From Novice to Expert Narratives of Dermatological Disease

Nse Obot
University at Buffalo
Buffalo, NY
nseobot@buffalo.edu

Laura O'Malley
Washington & Jefferson College
Washington, PA
omalleylc@jay.washjeff.edu

Ifeoma Nwogu
Rochester Institute of Technology
Rochester, NY
ion@cs.rit.edu

Qi Yu
Rochester Institute of Technology
Rochester, NY
qi.yu@rit.edu

Wei Shi Shi
Rochester Institute of Technology
Rochester, NY
ws7586@rit.edu

Xuan Guo
Rochester Institute of Technology
Rochester, NY
xxg3358@g.rit.edu

***Abstract*—Medical diagnosis requires extensive knowledge of pathological characteristics and the clinical training necessary to master specific domains. In the domain of dermatology, these properties include the size, shape, distribution, color, and location of symptoms. We have built and trained a tag recommender with expert narratives of dermatological disease, then explored the effectiveness of this system on novice narratives. The system mitigates the prerequisite of domain knowledge, empowering a novice to enhance their medical descriptions or even reach an accurate diagnosis. After collecting novice narratives, we explored word alignment to provide a mapping between expert and novice vocabulary allowing novice input to be augmented with expert terminology. Ultimately, we found that our system is an effective educational tool, which could be improved by word alignment and other techniques.**

***Index Terms*—Expert systems, Support vector machines, Natural language processing, Graphical user interfaces, Recommender systems, Medical diagnosis, Dermatology, Semantics**

I. Introduction

Performing accurate and timely diagnostics requires years of arduous study, in which physicians obtain the necessary knowledge and skill to arrive at a diagnosis. Due to the nature of medical diagnosis, demand far exceeds supply, increasing the economic burden on both patients and doctors. Physicians must endure heavier workloads, reducing the quality of their service. Furthermore, patients face greater financial risk in seeking medical attention.

To alleviate a portion of this societal burden, our research aims to provide a platform to educate people without prior exposure to medical knowledge. Through automating a portion of what is typically a doctor's responsibility, providing medical attention to people will not only become more efficient, but also more affordable and accessible to more people.

Although many factors go into a diagnosis, we focused primarily on the language of physicians and novices and observed the relationship that these narratives share in reference to photos of specific diseases. We chose dermatological diseases to avoid other aspects of a diagnosis, such as smell or touch. Most of the necessary information for diagnosis can be captured in images and verbal or textual descriptions of those images thereby minimizing technical challenges.

With these physician descriptions of dermatological diseases, we used various machine learning and categorization techniques in combination with methods derived from python's NLTK to find commonalities between novice and physician narratives. These similarities can help us measure user learning as well as system performance in deciding whether this could be a feasible tool in educating novices about dermatological diagnosis.

In this paper, we review the foundational works that enabled the development of our tag recommender. Then, we explore the construction and experimental design for evaluation of the system. Finally, we analyze the results and consider possible improvements, as well as avenues for future endeavors.

We hope to answer two major questions in our study: Can our system accurately diagnose a dermatological disease based on a novice description? Do users learn from their interaction with the system?

II. Related Work

Guo et al. established metrics by which physician narratives about dermatological disease could be assessed. They recruited

16 physicians to diagnose and list the symptoms of 50 different images of dermatological diseases in hopes of progressing towards the automation of diagnosis. Guo et al. explored physicians' spoken narratives and eye tracking data when viewing an image of a disease, as well as what they revealed about the logic behind their process of determining an illness. Ultimately, the researchers found that physician dialogs held great insight regarding dermatological diagnosis. Our work builds upon those findings by investigating how a novice's low-level description can be used to accurately diagnose a dermatological disease through the relationship between the novices' and doctors' narratives [2].

Guo et al. built upon this work by presenting a human in the loop machine learning approach for grouping medical images. The study also explored expert narratives which discussed dermatological disease while gathering eye-tracking data. The results of that research show that this approach is more effective for improving the accuracy of a ML model without having to manually tune parameters [3].

These works utilized various analysis techniques, one of which is word embeddings. Word embeddings are vector representations of words that facilitate semantic analysis. Kenter's study was insightful since we must find commonalities between physician and novice narratives regardless of the differences in vocabulary. Word embeddings can help bridge this terminology gap by providing vector translations of words in order to find semantic similarities between expert and novice descriptions. They concluded that the conversion of words to vectors yielded optimal results in finding similarities between texts [4].

