Figure 3: The image of a halo or Sutton's nevus viewed by the observers and the subject of example narratives.

analytical system. Observer B steps through the diagnosis in a methodical process and uses evidence presented to rationalize the choice of final diagnosis. Observer B also provides a differential diagnosis unlike observer A. This suggests that observer B is taking advantage of a process of elimination to decide on a final diagnosis.

Another way to evaluate these narratives is in terms of correctness and the related concept of diag-

nostic completeness. Whereas these newly elicited narrative examples have not been annotated by doctors, some observations can still be made. From the point of view of final diagnosis, observer A is correct, unlike observer B. Assessment of diagnostic correctness and completeness can also be made on intermediate steps in the diagnostic process (e.g. differential diagnoses or medical lesion morphological description). Including such steps in the diagnostic process is considered good practice. Observer A does not supply a differential diagnosis and instead skips to the final diagnosis. Observer B provides the correct answer in the differential diagnosis but gives the incorrect final diagnosis. Observer B fully describes the medical lesion morphology presented. Observer A, however, only describes the pink lesion and does not discuss the other two brown lesions.

The speech of the diagnostic narratives was recorded. At the same time, the observers' eye-movements were tracked; the eye-tracking data are considered in another report (Li et al., 2010). We leave the integration of the linguistic and eye-tracking data for future work.

After the collection of the raw audio data, the utterances were manually transcribed and time-aligned at the word level with the speech analysis tool Praat (Boersma, 2001).[1] A sample of the transcription process output is shown in Figure 4. Given our experimental context, off-the-shelf automatic speech recognizers could not transcribe the narratives to the desired quality and resources were not available to create our own automatic transcriber. Manual transcription also preserved disfluencies, which we believe convey meaningful information. Disfluencies were transcribed to include filled pauses (e.g. *uh, um*), false starts (e.g. *purreddish purple*), repetitions, and click sounds.

[1] See `http://www.fon.hum.uva.nl/praat/`.

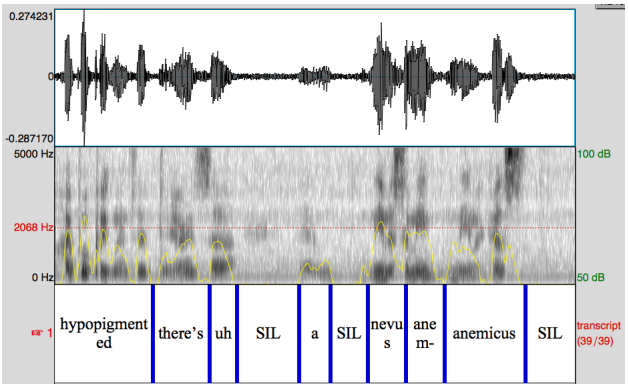

Figure 4: Transcripts were time-aligned in Praat which was also used to extract acoustic-prosodic features.

This study is strengthened by its involvement of medical experts. Trained dermatologists were recruited in the original elicitation experiment as well as the creation and application of both annotation schemes. This is crucial in a knowledge-rich domain such as medicine because the annotation scheme must reflect the domain knowledge. Another study reports on annotation details (McCoy et al., Forthcoming 2012).

4 Classification Study

This section discusses the classification work, first explaining the methodology for the initial pilot study followed by interpretation of results. Next, the methodology of the second pilot study is described.

4.1 Generic Model Overview

This work applies computational modeling designed to predict diagnostic correctness in physicians' narratives based on linguistic features from the acoustic-prosodic and lexical-structural modalities of language, shown in Table 2. Some tests discussed in 4.2 and 4.3 were performed with these modalities separated. These features are inspired by previous work conducted by Szarvas (2008), Szarvas et al. (2008), Litman et al. (2009), Liscombe et al. (2005), and Su et al. (2010).

We can formally express the created model in the following way: Let n_i be an instance in a set of narratives N, let j be a classification method, and let l_i be a label in a set of class labels L. We want to establish a function $f(n_i, j) : l_i$ where l_i is the label assigned to the narrative based on linguistic features from a set F, where $F = f_1, f_2, ...f_k$, as described in Table 2. The baseline for each classifier is defined as the majority class ratio. Using scripts in Praat (Boersma, 2001), Python, and NLTK (Bird et al., 2009), we automatically extracted features for each narrative. Each narrative was annotated with multiple labels relating to its diagnostic correctness. The labeling schemes used in the initial and second pilot studies, respectively, are described in subsec-

tions 4.2 and 4.3.

