Circumlocution in Diagnostic Medical Queries

Isabelle Stanton[*]
UC Berkeley
Berkeley, CA
isabelle@eecs.berkeley.edu

Samuel Ieong[†]
Microsoft Research
Mountain View, CA
sieong@microsoft.com

Nina Mishra
Microsoft Research
Mountain View, CA
ninam@microsoft.com

ABSTRACT

Circumlocution is when many words are used to describe what could be said with fewer, e.g., "a machine that takes moisture out of the air" instead of "dehumidifier". Web search is a perfect backdrop for circumlocution where people struggle to name what they seek. In some domains, not knowing the correct term can have a significant impact on the search results that are retrieved. We study the medical domain, where professional medical terms are not commonly known and where the consequence of not knowing the correct term can impact the accuracy of surfaced information, as well as escalation of anxiety, and ultimately the medical care sought. Given a free-form colloquial health search query, our objective is to find the underlying professional medical term. The problem is complicated by the fact that people issue quite varied queries to describe what they have. Machine-learning algorithms can be brought to bear on the problem, but there are two key complexities: creating high-quality training data and identifying predictive features. To our knowledge, no prior work has been able to crack this important problem due to the lack of training data. We give novel solutions and demonstrate their efficacy via extensive experiments, greatly improving over the prior art.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

1. INTRODUCTION

The act of "talking around" a term is known formally as circumlocution and happens when people do not know the correct term to describe what they seek. This is common among speakers learning a new language or entering an unfamiliar domain, e.g., in law [can't be tried twice for the same crime] for double jeopardy and [what is that pink liquid coming out my car] for transmission fluid leak.

In this paper, we study circumlocution in the medical domain. Accurate understanding in this domain is particularly important since the wrong information can have a large effect on people's actions. Inaccurate search results can lead to escalation of anxiety, as illustrated by work on cybercondria [34]. For example, a search for a common symptom such as [shoulder twitch] escalated to searches for [ALS symptoms] (also known as Lou Gehrig's disease), a serious disease that afflicts a tiny fraction of the population – 1-2 new cases a year per 100,000 people. Wrong self-diagnosis can lead to incorrect follow-on pursuit of both health information and consultation of a medical professional (or lack thereof).

Search engines are routinely used for medical self-diagnosis. The Pew Research Center [14] found that over 35% of respondents had searched online for medical information specifically to determine what medical condition they or someone else might have. Further, 8 in 10 of these self-diagnostic sessions start at a search engine and 46% of them lead the user to seek professional medical attention. Despite the prominent use of the web for self-diagnosis, studies on developing a *Consumer Health Vocabulary (CHV)* [39] show that most people are not familiar with professional medical terms, and particularly not for symptoms that are rarer than cough, fever or rash.

Given a freeform, unstructured medical circumlocutory query, our goal is to find its corresponding professional medical term, e.g., from [fluid accumulation around eye] to [periorbital edema]. The task is complicated by the fact that users issue varied queries, from the banal, to detailed, to vague, to describe their symptoms, e.g., [my 2 yo has been coughing since Thursday fever since Saturday] or [my shin is tight and shiny]. Many irrelevant terms are included in the search. Any solution must learn to ignore terms that are not relevant for determining the symptom. Our solutions heavily leverage structured sources of medical data to steer us towards terms that are more likely to matter.

Most prior work on information retrieval involves situations where a user enters the correct term. To our knowledge, little is known about what to do when the searcher lacks the vocabulary/knowledge to express what they desire. One exception in the medical domain is the work on DiaTM [36]. However, the goal of this work is to discover hidden topics, i.e., is unsupervised. Consequently, hidden topics need not correspond to symptoms, as their results suggest [36].

[*] Work performed while an intern at Microsoft Research. This author is now at Google Inc.

[†] This author is now at Google Inc.

SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.
ACM 978-1-4503-2257-7/14/07.:.$15.00
http://dx.doi.org/10.1145/2600428.2609589 .

One key barrier to supervised solutions is a lack of training data. In fact, one reason prior approaches were unsupervised was because no training data was available. Given a medical symptom, how can one find colloquial variants? We provide an innovative solution that involves a reverse use of crowdsourcing with carefully selected images and videos.

The goal of this work is to understand the concept that a user has in mind. We would like to improve our understanding of the query in order to enable future assistive technologies beyond simply returning web results. We do not envision mapping a query such as [headache] to [cephalgia] and returning medical journal articles on cephalgia as we do not believe this will often help the user. Rather, we do envision mapping queries to concepts, as well as articles to concepts so as to improve the search engines ability to provide user-friendly, authoritative content.

Contributions

Our work has a key insight that enables a different solution to this difficult and important problem – it involves a new way of generating training data. The customary way to generate training data is to give a human judge a search query and ask them to identify the corresponding medical symptom. However, a typical human judge may not be able to name the medical symptom, either because the medical term is unfamiliar or because they lack the needed medical background to determine the symptom. Instead, we reverse the task by fixing the symptom and asking the judge to tell us what search queries they would issue to determine what was wrong. We prompt the judges with carefully selected multimedia content. As mentioned, the best prior solutions are unsupervised [36] and may be due in part to the difficulty of finding good training data. We demonstrate that queries generated are of high quality: 94% are good circumlocutions of the symptoms for images and 97% for videos. To determine how difficult the problem is, we posed ourselves the challenge of trying to identify the medical symptom name from a colloquial query. Despite our prior knowledge of the symptoms, we only labeled 56% of colloquial queries correctly. This low success rate demonstrates that the task is very challenging.

There are a large variety of knowledge sources readily available that we can bring to bear on the circumlocution problem. Greek and Latin roots enable us to better automatically understand a medical symptom name, as do encyclopedias and medical dictionaries. Body parts, colors, synonyms and paraphrases are also valuable as ways to obtain alternate expressions. We show how these publicly available knowledge sources can be suitably combined to solve the circumlocution problem. We represent the similarity of a query to a symptom by a collection of features that reflect how close the query or (knowledge-source expansions of the query) match the symptom (or knowledge-source expansions of the symptom).