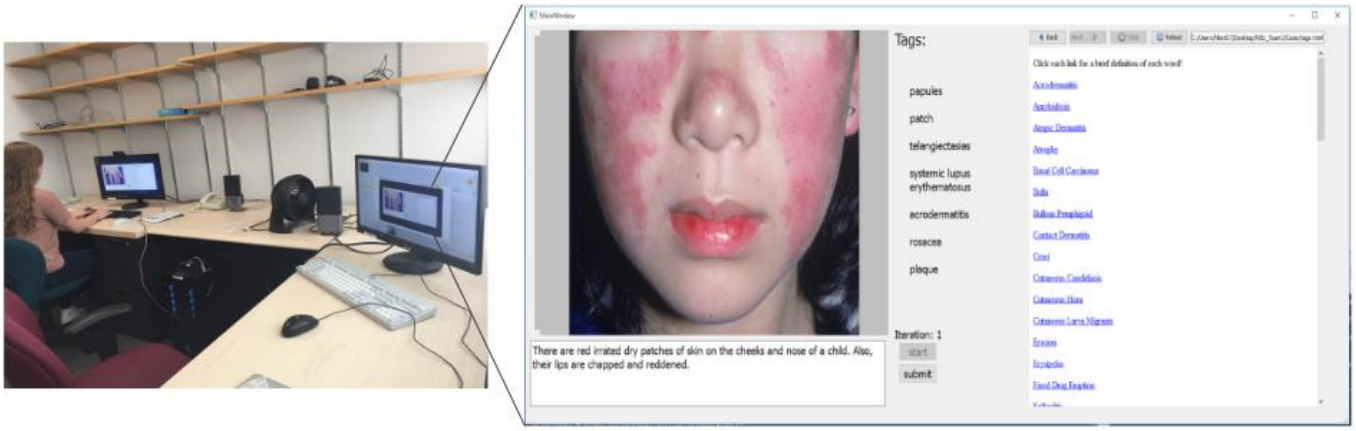
Another analysis technique was Gaussian Linear Discriminant Analysis (LDA), which provided a method for extracting the topics of documents even when those documents contain words unknown to that algorithm. This LDA based algorithm, which proved to be more optimal than other methods of topic modeling, is an example of data resolution. It groups terms into topics thereby reducing the number of units that must be processed to represent various documents. For example, instead of a thousand-word vector, a list of one hundred topics could represent a corpus [1].

Our tag recommender system directly builds upon these works. It uses the expert narratives recorded by Guo et al. as the training data set for the ML model, and integrates novice narratives with more analysis techniques. We also explore word alignment to improve the system. Word alignment has been shown to be effective for translating between languages or from the vernacular of one narrator to another regardless of word order. This makes it an ideal candidate for translating between novice and expert narratives [5].

III. Experimental Design

Xuan Guo et al. provided a set of 800 expert narratives from 16 physicians based on descriptions of 50 distinct images of dermatological diseases. Each narrative contains primary and secondary morphologies, which are skin lesions or symptoms that help in identifying certain diseases. Our first task was to locate and extract these terms from each narrative in order to find which terms corresponded to which disease.

Once these terms were extracted, we used stemming methods from the NLTK to combine duplicate words. For example, the word papule is a primary/secondary morphology found by our system, where they are well-circumscribed, elevated, solid lesions, less than one cm high. However, our system identified papule and papules as two separate terms, therefore these terms are combined. We continued this process of combining terminology and then we used a voting process to determine the morphologies of certain diseases. In the voting process, if these morphologies are seen in the majority of the physician narratives for that specific disease, we choose that as an identifying term for the disease. Then, we trained a binary classifier for each of the most relevant terms. Furthermore, we created a multi-label classifier that uses SVM based



Tags:

papules

patch

telangiectasias

systemic lupus erythematosus

acrodermatitis

rosacea

plaque

There are red irrated dry patches of skin on the cheeks and nose of a child. Also, their lips are chapped and reddened.

Click each link for a brief definition of each word!

Acrodermatitis

[illegible]

Atopic Dermatitis

Atrophy

Basal Cell Carcinoma

Bulla

Bullous Pemphigoid

Contact Dermatitis

Crust

Cutaneous Candidiasis

Cutaneous Horn

Cutaneous Larva Migrans

Erosion

Erysipelas

Fixed Drug Eruption

Figure 1: The UI presented to the user as well as the testing environment. Here, you see a sample image, an example description of the image, and the tags returned to the user based on this description. The user will navigate through the browser to be linked to photos and information about these diseases and morphologies.