4.2 Initial Pilot Study

The initial pilot classification study allowed the opportunity to refine the prediction target annotation scheme, as well as to explore a preliminary set of linguistic features. 160 narratives were assigned labels of *correct* or *incorrect* for two steps of the diagnostic process: *diagnostic category* and *final diagnosis*. These annotations were done by a dermatologist who did not participate in the elicitation study (co-author Cara Calvelli). For final diagnosis, 70% were marked as correct, and for diagnostic category, 80% were marked as correct. An outcome of the annotation study was learning that the initial annotation scheme needed to be refined. For example, *diagnostic category* had a fuzzy interpretation, and correctness and completeness of diagnoses are found along a gradient in medicine. This led us to pursue an improved annotation scheme with new class labels in the second pilot study, as well as the adoption of a gradient scale of correctness.

Linguistic Modality	Feature at the narrative level
Acoustic-prosodic	Total duration
	Percent silence
	Time silent
	# of silences *
	Time speaking
	# of utterances *
	Initial silence length
	F0 mean (avg. pitch) ○
	F0 min (min. pitch) ○
	F0 max (max. pitch) ○
	dB mean (avg. intensity) ○
	dB max (max. intensity) ○
Lexical-structural	# of words
	words per minute
	# of disfluencies •
	# of certainty amplifiers * •
	# of speculative words * •
	# of stop words * •
	# of content words * •
	# of negations * •
	# of nouns •
	# of verbs •
	# of adjectives •
	# of adverbs •
	Unigram of tokens
	Bigram of tokens
	Trigram of tokens

Table 2: Features used by their respective modalities. Features marked with a * were only included in the second pilot study. Features marked with ○ were included twice; once as their raw value and again as a z-score normalized to its speaker's data in the training set. Features marked with • were also included twice; once as their raw count and again as their value divided by the total number of words in that narrative. Disfluencies were considered as words towards the total word count, silences were not. No feature selection was applied.

For the initial pilot study, basic features were extracted from the diagnostic narratives in two modalities: acoustic-prosodic and lexical-structural (see Table 2). To understand the fundamental aspects of the problem, the initial pilot study experimented with the linguistic modalities separately and together, using three foundational algorithms, as implemented in NLTK (Naive Bayes, Maximum Entropy, Decision Tree), and a maximum vote classifier based on majority consensus of the three basic classifiers. The majority class baselines were 70% for diagnosis and 80% for diagnostic category. The small pilot data set was split into an 80% training set and a 20% testing set. The following results were obtained with the maximum vote classifier.

Utilizing only acoustic-prosodic features, the maximum vote classifier performed 5% above the baseline when testing final diagnosis and 6% below it for diagnostic category. *F0 min* and *initial silence length* appeared as important features. This initial silence length could signal that the observers are able to glean more information from the image, and using this information, they can make a more accurate diagnosis.

Utilizing only lexical-structural features, the model performed near the baseline (+1%) for final diagnosis and 9% better than the baseline for diagnostic category. When combining acoustic-prosodic and lexical-structural modalities, the majority vote classifier performed above the baseline by 5% for final diagnosis and 9% for diagnostic category. We are cautious in our interpretation of these findings. For example, the small size of the data set and the

particulars of the data split may have guided the results, and the concept of diagnostic category turned out to be fuzzy and problematic. Nevertheless, the study helped us refine our approach for the second pilot study and redefine the annotation scheme.

4.3 Second Pilot Study

For the second pilot study, we hoped to gain further insight into primarily two questions: (1) How accurately do the tested models perform on three steps of the diagnostic process, and what might influence the performance? (2) In our study scenario, is a certain linguistic modality more important for the classification problem?

The annotation scheme was revised according to findings from the initial pilot study. These revisions were guided by dermatologist and co-author Cara Calvelli. The initial pilot study scheme only annotated for *diagnostic category* and *final diagnosis*. We realized that *diagnostic category* was too slippery of a concept, prone to misunderstanding, to be useful. Instead, we replaced it with two new and more explicit parts of the diagnostic process: medical lesion *morphology* and *differential diagnosis*.