We report on a collection of machine-learning experiments using this reverse crowdsourced labeled training data and features derived from these knowledge sources. We begin by training and testing on the same set of symptoms. We train a baseline bag of words multi-class classifier and compare to our approach of training with knowledge sources. Our approach is a 33% (26%) improvement over the baseline for images (respectively, videos). Next, we generalize to training and testing on a different set of symptoms. This question is crucial because we can not generate training data for an arbitrary symptom. Thus, bag of words symptom-dependent approaches will not succeed. We translate the circumlocution problem to a two-class learning problem and then show how one can learn the relative weight of combined knowledge sources. Our results improve over a baseline by 20% (26%) for images (respectively, videos). Our solution allows us to learn from one set of symptoms and to predict on an unseen set of symptoms.

2. RELATED WORK

Consumer health search has received much attention in both the medical and the information retrieval communities. One main challenge is that of the language gap—the document that can answer the user's question is written in medical terminology which the user is unfamiliar with, and not knowing the keywords to search for, the user was unable to learn the terminology from the document to aid her search in the first place. We now review work in this area.

The Consumer Health Vocabulary (CHV) is an important research project in the medical community to establish a vocabulary for communicating medical information to the public [39]. The vocabulary is selected based on perceived familiarity of the general public with certain medical terms. Many approaches have been used for estimating familiarity, including surveys [21], predictive modeling [41], query log analysis [42], and text mining [40]. While this work helps with the choice of terminology medical professionals can use to communicate to the lay person, it does not directly help with the task of understanding a lay person's query.

More directly related to closing the language gap is HIQuA, a query assistant that suggests and expands medical queries [38]. Given a query, the assistant provides a tree of facets to explore. For example, the query [skin] has refinement suggestions such as [itching], [psoriasis], and [atopic dermatitis], or 'Symptoms' and 'Treatment'. In evaluation, while users liked the suggestions made by the assistant, they were not any more successful than the control group of users without HIQuA at completing informational tasks.

MetaMap [4] is a linguistic system for mapping natural language text to the UMLS Metathesaurus developed at the National Library of Medicine. As MetaMap is widely available, we compared with our system but found that it is not tuned for understanding circumlocutions, and maps individual words ignoring context. Our experiments provide further explanation.

Another system that addresses the language gap problem is DiaTM [36]. The system extends Latent Dirichlet Allocation with a 'technicality' feature to improve retrieval for documents written in different vocabularies. The technicality feature is interpreted as a dialect and can be used to capture the notion of colloquial, lay, and professional terminology. However, as the method is based on unsupervised learning, there is no guarantee that the learned topics align to a given set of definitions (see Figure 2 in [36]). Even a perfect clustering solution would not solve the problem we are after. Ultimately, we seek to map queries to a specific symptom name – as well as documents to a symptom name – so as to better facilitate the information retrieval process.

There are also vertical search engines specialized for medical domains. MedSearch [23] focuses on better handling long, natural language queries through improved extraction of important keywords, diversifying results, and generating related medical phrases for assistance in query refinement. However, building a vertical search engine does not address the fundamental mismatch between the language used on reputable medical websites and how users search for information about symptoms and diseases they can not name.

If one is to address the language gap problem via a supervised learning approach, one needs training data. For related problems such as query suggestions, spell correction, and synonym detection, a popular approach for obtaining training data is using query refinements [26, 5, 9, 24, 33]. This is commonly done by examining sequences of queries and using co-occurring pairs for labels. Unfortunately, for our problem, the queries we are interested in labeling tend to be rare, as few users know the professional terms. As a result, the training data created from refinements often lead to the professional terms being associated with other unrelated professional terms, rather than the colloquial ones.

Finally, the problem of mapping colloquial queries to professional medical terms is a special case of the textual entailment problem in natural language processing – given a text snippet t and a hypothesis h, decide weather t entails h. A version of the problem is to decide whether one piece of text paraphrases another [35, 22, 11]. We use paraphrasing tools [3] in our solution for circumlocution.

3. PRELIMINARIES

We define some terms used in our paper, particularly the difference between a colloquial and circumlocutory query. We also describe the three problems we set out to solve: how to generate training data, how to represent the similarity between a query and a symptom and how to predict the symptom from a query.

Definitions

Throughout the text, a *medical symptom* is evidence of disease or physical disturbance observed by the patient. A *medical sign* is an objective indication that can be observed by someone other than the patient. For example, pain is a symptom observable only by the patient, while flushed cheeks are a sign observable by a physician or concerned parent. Signs and symptoms have *professional* medical names, such as cephalalgia, and *colloquial* names, such as headache. Signs and symptoms can also be discussed in a *circumlocutory* fashion, where neither the professional medical name nor a common colloquial name is used. Instead, a description is given such as "my head is pounding". Throughout the paper the phrase *colloquial language or query* will refer to a set of queries that includes the colloquial names, while *circumlocutory* refers to only queries that describe a sign or symptom.

Problem Statement

Our goal is to map circumlocutory language to professional medical language. Medical search is a large domain, so we focus only on the problem of medical signs and symptoms. We consider a few key subproblems.

PROBLEM 1. *Given a medical sign or symptom s, find colloquial and circumlocutory ways of expressing s.*

Previous work approached this problem from an unsupervised perspective. This is because finding labeled training data is a non-trivial problem. We believe that it is one of the major barriers to significant progress. Given training data, the next question is how to represent the similarity of a symptom and a circumlocutory expression. Answering this question forms the backbone for feature construction crucial to a machine learned solution.

PROBLEM 2. *Given a medical sign or symptom s and a colloquial or circumlocutory expression c, identify ways to automatically represent the similarity of s to c.*

Our ultimate goal is to map user queries to the correct medical terminology. With the correct terms, we can facilitate new information retrieval applications. Note that this problem primarily applies to circumlocutory language since the colloquial names for queries are well known.

PROBLEM 3. *Given a circumlocutory expression c of a symptom s, design a way to automatically infer s from c.*

There are several follow-up problems including how to combine symptoms to form a diagnosis. This is not our focus. Automated medical diagnosis systems have been previously studied e.g., [17, 12, 18, 13], and are an interesting and challenging research area. We focus on a phase that precedes diagnosis. A good solution to our problem can facilitate the longer term vision of automated diagnosis.