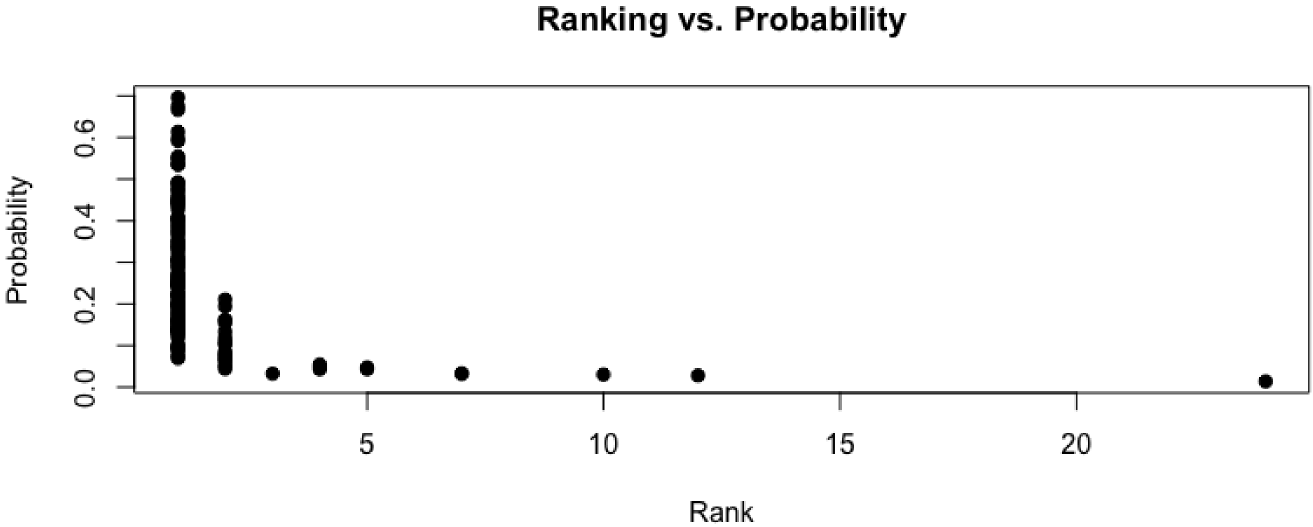


Figure 2: The ranking and probability of each of the test-expert narrative inputs.

classification in order to make predictions. This classifier offers probabilities for each disease when given a description of a photo, and returns the diseases in order of the highest probability that pertains to that description. To provide more information to the user, we expanded the model by returning the top five disease names from the multi-label classifier.

After creating the model, we designed a UI that presents photos of skin diseases and allows subjects to type and submit descriptions of the picture. Once they submit this description, the model returns tags that the UI displays to the participant. The user then learns from these tags by clicking associated links in the UI's web browser that contain a brief definition and photo of the disease. The user modifies their definition through incorporating terminology from these definitions and tags to use in their own description. Participants iterate this process of reading links and changing their definition until they feel they can no longer improve their description based on the terminology they have read. Subjects repeated this process for six photos.

We collected data from 25 participants, 17 being male and 8 being female between the ages of 18 and 34. While the subject interacted with the UI, a csv file was constructed that contained their descriptions, tags returned, photo IDs of the diseases being displayed to the user, and the probabilities of the 50 disease classes.

IV. Data Analysis

Upon the finalization of our data collection, we decided to measure system performance and novice improvement through first observing the ground truth disease probabilities at each iteration. However, we found that probability was not the best indicator of learning.

When diseases are returned to the subject, our multi-label classifier decides which diseases are chosen based on ground truth probabilities. These probabilities measure the likelihood of a disease class's relevance to the novice description by creating a distribution among the 50 classes. This distribution is measured based on a scale of one, where the probabilities of all 50 disease classes would sum to one. We chose a threshold of 50%, where if a disease's probability of being the described disease is at least 50%, it is the only disease returned along with additional primary/secondary morphologies. Otherwise, we return up to five diseases, or until their totaled percentages meets our threshold. There are many factors behind why this could occur, a few of them being if a description is vague or if the language used simply shares no commonalities with the expert narratives in the training data. In this scenario, the probabilities will be evenly distributed, allowing for disease classes with extremely low probabilities to be highly ranked, which could interfere with measuring improvement.