For *final diagnosis*, the class label options of *correct* and *incorrect* could not characterize narratives in which observers had not provided a final diagnosis. Therefore, a third class label of *none* was added. New class labels were also created that corresponded to the diagnostic steps of medical lesion *morphology* and *differential diagnosis*. Medical lesion *morphology*, which is often descriptively complex, allowed the label options *correct*, *incorrect*, and *none*, as well as *correct but incomplete* to deal with correct but under-described medical morphologies. *Differential diagnosis* considered whether or not the final diagnosis appeared in the differential and thus involved the labels *yes*, *no*, and *no differential given*. Table 3 summarizes the refined annotation scheme.

The examples in Figure 2 above can now be analyzed according to the new annotation scheme. Observer A has a *final diagnosis* which should be labeled as *correct* but does not give a differential diagnosis, so the *differential diagnosis* label should be *no differential given*. Observer A also misses parts of the morphological description so the assigned *medical lesion morphology* would likely be *correct but incomplete*. Observer B provides what seems to be a full morphological description as well as lists the correct final diagnosis in the differential diagnosis, yet is incorrect regarding *final diagnosis*. This narrative's labels for medical lesion *morphology* and *differential diagnosis* would most likely be *correct* and *yes* respectively. Further refinements may turn out useful as the data set expands.

Diagnostic step	Possible labels	Count	Ratio
Medical Lesion *Morphology*	*Correct*	537	.83
	Incorrect	36	.06
	None Given	40	.06
	Incomplete	32	.05
Differential Diagnosis	*Yes*	167	.24
	No	101	.14
	No Differential	434	.62
Final Diagnosis	*Correct*	428	.62
	Incorrect	229	.33
	None Given	35	.05

Table 3: Labels for various steps of the diagnostic process as well as their count and ratios of the total narratives, after eliminating those with no annotator agreement. These labels are explained in section 4.3.

Three dermatologists annotated the narratives, assigning a label of correctness for each step in the diagnostic process for a given narrative. Table 3 shows the ratios of labels in the collected annotations. Medical lesion *morphology* is largely correct with only smaller ratios being assigned to other categories. Secondly, a large ratio of narratives were assigned *no differential given* but of those that did provide a differential diagnosis, the correct final diagnosis was more likely to be included than not. Regarding *final diagnosis*, a label of *correct* was most often assigned and few narratives did not provide any *final diagnosis*. These class imbalances, existing at each level, indicated that the smaller classes with fewer instances would be quite challenging for a computational classifier to learn.

Any narrative for which there was not agreement for at least 2 of the 3 dermatologists in a diagnostic step was discarded from the set of narratives considered in that diagnostic step.[2]

[2]Because narratives with disagreement were removed, the total numbers of narratives in the experiment sets differ slightly on the various step of the diagnostic process.

Comparing classification in terms of algorithms, diagnostic steps, and individual classes

Weka (Witten and Frank, 2005)[3] was used with four classification algorithms, which have a widely accepted use in computational linguistics.[4]

Standard performance measures were used to evaluate the classifiers. Both acoustic-prosodic and lexical-structural features were used in a leave-one-out cross-validation scenario, given the small size of the data set. The results are shown in Table 4. Accuracy is considered in relation to the majority class baseline in each case. With this in mind, the high accuracies found when testing medical lesion *morphology* are caused by a large class imbalance. *Differential diagnosis*' best result is 5% more accurate than its baseline while *final diagnosis* and medical lesion *morphology* are closer to their baselines.

	Final Dx	*Diff. Dx*	M. L. *M.*
Baseline	*.62*	*.62*	*.83*
C4.5	.57	.62	.77
SVM	.63	.67	.83
Naive Bayes	.55	.61	.51
Log Regression	.53	.64	.66

Table 4: Accuracy ratios of four algorithms (implemented in Weka) as well as diagnostic steps' majority class baselines. Experiments used algorithms' default parameters for *final diagnosis* (3 labels), *differential diagnosis* (3 labels), and medical lesion *morphology* (4 labels) using leave-one-out cross-validation.

In all scenarios, the SVM algorithm reached or exceeded the majority class baseline. For this reason, other experiments used SVM. The results for the SVM algorithm when considering precision and recall for each class label, at each diagnostic step, are shown in Table 5. Precision is calculated as the number of true positives for a given class divided by the number of narratives classified as the given class. Recall is calculated as the number of true positives for a given class divided by the number of narratives belonging to the given class. As Table 5 shows, and as expected, labels representing large proportions were better identified than labels representing intermediate proportions, and classes with few instances did poorly.