Search Activity

A common first response for how to identify circumlocutory expressions and their corresponding medical symptom label (Problems 1 and 3) is to mine search activity. We could not find a way to use the search logs, though it may be possible. One commonly used method in understanding search queries is to identify frequently co-occurring circumlocutions that precede a search for a medical symptom within a session. However, this approach may have limitations. For example, the nature of our queries is such that many words describe what could be said with fewer. This implies that circumlocutory searches can be long, and therefore possibly rare – complicating a frequency calculation. Further, since people can be unfamiliar with professional medical terminology [42] when they find content that may answer their diagnostic questions, they may issue symptom search queries to understand the definition, but not actually possess the symptom. Search behavior like this is supported by observations of Cartwright et. al [7] who found that medical search sessions alternate between two phases, *evidence-based* and *hypothesis-directed*, where evidence-based queries are of the form [back pain] or [dry cough with chest pain] while hypothesis-directed queries test diagnoses and treatments like [ointments for eczema] and [back stretching exercises]. This suggests that the search logs may be quite noisy for this task. We leave the question of how the search logs can be used in the context of understanding circumlocutory queries as a direction for future work.

4. GENERATING TRAINING EXAMPLES VIA CROWDSOURCING

We now show how given the correct priming and experience anyone can generate the training data that we require. With the advent of crowdsourcing platforms such as Amazon Mechanical Turk, Taskcn, Topcoder, 99designs, Innocentive, CrowdCloud, CrowdFlower, it has never been easier or more cost effective to obtain labeled data. A very simple task to crowdsource is obtaining and verifying labels. How to do this accurately and effectively for machine learning tasks has been extensively studied [31, 16, 27, 29, 20, 32, 8, 37]. While search engines often use human rated feedback, it is quite difficult for human raters to infer intent after the fact [1]. Moreover, considering the average layperson is unable to self-diagnose themselves, it is hopeless to ask crowdsourcing participants to label a set of queries with the symptoms that they describe.

Instead of asking the crowdsourcing participants to generate the labels for the training data, we ask them to solve the inverse problem. Given a label, generate the data. This is possible in our setting because the participants are a good approximation of the average user of a search engine, i.e. the population generating the circumlocutory queries that we are trying to classify. The challenge lies in finding the appropriate material to allow a crowdsourcing participant to create appropriate queries. This approach dramatically reduces the error in the labels.

We can not show the crowdsourcing participant a label such as *scleral icterus* and ask them to generate a query as we do not expect them to be familiar with that term. Instead, we observed that users very often describe what they see and feel, so we set up Amazon Mechanical Turk tasks that simulate the experience of having a given symptom. We identified medical symptoms with a strong visual component, and found three diverse pictures that demonstrated the symptom. The Mechanical Turkers were then shown one of these pictures and asked "If you were the patient in this picture, what queries would you issue to find out what is wrong with you?". We also found 10 - 30 second clips from YouTube of patients experiencing a symptom without extra commentary and repeated the task with this training data. The videos covered auditory symptoms such as stridor (a type of troubled breathing), whooping cough, and hypoventilation (slow breathing), and movement symptoms such as nystagmus (uncontrolled eye movements), parkinsonism, and raynaud's phenomenon (a discoloration of the hands due to problems with blood flow). The task set up is in Figure 1.

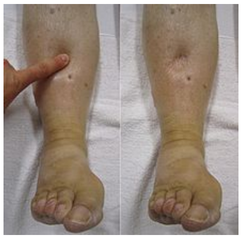
We restricted our set of allowed workers to only Mechanical Turkers located in the United States and set an average payment rate of $10 an hour to encourage high quality work. In our analysis of the data obtained, we found that over 94% of the provided queries were of high quality when prompted by images, and over 97% was of high quality when prompted by a video. We demonstrate some of the data obtained in column 2 of Table 1.

There are two major drawbacks to this approach. The first is that crowdsourcing participants are not a representative sample of internet users, let alone the population. There is not strong consensus about the demographics of crowdworkers - it differs from platform to platform and hour of the day, though the most reliable data may come from Ross *et al.* [28] who show that it combines moderate-income US workers with an international force consisting primarily of well-educated Indian workers.

The more crucial issue with this approach is that it is impossible for anyone not experiencing the symptom to accurately describe how the symptom feels, the pain experienced, and so on, as well as the additional signals that would be available from someone in the same room as the patient (as in a parent issuing a query on behalf of their child). Fortunately, this drawback exists only in situations where personal experience is necessary. Even there, it can be somewhat mitigated through the use of video (e.g. results in Section 7 show that crowdsourced queries from videos are of higher quality). Despite the fact that the Turkers were unable to experience the symptoms, they still projected their own experiences onto the task and did mention pain words such as 'sore'. Unlike the queries that we observed in query logs, the provided queries tend to be shorter and not mention as many details about the situation, such as how long the symptom has been occurring for, the age of the patient, or a combination of multiple symptoms and conditions.

Crowdsourced Task

Imagine that you are experiencing the health problem shown below. Please provide three search queries that you would issue to find out what is wrong.
Instructions:
You must provide 3 distinct search queries.
The search queries must relate to what you see below.
Each query must completely describe what you see below.



Search Query 1: ____________
Search Query 2: ____________
Search Query 3: ____________

Figure 1: The crowdsourcing task for obtaining training data. The symptom shown is *edema.*

We believe that this approach of crowdsourcing data generation can be easily generalized to other domains where the issue with personal experience is a significantly smaller factor. For example, one could easily imagine showing a Mechanical Turker a video of how a broken car starter motor sounds and asking how they would search for information about the observed problem.

5. OUR APPROACH

In this section, we describe the collection of knowledge sources used to understand what a medical symptom is, as well as how they were used to construct a feature space for machine learning. Our focus is on using publicly available datasets to demonstrate the power of existing knowledge sources.

5.1 Knowledge Sources

There is a wealth of knowledge readily available about medical symptoms that should be exploited in any solution to Problems 2 and 3. Prior unsupervised approaches do not seem to use these sources. We also describe knowledge sources that help us understand a query.