To confirm our suspicions, we randomly chose ten percent of the expert narratives as testing data while the other ninety percent was used as training data. Our system determined the ground truth probabilities for each test narrative and from those probabilities we were able to extract the ranking.

The results displayed that even when the ground truth class of an expert narrative was highly ranked, it was still possible to have its ground truth probability be extremely low. Due to this variance in probability using the expert narratives, we determined evaluating the ground truth ranking rather than the ground truth probability would better reflect a novice's improvement/learning.

We measured the rank by averaging the ground truth ranking of the initial novice narratives (initial meaning the subject's first description of a photo), and comparing that to the average ground truth ranking of the final novice narratives (final meaning the subject's description at the last iteration of a photo). After N/A's were removed, we found the ranking generally increased, but only by a small amount.

Table I : Initial & Final Ground Truth Ranking

Photo Set	Initial Ground Truth Ranking	Final Ground Truth Ranking
All Photos	11.340	10.850
Disease 1	9.217	7.217
Disease 2	1.000	1.000
Disease 3	12.450	10.180
Disease 4	27.410	25.550
Disease 5	4.273	7.227
Disease 6	12.300	12.650

According to the table above, half of the diseases experienced an increase in their ranking, two of the diseases grew further from being highly ranked, and disease 2 experienced neither an increase nor decrease in ranking.

Using cosine similarity, we sought to measure learning in novice descriptions. The cosine similarity was found by using a tfidf vectorizer from scikit-learn on each of the narratives to create a matrix of the phrase, and then using the cosine similarity function on the matrices created. We first found the cosine similarities between the initial novice narrative to every novice narrative at every iteration broken down on a photo and participant level. The results displayed the cosine similarities decreasing with each iteration, proving the novice made an effort to edit their description with each iteration.

To test whether novice language became closer to the expert descriptions, we also measured the average cosine similarity between every initial novice narrative and the combined expert narratives, then compared those values to the cosine similarities between the final novice narratives and the expert narratives. For each of these narrative sets, they were combined over all photos and all participants and we saw a general increase in similarity. The increase is small, however proportional to the improvement in ranking.

Table II : Cosine Similarities

Photo Set	Cos Similarity for Initial and Expert	Cos Similarity for Final and Expert
All Photos	0.16050	0.16920
Disease 1	0.16250	0.19340
Disease 2	0.11970	0.12641
Disease 3	0.23061	0.25295
Disease 4	0.13065	0.15201
Disease 5	0.15501	0.13680
Disease 6	0.16305	0.16967

The values in the above table likewise demonstrate a correlation with the ranking, where diseases that improve in rank also improve in cosine similarity and there is a decrease in both cosine similarity and ranking for disease 5. Also, diseases 2 and 6 experience little improvement in cosine similarity and remain about the same rank.

In addition to evaluating if users learned from their interactions with the system, we also sought for ways to improve the system. In particular, we strove to bridge the gap between novice and expert descriptions by automatically incorporating physicians' terminology into novice descriptions. Tantamount to investigating novice learning, the reliability of our system with expert narratives compelled us to explore the effect of augmenting novice narratives to utilize more medical terminology.

Word alignment's effectiveness in translating between languages prompted us to explore its potential for improving our system. Since our system is trained from expert narratives, modifying the novice input to incorporate more medical terminology could enhance the quality of the initial tags returned. Consequentially, the novice would more likely be directed towards the correct diagnosis. The aligner can create such a mapping between the novice and expert vocabulary. Using this mapping, input from a novice description could be augmented to directly utilize the physicians terminology as input to the recommender system.

Mimicking the French to English examples available by default from the Berkeley Aligner, we structured a parallel corpus from our data. In a parallel corpus, sentences are arranged in pairs such that equivalent sentences are adjacent. For example, in the French to English corpus each French sentence is adjacent to an equivalent English sentence. Unlike the French and English parallel corpus, semantic equivalence and a close ordering of the ideas of adjacent sentences were not guaranteed between expert and novice narratives. Therefore, we defined equivalence in our corpus to mean any two sentences that were created to describe the same image. Instead of French, we used expert narratives, and substituted the novice narratives for the English.