[3]See `http://www.cs.waikato.ac.nz/ml/weka/`.

[4]In this initial experimentation, not all features used were independent, although this is not ideal for some algorithms.

Dx step	Labels	Precision	Recall
Medical Lesion *Morphology*	*Correct*	.83	.99
	Incorrect	0	0
	None Given	0	0
	Incomplete	0	0
Differential Diagnosis	*Yes*	.49	.44
	No	.26	.10
	No Diff.	.76	.89
Final Diagnosis	*Correct*	.67	.84
	Incorrect	.32	.47
	None Given	0	0

Table 5: Precision and recall of class labels. These were obtained using the Weka SVM algorithm with default parameters using leave-one-out cross-validation. These correspond to the experiment for SVM in Table 4.

	Final Diagnosis	*Diff. Diagnosis*
Baseline	*.62*	*.62*
Lex.-struct.	.62	.67
Acous.-pros.	.65	.62
All	.63	.67

Table 6: Accuracy ratios for various modalities. Tests were performed for *final diagnosis* and *differential diagnosis* tags with Weka's SVM algorithm using a leave-out-out cross-validation method. Lexical-structural and acoustic-prosodic cases used *only* features in their respective set.

Experimentation with types of feature

To test if one linguistic modality was more important for classification, experiments were run in each of three different ways: with only lexical-structural features, with only acoustic-prosodic features, and with all features. We considered the *final diagnosis* and *differential diagnosis* scenarios. It was decided not to run this experiment in terms of medical lesion *morphology* because of its extreme class imbalance with a high baseline of 83%. Medical lesion *morphology* also differs in being a descriptive step unlike the other two which are more like conclusions. Again, a leave-one-out cross-validation method was used. The results are shown in Table 6.

These results show that, regarding *final diagnosis*, considering only acoustic-prosodic features seemed

to yield somewhat higher accuracy than when features were combined. This might reflect that, conceptually, *final diagnosis* captures a global end step in the decision-making process, and we extracted voice features at a global level (across the narrative). In the case of *differential diagnosis*, the lexical-structural features performed best, matching the accuracy of the combined feature set (5% over the majority class baseline). Future study could determine which individual features in these sets were most important.

Experiments with alternative label groupings for some diagnostic steps

Another set of experiments examined performance for adjusted label combinations. To learn more about the model, experiments were run in which selected classes were combined or only certain classes were considered. The class proportions thus changed due to the combinations and/or removal of classes. This was done utilizing all features, the Weka SVM algorithm, and a leave-one-out methodology. Only logically relevant tests that increased class balance are reported here.[5]

An experiment was run on the *differential diagnosis* step. The *no differential given* label was ignored to allow the binary classification of narratives that included differential diagnoses. The new majority class baseline for this test was 62% and this classification performed 1% over its baseline. A similar experiment was run on the *final diagnosis* diagnostic step. Class labels of *incorrect* and *none given* were combined to form binary set of class labels with a 62% baseline. This classification performed 6% over the baseline, i.e., slightly improved performance compared to the scenario with three class labels.

5 Conclusion

In these pilot studies, initial insight has been gained regarding the computational linguistic modeling of extra-propositional meaning but we acknowledge that these results need to be confirmed with new data.

This paper extracted features, which could possibly relate to uncertainty, at the global level of a narrative to classify correctness of three diagnostic reasoning steps. These steps are in essence local phenomena and a better understanding of how uncertainty is locally expressed in the diagnostic process is needed. Also, this work does not consider parametrization of algorithms or the role of feature selection. In future work, by considering only the features that are most important, a better understanding of linguistic expression in relation to diagnostic correctness could be achieved, and likely result in better performing models. One possible future adaptation would be the utilization of the Unified Medical Language System to improve the lexical features used Woods et al. (2006).

Other future work includes integrating eye movement data into prediction models. The gaze modality informs us as to where the observers were looking when they were verbalizing their diagnostic process. We can thus map the narratives to how gaze was positioned on an image. Behavioral indicators of doctors' diagnostic reasoning likely extend beyond language. By integrating gaze and linguistic information, much could be learned regarding perceptual and conceptual knowledge.

Through this study, we have moved towards understanding reasoning in medical narratives, and we have come one step closer to linking the spoken words of doctors to their cognitive processes. In a much more refined, future form, certainty or correctness detection could become useful to help understanding medical reasoning or help guide medical reasoning or detect misdiagnosis.