Symptom Name	Crowdsourced Queries	Wikipedia Redirects
alopecia	[baldness in multiple spots] [circular bald spots] [loss of hair on scalp in an inch width round]	[elderly hair loss] [loss of hair] [ways to fight hair loss] [alopetia]
angular cheilitis	[broken lips] [dry cracked lips] [lip sores] [sores around mouth]	no redirects
edema	[fluid in leg] [puffy sore calf] [swollen legs]	[dropsical] [dropsy] [hydropsy]
exophthalmos	[bulging eye] [eye balls coming out] [swollen eye] [swollen eye balls]	[bulging eyes] [eye popping] [exophthalmia]
hematoma	[hand turned dark blue] [neck hematoma] [large purple bruise on arm]	[haematoma] [hemotoma] [organizing haematoma]
jaundice	[yellow eyes] [eye illness] [white part of eye turned green]	[icteric] [yellowed] [scleral icterus]
psoriasis	[red dry skin] [dry irritated skin on scalp] [silvery-white scalp + inner ear]	[oil drop nail] [psoriatic] [cerises]
urticaria	[hives all over body] [skin rash on chest] [extreme red rash on arm]	[hives] [nettle rash] [autographism]

Table 1: A sample of the training and testing data

5.1.1 Symptom Data

Medical symptoms have strict formal definitions and are often carefully named from Greek and Latin sources. This information is contained in medical dictionaries and encyclopedias that are widely available. We describe these sources, together with their strengths and limitations.

Greek and Latin roots

A good knowledge of Greek and Latin is helpful in understanding English words. For example, *circumlocution* can be broken into two parts, *circum* meaning around, and *locution* from the Latin word for speech, giving us a definition of "round-about speech". Many medical symptom names derive from Greek or Latin. e.g. the symptom *hematoma* translates into "a mass or collection of blood".

We obtained a list of common Greek and Latin prefixes, suffixes and roots used in medical words, for example, words beginning with *adip-* refer to fat or fatty tissue, and *ocul-* pertains to the eye, and use this to expand the symptom names. For a symptom s, let $\mathit{root}(s)$ denote the set of words that match a prefix, suffix or stem of the words in s.

The drawback to this approach is that some words may have a prefix or suffix, but not truly match the meaning. For example, *herpes* has the Latin suffix "pes" which means "of the foot", but the word herpes is Greek.

For our experimental section, we retrieved a list of 762 Greek and Latin roots, which comprised of 109 suffixes, 645 prefixes and the remainder stems from Wikipedia.

Medical Dictionaries

There are a variety of medical dictionaries online that define symptoms. One example is MeSH (Medical Subject Headings) whose primary purpose is to classify medical articles and books into a taxonomy. MeSH also contains symptom definitions in a field called scope notes. For each symptom, we crawled MeSH for the scope note definition. For a symptom s, we define $\mathit{dict}(s)$ to be the set of words in the dictionary definition of the symptom.

One limitation of dictionaries is that the definition may be more convoluted than the symptom. For example, the scope note for *cough* expands to "a sudden, audible expulsion of air from the lungs through a partially closed glottis, preceded by inhalation." Overall, however, we believe that dictionary definitions will be helpful in shedding light on the symptom.

Encyclopedias

Whereas dictionary definitions are short, encyclopedias expound further on concepts. Wikipedia is a natural resource for such content. The first paragraph tends to have a definition and contain some distinguishing features about the symptom. For a symptom s, we define $\mathit{encyc}(s)$ to be the set of words in the encyclopedia description of s.

Utilizing data from Freebase and Wikipedia, we obtained a list of 1800 symptoms with 1100 Wikipedia entries. We removed the stop words, and extremely common and rare words from the first paragraph of these entries.

Synonyms

Synonyms can play a large role in understanding a query. WordNet [30] is a natural choice for a synonym service. The primary drawback with using WordNet on a search query (that is not a sentence) is that it is not context aware. For example, 'affection' is a valid synonym of 'heart', yet it is not appropriate when 'heart' is mentioned in a medical context. For a query q, let $\mathit{syn}(q)$ denote the set of synonyms of the words in q.

Paraphrases

There is a body of paraphrasing literature that benefits from Machine Translation data [22, 35]. The principal idea is to take a phrase translated from one language, say French, to English. If that phrase is translated in two different ways to English then these two translations may be paraphrases of each other. If these translations are repeated across many languages, there is further evidence to support paraphrasing. The tools in this space are better at translating from formal/uncommon language to more generic language, e.g., "I am suffering from tachycardia" paraphrases to "I am suffering from heart palpitations", but not vice versa. For a symptom s, we let $\mathit{para}(s)$ denote the set of words in the paraphrase of s. We used the Microsoft Translator's Paraphrase API [3] to find the top 6 paraphrases for *I'm suffering from* $\langle \mathit{symptom\_name} \rangle$.

Anatomy

Body parts are routinely mentioned in search queries and there are many resources on the web that contain lists of body parts. For a phrase p, let *body*(p) denote the body parts mentioned in the phrase.

One limitation of matching a search query with a body part is that people may be both overly general and specific when describing the body part. For example, a user may issue a query such as [swollen eye] when the actual body part affected is the *eye orbit* (see *periorbital edema*). Another common example is that skin conditions often mention the body part that the skin covers, such as [rash on arm]. The skin condition problem is particularly challenging as some symptoms affect only certain types of skin (akin to the skin lesions that appear in lupus) while others can appear anywhere on the body so sometimes the body part mentioned is helpful, while other times it is misleading. We obtained anatomy data from WordNet.

Colors

Colors play a significant role in diagnosing and distinguishing medical symptoms. One limitation of using color is that two users can describe the same color in different ways, e.g., *pale red* and *pink*. Some of these expressions are included in our synonym data but we also expect that the color substitutions used will be symptom dependent. For a set of words p, let *color*(p) denote the set of color words in p. A list of 949 colors were obtained from [25].

We opt to represent our data over words as opposed to phrases as phrases are not known to add predictive power over words [6]. While each of these data sources have their own imperfections, we will shortly see how in combination they can improve our understanding of a symptom and query.