After running the aligner on our corpus, we investigated methods for augmenting novice narratives with expert terminology. Using the aligner, we created a mapping between the novice and expert terminology, by selecting the most probable translations between the novice and expert words. Then, for each novice narrative, we replaced each word in the narrative with its expert mapping and recorded the ground truth ranking. As evident in the example of brightly, some of the physicians' words had a semantically reasonable word grouping. However, this was not the case overall. Furthermore, even if most of the novice words associated with a particular medical term were semantically reasonable, this did not guarantee that the novice word would be replaced with a semantically plausible medical term. For example, the probability list for periungual contains nails, fingernail, and finger, but the first word in the list is near. In addition, medical terminology is also more advanced in that multiple novice terms could be combined into a single medical term. For example, the novice phrase around the fingernails could be interpreted by an expert as periungual.

Table III : *brightly* Translation Probabilities

Words associated with *brightly*	Translation Probability
discoloration	0.23
patches	0.20
lighter	0.18
color	0.13
patterns	0.12
streaks	0.09
light	0.03

These complications in augmenting novice narratives with physicians words led us to broaden our approach. Using standard NLP techniques such as stemming, lemmatizing, and removing stop-words resulted in similar ground truth rankings. In addition to these techniques, we also incorporated a union list of words, both in the novice and physician narratives, into the training set for the aligner. We integrated this list by appending it to both the novice and expert sections of the parallel corpus, thereby having single words paired with themselves. Prior work with the aligner indicated that using a union list would decrease the probability of a word not having itself as the most probable translation. Furthermore, this improves the translation probability list, of other words, by essentially eliminating all of the union words as potential misalignments. Despite these approaches, the ground truth rankings did not improve. So, we decided to hand translate a small subset of novice narratives with our knowledge of the images and expert vocabulary. Training the aligner with these hand translations did not improve the ground truth rankings overall. However, we noticed significant improvement in the ground truth ranking of disease 4 from 27.41 to 16.96. Ground truth rankings from training the aligner with each corpus are in the table below.

Table IV : Ground Truth Rankings from Corpuses

Photo Set	Ground Truth Ranking (Hand Translated Corpus)	Ground Truth Ranking (Novice to Expert Corpus)
All Photos	21.82	23.13
Disease 1	18.30	21.48
Disease 2	24.04	26.83
Disease 3	28.29	21.21
Disease 4	16.96	22.38
Disease 5	17.87	24.52
Disease 6	25.21	22.5

V. Discussion

Our experiment yielded results displaying improvement in novice knowledge, however we realize we could have obtained superior data if our training set was much larger. The small corpus made it difficult to map expert descriptions to novice terminology when using the Berkeley Aligner. For comparison, the French to English corpus had approximately 3,000 sentences pairs. Yet, there were 141 narratives in the first iteration of novice descriptions and only 96 expert narratives for the six images that participants described. Furthermore, just 20 novice narratives were used for the hand translated corpus. However, obtaining novice and expert descriptions can be expensive as well as time consuming.

One alternative to modifying the system is to revise our experimental design. For example, instead of recruiting participants locally we could deploy the system online, enabling us to gather more data from a broader area. To mitigate expenses, we could utilize a system like Amazon's Mechanical Turk. The Turkers could provide inexpensive descriptions, or even hand translations of novice descriptions given a dictionary of expert terminology. For now, modifications to our system are necessary to avoid the issues that come with small training and testing sets.

Another modification to our system would be possibly considering an entirely different approach in classifying diseases where instead of depending on a novice for raw input, we would apply a question and answer approach. First, we would identify features specific to each disease in order to group them, possibly referring to the image grouping study[3] to find these differentiating features, and use this to come to a diagnosis. The initial question begins with every disease having an equal chance of being classified, but with each question we could narrow down the choices to ultimately come to a diagnosis. We would still incorporate tags in our questions and have a browser window containing the links to brief descriptions and photos of the terms. This will allow for user learning and instead of assessing their written narratives, we could use sensors to measure cognitive load and a short quiz following completion of the study. Less subjectivity would come with having more control over user input and less complications would arise even when it is a more difficult photo to characterize.

Overall, our system was able to measure slight user improvement, as well as lead novices towards the correct diagnosis. Subjects also displayed learning through the ground truth rankings, cosine similarities, and self-reported learning. Although our system did not exhibit as considerable of an improvement as we hoped, we have nonetheless spotted notable correlations in our data which indicate improvement and know which adaptations need to be made to our system to yield the best results.