[5]Other experiments were run but are not reported because they have no use in future implementations.

Acknowledgements

This research supported by NIH 1 R21 LM010039-01A1, NSF IIS-0941452, RIT GCCIS Seed Funding, and RIT Research Computing (http://rc.rit.edu). We would like to thank Lowell A. Goldsmith, M.D. and the anonymous reviewers for their comments.

References

Cecilia Ovesdotter Alm. 2011. Subjective Natural Language Problems: Motivations, Applications, Characterizations, and Implications. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 107–112.

Hugh Beyer and Karen Holtzblatt. 1997. *Contextual Design: Defining Customer-Centered Systems*. Morgan Kaufmann.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media.

Paul Boersma. 2001. Praat, a system for doing phonetics by computer. *Glot International*, pages 341–345.

Herbert Clark and Jean Fox Tree. 2002. Using uh and um in spontaneous speaking. *Cognition*, pages 73–111.

Martin Corley and Oliver Stewart. 2008. Hesitation disfluencies in spontaneous speech: The meaning of um. *Language and Linguistics Compass*, 5(2):589–602.

Pat Croskerry. 2009. A universal model of diagnostic reasoning. *Academic Medicine*, pages 1022–1028.

Mark Graber. 2005. Diagnostic errors in medicine: A case of neglect. *The Joint Commission Journal on Quality and Patient Safety*, pages 106–113.

Aron Henriksson and Sumithra Velupillai. 2010. Levels of certainty in knowledge-intensive corpora: An initial annotation study. *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pages 41–45.

Rui Li, Preethi Vaidyanathan, Sai Mulpuru, Jeff Pelz, Pengcheng Shi, Cara Calvelli, and Anne Haake. 2010. Human-centric approaches to image understanding and retrieval. *Image Processing Workshop, Western New York*, pages 62–65.

Jackson Liscombe, Julia Hirschberg, and Jennifer Venditti. 2005. Detecting certainness in spoken tutorial dialogues. *Proceedings of Interspeech*, pages 1837–1840.

Diane Litman, Mihail Rotaru, and Greg Nicholas. 2009. Classifying turn-level uncertainty using word-level prosody. *Proceedings of Interspeech*, pages 2003–2006.

Wilson McCoy, Cecilia Ovesdotter Alm, Cara Calvelli, Rui Li, Jeff Pelz, Pengcheng Shi, and Anne Haake. Forthcoming-2012. Annotation schemes to encode domain knowledge in medical narratives. *Proceedings of the Sixth Linguistic Annotation Workshop*.

Sergey Pakhomov. 1999. Modeling filled pauses in medical dictations. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 619–624.

Thierry Pelaccia, Jacques Tardif, Emmanuel Triby, and Bernard Charlin. 2011. An analysis of clinical reasoning through a recent and comprehensive approach: the dual-process theory. *Medical Education Online*, 16:5890.

Heather Pon-Barry and Stuart Shieber. 2009. The importance of sub-utterance prosody in predicting level of certainty. *Proceedings of NAACL HLT*, pages 105–108.

Qi Su, Chu-Ren Huang, and Helen Kai-yun Chen. 2010. Evidentiality for text trustworthiness detection. *Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground ACL 2010*, pages 10–17.

Gyorgy Szarvas, Veronika Vincze, Richard Farkas, and Janos Csirik. 2008. The bioscope corpus: annotation for negation, uncertainty and their scope in biomedical texts. *BioNLP 2008: Current Trends in Biomedical Natural Language Processing*, pages 38–45.

Gyorgy Szarvas. 2008. Hedge classification in biomedical texts with a weakly supervised selection of keywords. *Proceedings of 46th Annual Meeting of the Association of Computational Linguistics*, pages 281–289.

Sumithra Velupillai. 2010. Towards a better understanding of uncertainties and speculations in Swedish clinical text - analysis of an initial annotation trial. *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pages 14–22.

Ian H. Witten and Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.

James Woods, Charles Sneiderman, Karam Hameed, Michael Ackerman, and Charlie Hatton. 2006. Using umls metathesaurus concepts to describe medical images: dermatology vocabulary. *Computers in Biology and Medicine 36*, pages 89–100.

Simon Zwarts and Mark Johnson. 2011. The impact of language models and loss functions on repair disfluency detection. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 703–711.