5.2 From Knowledge Sources to Feature Representations

To represent the combination of a query and symptom, we make use of the knowledge sources just described to create a multidimensional space. Each feature of this space captures the similarity of a symptom and a query in a different aspect. For example, one feature could be the similarity of a query to the dictionary definition of a symptom. To compute the query feature, we turn the query q into a bag of words where the features are words and the count is the number of times the word appears in the query, denoted by *vector*(q). Similarly, for the symptom s, we convert dict(s) into a similar vector space over words. We then take the cosine of the angle between these two vectors to represent similarity in this regard, i.e., cos(vector(q), vector(dict(s))). We will often drop the vector notation and simply write cos(q,dict(s)). To compute a feature based on body part similarity, we could for a query q and a symptom s compute cos(body(q), body(encyc(s))).

5.3 Learning a symptom classifier

Armed with crowdsourced labels and feature vectors, we have reduced circumlocution to a classical machine learning problem. We learn two kinds of classifiers. In the multiclass setting, we learn a classifier that can only predict well on symptoms that it has previously seen. In this setting, the label of each feature vector is the symptom. The features are computed as described and the bag of words in the query can also be used as additional textual features. In the two-class setting, the goal is to understand what combinations of features can generalize to unseen classes. Every combination of query and symptom is used to create a feature vector. Queries that are circumlocutions of a symptom are labeled positive, and all other combinations are labeled negative.

Query	Symptom
q	s, dict(s), encyc(s), root(s), syn(root(s)), para(s)
syn(q)	root(s), syn(root(s))
body(q)	body(dict(s)), body(encyc(s))
body(syn(q))	body(dict(s)), body(encyc(s))
color(q)	color(encyc(s))

Table 2: Query × symptom combinations used to generate features.

6. EXPERIMENTS

The goal of our experiments is to demonstrate that we can generate high quality training and testing data reliably and that we can predict the symptom of a given circumlocutory query. We manually evaluated the training data and compared the performance of our learning algorithms with and without the addition of knowledge sources. Finally, we evaluated how well the learning algorithm is able to generalize from our feature space via a test set that was obtained from Wikipedia redirects.

6.1 Generating Labeled Data

We have two methods for obtaining labeled data. The first is the crowdsourcing experiment described in Section 5 and the second involves Wikipedia redirects. We describe the strengths and limitations of these methods, as well as how we used them in our experiments. In the crowdsourced data, we demonstrate two key findings: that the overall quality of the training data is high - the crowdworkers completed the task to the best of their ability - and that the task of finding the correct symptom from a circumlocution is quite hard - we, the authors, when faced with the task of labeling the queries with symptoms were only able to label 56% correctly, despite being knowledgeable about the space of symptoms.

6.1.1 Crowdsourced Labels

We selected a collection of 41 medical symptoms or conditions which have a strong visual component. Of this, 31 had images and 10 had videos. A sample of these symptoms are shown in the first column of Table 1. Our training data is skewed towards cutaneous, audible and movement based conditions since we require visible or auditory evidence for our query generation method.

To assess the quality of the training data, the authors randomly selected 10% of the (query, symptom) combinations and evaluated the quality of the generated data. We show a breakdown in Table 3. For our first analysis, we concerned ourselves with evaluating whether or not the crowdworkers were able to successfully complete the task of generating medical queries to the best of their ability. We took a random sample of 10% of the generated queries and rated whether the query was a good circumlocution, contained the correct symptom name, contained an incorrect symptom name, or was too general to classify.

Type	Percent
Images	
Good Circumlocutions	88%
Ground Truth Present	14%
Wrong Symptom	13%
Too General	6%
Videos	
Good Circumlocutions	92%
Ground Truth Present	2%
Wrong Symptom	6%
Too General	3%

Table 3: Quality of crowdsourced training data. The categories do not sum to 100% as the training data may mention the wrong symptom but also be a good circumlocution, or may mention the colloquial symptom name.

The overall quality of the crowdsourced data is quite high (94% with only 6% too general), and in some cases (14%) the exact professional or colloquial name was given. In 13% of the cases, the wrong symptom was given and 88% of the time a good circumlocution was provided. When the queries produced by the crowdworkers were incorrect, we do not believe that the goal was to mislead. For example, in a picture of a jaundiced patient, poor screen quality may have caused a user to not realize that the patient was yellow, and instead focus on a different aspect of the image, resulting in the query [lost hair on scalp]. Second, as previously noted, people who are not experiencing the symptom cannot accurately describe pain or feelings associated with the symptom. Finally, a small percentage (6%) the queries were too general to be useful, e.g., the query [eyes] or [yellow] to describe a jaundice patient. Note that the numbers do not sum to 100% because in some cases a query could be assigned to multiple categories, e.g., be a good circumlocution but name the wrong symptom, e.g., [athlete's foot] for psoriasis.

The problem of learning a symptom from a query is quite challenging. As an experiment, we utilized the same 10% sample of Turk queries that was categorized above (without the symptoms) and attempted to label the symptom ourselves. The labeler was very familiar with the symptom names and definitions of the symptoms and had even hand curated the pictures that were shown in the Mechanical Turk study, but had never been exposed to the generated queries. To our surprise, our own labels were only 56% accurate. Much of the challenge came from the fact that many of the symptoms are quite similar - iritis and conjunctivitis, or the various skin rashes. Vague queries about these symptoms without great detail are challenging to differentiate. The question is: can an automated algorithm do better?

6.1.2 *Wikipedia Redirects*

Our other source of labeled data is Wikipedia redirects. These redirects are quite powerful: issuing a query such as [fast heart rate] in Wikipedia redirects to the Wikipedia page on [tachycardia]. We scraped redirects from the Wikipedia link graph for 370 symptoms with an average of 12 redirects per symptom. Some examples of redirects are shown in the last column of Table 1. There are only 10 symptoms that occur in both the Wikipedia redirect data set and the one generated by the image Mechanical Turk task.

Wikipedia redirects are not designed solely to deal with circumlocution and often redirect to a professional term from an even more specific professional term e.g. *hyperbilirubinaemia* to *jaundice*. There are foreign language redirects, e.g., from *signe de la tache de bougie* to *psoriasis*, as well as spelling redirects e.g. from *alopetia* to *alopecia*. Ultimately, these redirects exist to connect people to the best content available in Wikipedia. As a result, the data is not ideally suited for training purposes. However, improvement in performance on this data as a test set suggests an improvement in our ability to predict symptoms. Since Wikipedia is manually curated, the quality of the redirects is high.