VI. Conclusion

We built upon the foundational works of this study through not only furthering the effort to automate certain aspects of diagnosis, but giving novices the potential to come to a diagnosis. Our work provides a unique contribution by finding similarities between physician and novice language to determine a disease. Additionally, no work has been done in automating the process of teaching novices about skin related illness specifically.

Previous research has investigated automating image grouping of diseases in order to identify problem areas of skin disorders efficiently. We would not have been able to perform our study if those same researchers had not found that expert narratives pertaining to skin diseases held great insight in what determines a dermatological diagnosis[2]. We used this discovery in order to find if novice narratives could be utilized in a system to accurately identify a disease and its morphologies, as well as serve as an educational tool.

Given our time constraints, we were also unable to apply word2vec in our analysis or back-end of our recommender. Word2vec would have given us a more dramatic increase in cosine similarity since this method is able to equate words such as queen and woman, and therefore would liken a term most frequently used by experts such as annular to the novice term circular. In a future application of this system, utilizing this method would offer a measure of semantic similarity rather than sentence/word structure similarities amongst the novice and expert narratives. This could possibly allow our system to provide more accurate tags to the user in eventually reaching the proper diagnosis.

In the future, we hope to expand on this work by exploring other semantic tools such as topic modeling for analysis and other statistical translation systems such as Moses. Semantic analysis is essential to understand the extent to which meaning is preserved between image descriptions. Although cosine similarity between word vectors provided a sense of semantic similarity, many details pertaining to the relationship between the meaning of the novice narratives and the physicians' narratives are obscured. It would be interesting to see what topics are shared between novice and expert narratives and generate a semantic structure from these topics. It may be the case that the more advanced topics of the expert narratives would become subclasses of the more general topics of the novice narratives. However, this it yet to be explored. Similarly, the Berkeley Aligner is designed specifically for word alignment, not necessarily translation. Therefore, the semantics between descriptions may be better preserved by a system that is designed for translation. Thus. to improve our system, it may be beneficial to explore machine translation systems.

Our work, both present and future, aims to alleviate the societal burden incurred by the nature of diagnosis. Therefore, we have considered many possible implementations for our work. Among them is online diagnosis where our system is made available through a website. This enables anyone with internet access to efficiently reach a dermatological diagnosis without incurring a financial burden through medical expenses. Another application is to have patients use our system while waiting for a doctor in a hospital. This way, regardless of the accuracy of the diagnosis, the patient will be more informed about dermatological disease, and will likely have better ways of describing their condition to a doctor. With the patient better educated about skin disease, physicians may reach a diagnosis sooner, enabling them to provide care to more patients. Finally, with a brief modification to our system, medical students could use it for training. Unlike our experiment, which did not indicate to the user the ground truth ranking, medical students could be shown these rankings. As they interact with the system, they could strive to improve the ground truth ranking and know when their description sufficiently arrives at an expert level. Whether a globally available system, a tool to assist doctors, or a training tool for doctors-to-be, our system has the potential to improve the process of accurate medical diagnosis, especially in the domain of dermatology.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Award No. IIS-1559889. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] R. Das, M. Zaheer, and C. Dyer. “Gaussian lda for topic models with word embeddings,” In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 795–804, Beijing, China, July 2015. Association for Computational Linguistics.

[2] X. Guo, Q. Yu, C. O. Alm, C. Calvelli, J. B. Pelz, P. Shi, and A. R. Haake. “From spoken narratives to domain knowledge: Mining linguistic data for medical image understanding,” Artificial intelligence in medicine, 62 2:79–90, 2014.

[3] X. Guo, Q. Yu, R. Li, C. O. Alm, C. Calvelli, P. Shi, and A. Haake. “An Expert-in-the-loop Paradigm for Learning Medical Image Grouping,” pp. 477–488. Springer International Publishing, Cham, 2016.

[4] T. Kenter and M. de Rijke. “Short text similarity with word embeddings,” In Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, CIKM ’15, pp. 1411–1420, New York, NY, USA, 2015. ACM.

[5] E. Prud’hommeaux and B. Roark. “Graph based word alignment for clinical language evaluation,” Comput. Linguist., 41(4):549–578, December 2015.

[6] Bird, Steven, Ewan Klein, and Edward Loper (2009), Natural Language Processing with Python, O'Reilly Media.

[7] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.