Preclude: Conflict Detection in Textual Health Advice

Sarah Masud Preum*, Abu Sayeed Mondol†, Meiyi Ma‡, Hongning Wang§, and John A. Stankovic¶
Department of Computer Science, University of Virginia
Charlottesville, Virginia 22904
Email: *preum@virginia.edu, †mm5gg@virginia.edu, ‡mm5tk@virginia.edu, §hw5x@virginia.edu, ¶jas9f@virginia.edu

***Abstract*—With the rapid digitalization of the health sector, people often turn to mobile apps and online health websites for health advice. Health advice generated from different sources can be conflicting as they address different aspects of health (e.g., weight loss, diet, disease) or as they are unaware of the context of a user (e.g., age, gender, physiological condition). Conflicts can occur due to lexical features, (such as, negation, antonyms, or numerical mismatch) or can be conditioned upon time and/or physiological status. We formulate the problem of finding conflicting health advice and develop a comprehensive taxonomy of conflicts. While a similar research area in the natural language processing domain explores the problem of textual contradiction identification, finding conflicts in health advice poses its own unique lexical and semantic challenges. These include large structural variation between text and hypothesis pairs, finding conceptual overlap between pairs of advice, and inference of the semantics of an advice (i.e., what to do, why and how). Hence, we develop *Preclude*, a novel semantic rule-based solution to detect conflicting health advice derived from heterogeneous sources utilizing linguistic rules and external knowledge bases. As our solution is interpretable and comprehensive, it can guide users towards conflict resolution too. We evaluate *Preclude* using 1156 real advice statements covering 8 important health topics that are collected from smart phone health apps and popular health websites. *Preclude* results in 90% accuracy and outperforms the accuracy and F1 score of the baseline approach by about 1.5 times and 3 times, respectively.**

I. Introduction

The growing interest in digital health [1] and the increasing trend of health consumerism have accelerated the development of numerous health apps and websites. In 2014, one fifth of U.S. adults have regularly used at least one mobile health app [2]. Internet usage for health information has also increased: as of June 2016, WebMD receives about 80 million unique visitors every month. Such digital resources are gaining more trust as people often use advice derived from these for decision making, e.g., deciding diets or planning exercise routines. Often people use multiple health apps or online resources simultaneously to educate themselves [3], manage different physiological conditions (e.g., pregnancy and food allergy), track multiple chronic conditions, or even to achieve different personal goals (e.g., weight loss and healthy diet) [4], [5].

Advice generated from multiple health apps/websites can be **conflicting** [3], [6] due to three factors. Firstly, when two information sources (i.e., app/website) correspond to different health topics (e.g., pregnancy and weight loss) they might be conflicting. Secondly, even when two sources are related to the same topic, conflicts may occur due to conflicting findings from the underlying research corresponding to each source [7]. Finally, an app/website may lack contextual awareness of a user and suggest an advice that adversely interacts with the lifestyle, diet, disease, or medications of the user and thus causes a conflict. For instance, in case 1 of Table I a diet app suggests eating green leafy vegetables while a medication app is reminding the user about potential negative interaction between green leafy vegetables and a drug (i.e., Coumadin[1]) the user is currently prescribed. Here the two apps suggest directly opposite actions and thus result in conflict. Similarly, conflicts can arise from multiple advice coming from heterogeneous sources where the suggested *actions* of the advice are directly opposite (e.g., due to negation, antonym) or quantitatively different. For example, the amount of required dietary fiber intake is different for a healthy adult and an adult suffering from abdominal bloating. This also emphasizes the need to infer the **contexts** of conflicts so that people are not bothered with false alarms.

Manually detecting conflicts in health advice generated from multiple sources is challenging. Because, (i) people often do not process advice statements [2] at the same time as advice from different sources can appear intermittently. So one needs to thoroughly recall all advice to detect potential conflicts, which is unrealistic. Manually detecting conflicts in textual health advice also requires increased human attention, which is a scare resource for pervasive applications, as identified in existing research [8], [9]. (ii) People often lack the necessary domain knowledge for conflict detection. For example, a diet app suggests to increase intake of Kale while an online article on digestive health suggests *heavy consumption of cruciferous vegetables can lead to hypothyroidism*. Now to detect a conflict, one needs to be aware of the fact that Kale is a cruciferous vegetable.