6.2 Learning

We conduct two learning experiments to determine whether our knowledge sources have value in symptom prediction. In the first experiment, we train and test on the crowdsourced data. This experiment fixes the symptom set and does not require any generalization, so we are able to determine if our knowledge sources improve upon our ability to predict over a bag of words baseline. In the second experiment we train on crowdsourced data and test on Wikipedia redirects, effectively evaluating how well the feature space allows us to generalize our predictions to unseen symptoms.

6.2.1 *Baseline*

We evaluate two baseline approaches, Batch MetaMap (available at http://metamap.nlm.nih.gov/) and bag of words, i.e., where each word in the query is a feature in a high-dimensional space. Our goal is to demonstrate that we have an effective way of using knowledge sources to solve the problem. Note that we are not comparing to unsupervised prior solutions [36]. An unsupervised approach of clustering queries into latent topics, even perfect latent topics, simply does not solve the problem of identifying the symptom name. In fact, we are not aware of prior solutions that identify the symptom name – in part because of the difficulty in generating training and testing data.

6.2.2 *Train and Test on a fixed set of symptoms*

We learn a multi-class classifier that can predict the symptom of a circumlocutory query. In the baseline, crowdsourced data is used to create a bag of words representation: given a symptom and given a crowdsourced query, we construct a vector space over the words in the query. The dimensions of the vector are words and the values correspond to how many times the word appears. Our approach uses the features described in Table 4. In the case of images, since there are 31 symptoms, the classifier must identify which of the 31 classes is the most likely candidate. In the case of videos, there were 10 symptoms. A multiclass linear regression algorithm was used for both feature sets [2].

We repeat 10 runs of 10-fold cross validation. The average and standard deviation over the runs is reported in Table 4. Results are reported using both micro-average and macro-average accuracy. Micro-average reports the total fraction of correct symptoms while macro-average averages on a per-symptom basis. Random guessing is expected to be 3% accurate for images (31 symptoms) and 10% accurate for videos (10 videos).

The performance of bag of words is 47% micro-average. For images, our approach shows an improvement to 61% micro-average, see Table 2. This is a 33% improvement over

Type	Micro-Average Accuracy (stdev)	Macro-Average Accuracy (stdev)
Image: Random Guessing	0.03 (—–)	0.03 (—–)
Image: Bag of Words	0.47 (0.05)	0.41 (0.05)
Image: Bag of Word plus Table 2	0.61 (0.05)	0.54 (0.06)
Video: Random Guessing	0.10 (—–)	0.10 (—–)
Video: Bag of Words	0.69 (0.04)	0.57 (0.04)
Video: Bag of Words plus Table 2	0.85 (0.03)	0.72 (0.04)

Table 4: Training and testing a multi-class classifier on the Mechanical Turk generated data. These results compare bag of words with bag of words with additional knowledge sources listed in Table 2. Accuracy is averaged over 10 runs of 10-fold cross validation. Additional knowledge sources improve the final accuracy.

the baseline bag of words approach. The difference in accuracy is statistically significant: under a paired t-test, the p-value is less than 0.001. The method also improves over the 56% accuracy that was obtained by human labeling. For videos, our approach shows an improvement to 85% micro-average, which is a 26% improvement over the baseline. The difference in accuracy in both cases is strongly statistically significant: under a paired t-test, $p < 0.001$.

For MetaMap, we selected a random 10% sample of our crowdsourced data and restricted MetaMap to make predictions of Symptoms, Findings and Diseases. Of these, we graded MetaMap as correct if it had a strong confidence in mapping to a correct symptom or variant, e.g., hand tremors and tremors were both considered correct for our tremors training data. Within this set, MetaMap produced no mapping for 15% of the data, was correct on 12% of the data and produced an incorrect prediction on 73% of the data. Within the correct predictions, 10% of the correct predictions were for only two symptoms and the other 2% correct predictions had the symptom name in the query. Some of MetaMap's failures stem from ignoring the surrounding context, e.g., [trouble falling asleep] was mapped to Finding: Falls and Finding: Sleepiness. Given that bag of words performed better, we do not continue evaluating MetaMap.

6.2.3 Train and Test on a different set of symptoms

For symptoms where Problem 1 can be solved and training data can be generated, the learning approach just described is valuable. However, the major drawback of our approach to generating labeled data is that it requires visual evidence and many symptoms do not have an image or video that is appropriate, e.g., *low blood sugar*. Fortunately, for these symptoms, we still have our underlying knowledge base. We now show how our combined query, symptom feature vectors enable generalization to symptoms where we have no training data.

Our approach translates Problem 3 into a two-class learning problem. For every query, symptom combined feature vector, we label the vector as follows:

$$\text{label}(q, s) = \begin{cases} + & \text{if } q \text{ is a circumlocution of } s \\ - & \text{otherwise} \end{cases}$$

We then learn the weight of the features using a two-class prediction algorithm, specifically Friedman's boosted tree approach [15]. We train on crowdsourced data and test on Wikipedia redirects. Note that using the above scheme, most of the data is labeled negative (99.4%). Consequently, a very accurate classifier would label every query × symptom combination negative. We overcome the class imbalance problem by down sampling the majority class [19]. In our experiment, we down sampled the negatives in the Wikipedia redirects so as to have an equal number of positives and negatives in the test set. This transformation means that a random guessing baseline is expected to be 50% accurate.