With the advent of pervasive health information sources, an increasing number of people are using these sources for health related decision making on a daily basis [10], [11], sometimes even without consulting a professional health-care provider [11]. As people trust these sources of advice [10], [11], they may follow conflicting advice statements without being aware of potential consequences. If left undetected, these conflicts can sometimes pose serious threats to a user's health by adversely affecting different physiological parameters, (e.g., increasing heart rate / blood pressure too much) and making them susceptible to adverse physiological conditions. To the best of our knowledge, no existing system intercepts advice to detect or resolve conflicts across textual advice statements

[1]It is used to treat blood clots in veins or arteries.
[2]We use the terms *advice* and *advice statement* interchangeably

978-1-5090-4327-9/17/$31.00 ©2017 IEEE

from heterogeneous sources. ***Preclude*** focuses on detecting conflicts in health advice from health apps and websites in an interpretable manner using linguistic features and external knowledge bases. While our solution does not guarantee safety, our end goal is to minimize the risk of conflicting advice as much as possible. This paper focuses only on detecting conflicts in textual health interventions generated from smart phone apps and websites. But in the future the number of conflicting health interventions may increase with the growing interest in pervasive health and medical cyber physical system applications [12], [13], [14]. ***Preclude*** can serve as the building block to detect potential conflicts in the interventions generated from such platforms.

Detecting conflicts in textual health advice poses both **lexical** and **semantic** challenges. This task is lexically challenging as the lexical structure of advice text can vary significantly in terms of length of advice and/or tone of advice. The semantic challenges of conflict detection are multifold. Firstly, we need to extract the implied action and resulting effects of an advice from the text. Secondly, we need to detect whether two or more advice statements have any conceptual overlap (e.g., Kale and cruciferous vegetables). Detecting conceptual overlap often requires inferring the hierarchical relationships between different topics, such as, foods, drugs, and exercise. Finally, often conflicts are temporal or conditional, i.e., a conflict occurs if a temporal/physiological condition holds true. Hence, it requires thorough inference of the semantics of an advice.

There are some existing works that focus on detecting contradiction in a given pair of sentences/texts [15], [16]. They model this problem as a binary classification task and apply statistical learning models. But detecting a contradiction in a given pair of sentences/texts is different from detecting conflicting advice. Because, the former does not require (i) detecting conceptual overlap to find potential candidates of conflict and (ii) understanding the semantics of an advice, e.g., action, effect, condition. Also, statistical learning models require a lot of labeled training data which is currently unavailable for textual health advice.

We present ***Preclude***, a semantic rule based system to detect conflicts in text advice derived from mobile health apps and websites. The main contributions of this work are as follows.

- *Preclude* is the first to formulate the problem of detecting conflicts in textual health advice derived from heterogeneous sources. It provides a comprehensive **taxonomy** of potential conflicts and a detail semantics of advice that guides the conflict detection process.
- *Preclude* combines linguistic features with multiple, rich external knowledge bases (e.g., MetaMap[17], WordNet[18], Wikipedia) to generate **semantic** rules and detect conflicts in an **interpretable** manner.
- We are the first to create and release a health advice dataset comprising of **1156** health advice statements covering 8 important health topics. These advice statements are collected from several real mobile health apps and popular health websites[3]. The dataset is annotated for conflicting pair of advice statements.
- Based on the evaluation using the aforementioned dataset, on average *Preclude* results in 90% accuracy in conflict detection and outperforms the accuracy and F1 score of the baseline approach by about 1.5 times and 3 times, respectively. In addition, we demonstrate the effect of context awareness in *Preclude*. Also a controlled user survey of 24 participants is conducted to emphasize the importance of detecting conflicting health advice.

[3]This dataset is available upon request to the authors.

II. Problem Formulation

In order to position our solution we first carefully define *conflict* of advice. Each advice involves a set of *action*s and each action results in a set of *effects*. Before formally defining *conflict*, we need to define *object*, *action*, and *effect* of an advice, $advice_i$.

Object (o_i): Each health advice suggests either in favor of or against each in a set of objects. For example, from Table I, objects of advice 1 of case 1 are citrus fruits and green leafy vegetables. Each object of an advice can contain **sub-typical** (s_i) semantics (e.g., *green leafy* vegetable).

Action (a_i): Action is the intervention that is implied by an advice either directly as in an imperative sentence or indirectly as in a declarative sentence. Referring to Table I, the action (i.e., eating) is directly mentioned in case 1 while it is implied in advice 2 from case 4. Each action of an advice is often associated with different semantics that suggest people when and how to perform the action. An action can specify **quantity** (q_i) (e.g., *200 mg* of coffee) of corresponding object(s). Quantity can be specified using numerical (n_i) or adverbial quantifier (f_i) (e.g., more, few). An action (a_i) can be conditional or temporal. This is specified by one or more **conditional** clauses (c_i) and/or **temporal** clauses (t_i) in an advice text. Such as, in Table I advice 1 from case 3 and case 4 suggest the time and physiological condition of the corresponding action, respectively.