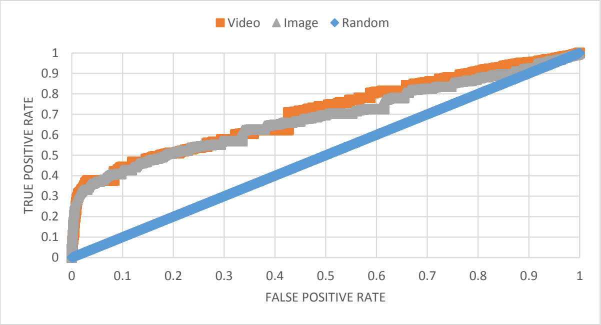


Figure 2: ROC curve for training on Mechanical Turk data and testing on unseen Wikipedia redirects. The negative examples were down sampled to have an equal number of positives and negatives. The charts show the performance with enriched features only, i.e., bag of words features removed

We compared prediction performance with and without textual features and found the results to be very similar. Consequently, the results we report here are based on only the features in Table 2. The prediction results are shown in Table 5. The overall accuracy is 59%. The strength of the approach is positive precision and negative recall which reach 99%. The limitation is positive recall, in other words, there are many query, symptom combinations that are truly positive that the learned classifier fails to connect. Some of these errors are due to the complex nature of the problem. Others may be fixed with additional features and training data. But the primary issue is that many of the redirects serve an alternate purpose than Problem 3. We do not expect our classifier to perform well on redirects from professional to more professional language, e.g., *scleral icterus* to *jaundice*, as the former is not a query we expect a user to issue. But such a redirect helps someone find relevant content in Wikipedia. Similarly, we do not expect our method to perform well on foreign language redirects as our training data was derived from US crowdworkers. The ROC curve is shown in Figure 2 – our method does indeed provide lift over random guessing. Note that the difference between image and video training is negligible in this case.

Finally, in Table 6, we show the top ten features with the most predictive power. The very top features capture the match of the query or synonym of the query to the symptom

Method	Random	Image	Video
Accuracy	0.50	0.59	0.63
Pos. Precision	0.50	0.99	0.97
Pos. Recall	0.50	0.18	0.27
Neg. Precision	0.50	0.55	0.58
Neg. Recall	0.50	0.99	0.99

Table 5: Training on Mechanical Turk generated data and testing on the unseen Wikipedia redirects. The test set was down sampled to have an equal number of positives and negatives. The bag of words only algorithm has no ability to generalize to unseen symptoms and can not do better than random guessing, so is not included. Our additional knowledge sources enabled generalization to unseen categories.

Rank	Query Vector	Symptom Vector	Per-Feature Gain
1	syn(q)	encyc(s)	1
2	q	encyc(s)	0.93
3	q	s	0.78
4	body(syn(q))	body(encyc(s))	0.71
5	syn(q)	syn(root(s))	0.61
6	syn(q)	dict(s)	0.51
7	q	dict(s)	0.49
8	body(q)	encyc(s)	0.45
9	syn(q)	root(s)	0.44
10	q	para(s)	0.32

Table 6: Top ten features with the most predictive power. Per-feature gain summary for the boosted tree ensemble.

or encyclopedia expansions of the symptom. While body parts play a role in prediction, colors are notably absent. This may be due to the fact that colors are already represented in the match between a query and the encyclopedia expansion of a symptom. Thus color may not provide predictive power beyond the top features.

Unlike the previous learning experiment, we are not able to compare to a bag of words baseline here. Crucially, if the bag of words classifier learns what words are relevant for predicting a set of symptoms S, it has no ability to predict which words are important for a set of symptoms T when $S \cap T = \emptyset$. Our combined query, symptom feature vectors facilitate such generalization.

7. CONCLUSIONS AND FUTURE WORK

Our goal is to improve the understanding of the symptoms that users seek to express when generating circumlocutory evidence-based queries. Our motivation for this work and the direction that we have taken - mapping queries to concepts, rather than directly trying to improve a ranking algorithm - is motivated by the observation that web search is increasingly moving in the direction of a 'web of concepts' [10], rather than keyword search. As such, while we have focused on mapping these queries to their professional medical names, we do not intend for these names to be shown to users. It is not helpful to just rewrite a user's query from [headache] to [cephalalgia] since the results will be written in professional language and less understandable to the average user. We believe that mapping to the medical concept can help us design new technologies that can assist the user's health information needs. For example, familiar web search applications may be improved synonym systems where we can identify full colloquial phrasings as synonyms of the medical terms, or more advanced refinement suggestion systems. Identifying colloquial refinements that would assist in diagnosing the problem at hand would be very helpful to the user, e.g. *conjunctivitis* and *iritis* are both eye infections that are distinguished by pain. Web search engines are increasingly adding functionality that goes beyond a set of web results, such as flight status, sports results, and answers to the queries, and we can envision such additional functionality for the medical domain that depends crucially on understanding the users' query as a set of concepts instead of words. Finally, our work may help form the first steps of a conversational input to an automated medical diagnosis system, such as IBM's Watson system [13]. These are but a few of the possibilities within the existing framework of search. Given the sheer number of users issuing medical queries, we expect other applications.

In summary, our contributions in this paper are as follows. We gave a method for generating training data for the problem of finding a medical symptom associated with a circumlocutory search query. We demonstrated that crowdsourcing could be used to create high quality training data. Additional data sources that are freely and readily available on the web were used to create features with predictive power. Our methods are general enough that what we learn can apply to unseen medical symptoms.

Expanding our work to search sessions (a sequence of queries with a short gap between each query) may be valuable, particularly from the viewpoint of improving our ability to identify a collection of symptoms the user is experiencing. Closing the loop to a final medical disease may be helpful to a user, but this is a challenging problem.

The tools developed in this work are not specific to the medical domain. Circumlocution in automobile, computer and legal domains may be solved with similar approaches. For example, images of unusual fluids flowing out of a car or video snippets of transmission problems could be used to generate training data. Latin roots are at the heart of understanding legal terms such as *habeas corpus* and *pro bono*. Expanding our techniques to other domains is a promising direction for future work.

Acknowledgments

We are grateful to Rina Panigrahy for thoughtful discussions, John Stanton for background knowledge and advice on medical symptoms, and Adrienne Porter Felt for advice on experiment design with Mechanical Turk.

8. REFERENCES

[1] R. Agrawal, A. Halverson, K. Kenthapadi, N. Mishra, and P. Tsaparas. Generating labels from clicks. In *Web Search and Data Mining*, pages 172–181, 2009.

[2] G. Andrew and J. Gao. Scalable training of l1 regularized log-linear models. In *International Conference on Machine Learning*, pages 33–40, 2007.

[3] Microsoft Translator's Paraphrase API. http://msdn.microsoft.com/en-us/library/hh847648.aspx.

[4] A. R. Aronson and F. Lang. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3), 2010.

[5] J. Bai, D. Song, P. Bruza, J-Y Nie, and G. Cao. Query expansion using term relationships in language models for information retrieval. In *CIKM*, pages 688–695. Association for Computing Machinery, 2005.

[6] R. Bekkerman and J. Allan. Using bigrams in text categorization. Technical report, Department of CS, University of Massachusetts, Amherst, 2004.

[7] M-A Cartright, R. W. White, and E. Horvitz. Intentions and attention in exploratory health search. In *SIGIR*, pages 65–74, 2011.

[8] X. Chen, P. Bennett, K. Collins-Thompson, and E. Horvitz. Pairwise ranking aggregation in a crowdsourced setting. In *Web Search and Data Mining*, pages 193–202, 2013.

[9] H. Cui, J-R Wen, J-Y Nie, and W-Y Ma. Probabilistic query expansion using query logs. In *WWW*, January 01 2002.

[10] N. Dalvi, R. Kumar, B. Pang, R. Ramakrishnan, A. Tomkins, P. Bohannon, S. Keerthi, and S. Merugu. A web of concepts. In *PODS*, pages 1–12, 2009.

[11] L. Deleger and P. Zweigenbaum. Extracting lay paraphrases of specialized expressions from monolingual comparable medical corpora. In *Workshop on Building and Using Comparable Corpora*, 2009.

[12] A. Elstein, L. Shulman, S. Sprafka, and L. Allal. *Medical problem solving: an analysis of clinical reasoning*, volume 2. Harvard University Press Cambridge, MA, 1978.

[13] D. Ferrucci, A. Levas, S. Bagchi, D. Gondek, and E. Mueller. Watson: Beyond jeopardy. *Artificial Intelligence*, 2012.

[14] Susannah Fox and Maeve Duggan. Health online 2013.

[15] J. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 2001.

[16] R. Gomes, P. Welinder, A. Krause, and P. Perona. Crowdclustering. In *NIPS*, pages 558–566, 2011.

[17] G. Gorry and G. Barnett. Sequential diagnosis by computer. *JAMA*, 205(12):849–854, 1968.

[18] E. Horvitz, D. Heckerman, B. Nathwani, and L. Fagan. Diagnostic strategies in the hypothesis-directed pathfinder system. In *Artificial Intelligence Applications*, 1984.

[19] N. Japkowicz. The class imbalance problem: Significance and strategies. In *IJCAI*, 2000.

[20] D. R. Karger, S. Oh, and D. Shah. Iterative learning for reliable crowdsourcing systems. In *NIPS*, pages 1953–1961, 2011.

[21] A. Keselman, T. Tse, J Crowell, A Browne, L Ngo, and Q Zeng. Assessing consumer health vocabulary familiarity: An exploratory approach. In *J Medical Internet Res*, 2007.

[22] X. Liu, R. Sarikaya, C. Brockett, C. Quirk, and W. Dolan. Paraphrase features to improve natural language understanding. In *Interspeech*, 2013.

[23] G. Luo, C. Tang, H. Yang, and X. Wei. Medsearch: a specialized search engine for medical information retrieval. In *CIKM*, pages 143–152, 2008.

[24] D. Metzler and W. B. Croft. Latent concept expansion using markov random fields. In *SIGIR*, 2007.

[25] List of Colors. http://xkcd.com/color/rgb/.

[26] F. Radlinski, M. Szummer, and N. Craswell. Inferring query intent from reformulations and clicks. In *WWW*, pages 1171–1172. ACM, 2010.

[27] V. C. Raykar and S. Yu. Eliminating spammers and ranking annotators for crowdsourced labeling tasks. *Journal of ML Research*, 13:491–518, 2012.

[28] J. Ross, L. Irani, M. S. Silberman, A. Zaldivar, and B. Tomlinson. Who are the crowdworkers?: shifting demographics in mechanical turk. In *CHI*, 2010.

[29] O. Tamuz, C. Liu, S. Belongie, O. Shamir, and A. Kalai. Adaptively learning the crowd kernel. *International Conference on Machine Learning*, 2011.

[30] Princeton University. About wordnet, http://wordnet.princeton.edu, 2010.

[31] F. L. Wauthier and M. I. Jordan. Bayesian bias mitigation for crowdsourcing. In *NIPS*, 2011.

[32] P. Welinder, S. Branson, S. Belongie, and P. Perona. The multidimensional wisdom of crowds. In *NIPS*, 2010.

[33] J. Wen, J-Y. Nie, , and H. Zhang. Query clustering using user logs. *ACM Trans. on IS*, 20(1):59–81, 2002.

[34] R. W. White and E. Horvitz. Cyberchondria: Studies of the escalation of medical concerns in web search. *ACM Trans. Inf. Syst*, 27(4), 2009.

[35] W. Xu, A. Ritter, B. Dolan, R. Grishman, and C. Cherry. Paraphrasing for style. In *COLING*, 2012.

[36] S. Yang, S. P. Crain, and H. Zha. Bridging the language gap: Topic adaptation for documents with different technicality. *JMLR*, 15:823–831, 2011.

[37] J. Yi, R. Jin, A. Jain, S. Jain, and T. Yang. Semi-crowdsourced clustering: Generalizing crowd labeling by robust distance metric learning. In *NIPS*, pages 1781–1789, 2012.

[38] Q. T. Zeng, J. Crowell, R. M. Plovnick, E. Kim, L. H. Ngo, and E. Dibble. Assisting consumer health information retrieval with query recommendations. *Journal of the American Medical Informatics Association*, 13(1):80–90, 2006.

[39] Q. T. Zeng and T. Tse. Exploring and developing consumer health vocabularies. *Journal of the American Medical Informatics Association*, 13(1):24–29, 2006.

[40] Q.T. Zeng, E Kim, J Crowell, and T Tse. A text corpora-based estimation of the familiarity of health terminology. In *ISBMDA*, volume 3745, 2005.

[41] Q.T. Zeng, T. Tse, G Divita, A Keselman, J Crowell, S Goryachev, and L Ngo. Term identification methods for consumer health vocabulary development. In *J. Med Internet Res*, volume 9, 2007.

[42] Q Zeng-Treitler, S. Goryachev, T. Tse, A. Keselman, and A. Boxwala. Estimating consumer familiarity with health terminology: A context-based approach. *Journal of the American Medical Informatics Association*, 15(3):349–356, 2008.