Effect (e_i): An effect refers to the purpose or resulting physiological effect of an action. For example, in case 2 of Table I, a potential effect of consuming Pate made from meats is Listeriosis. Often effect can create a chain of subsequent effects, such as, primary effect, secondary effect, tertiary effect, and so on.

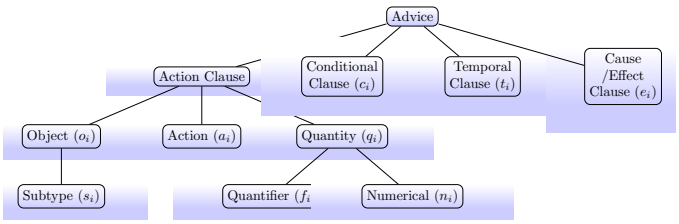

Fig. 1: Semantic decomposition of textual health advice

Thus an advice statement $advice_i$ can be expressed as a tuple of **semantic tokens**: $advice_i$: $<s_i^m, o_i^m, a_i^m, q_i^m, c_i^m, t_i^m, e_i^m>$as presented in Figure 1. Here m denotes the index of tuple m. Often a single advice can contain multiple objects. Then each object results in a tuple. Also, q_i^m can be either adverbial quantifier (f_i^m) or numerical (n_i^m) or both. Note

978-1-5090-4327-9/17/$31.00 ©2017 IEEE

	Cases	Advice 1	Advice 2
1	Opposite polarity (actions)	Eat citrus fruits and green leafy vegetables as they are rich in Vitamin C.	Be careful about green leafy vegetables if you are on Coumadin or ACE Inhibitors.
2	Opposite polarity (effects)	Pate made from meats may carry the listeria bacteria and cause listeriosis. Avoid eating it while pregnant.	Consume red meat at least two to three times a week to fight anemia.
3	Temporal	Do stretching exercises when you wake up.	Avoid stretching or similar exercises after the end of week 12 of your pregnancy.
4	Conditional	Alcohol may severely affect your baby's development. Avoid alcohol if pregnant or trying to conceive.	Small amounts of alcohol increase the body's metabolic rate, causing more calories to be burned.
5	Sub-typical	Eat calcium-rich foods like milk, cheese and green vegetables.	Use skimmed milk instead of whole milk as dairy products often cause bloating and gas.
6	Quantitative	Limit your caffeine intake to less than 200 milligrams per day during pregnancy.	Up to 400 milligrams (mg) of caffeine a day appears to be safe for most healthy adults.
7	Cumulative effect	Run for at least 30 minutes a day.	Take Salmeterol 1 inhalation (50 mcg) twice daily.

TABLE I: Possible cases of conflict: all advice are taken from real health apps or authentic medical sites

that, an action and an effect of an advice can be mapped into positive or negative polarity with respect to the corresponding object. Such as, in case 2 of Table I, action in advice 1 has negative polarity with respect to *meat* while effect in advice 2 has positive polarity with respect to *red meat*.

At first we define pair-wise conflict between two statements of advice. This definition can be extended to define conflict among any size set of advice.

Conflict: Two pieces of advice $advice_i$ and $advice_j$ are conflicting with each other if they have at least one common object ($o_i^m = o_j^n$) and at least one of the following is true:

1) Opposite polarity of actions, i.e., a_i^m and a_j^n have opposite polarity (case 1 of Table I).
2) Opposite polarity of effects, i.e., e_i^m and e_j^n have opposite polarity (case 2 of Table I).
3) Opposite polarity of action-effect, i.e., a_i^m and e_j^n have opposite polarity or e_i^m and a_j^n have opposite polarity (case 4 of Table I).
4) Both of the advice have the same polarity but they are quantitatively different from each other, i.e., q_i^m is not compatible with q_j^n (case 6 of Table I).
5) The cumulative effect(s) of the actions of the pair of advice statements exceeds a safety threshold, although when performed individually none of the advice results in exceeding the safety threshold (case 7 of Table I).

The above cases demonstrate **direct** and **quantitative** conflicts. In addition, conflicts between a pair of advice statements can be **conditional, temporal, sub-typical** based on the semantics of advice tokens. For example, the fourth pair of Table I have opposite polarity (according to rule 3 presented above) and one advice of the pair has a condition. So it is a conditional conflict. In addition to detecting conflicts, *Preclude* also extracts such semantic refinements from potential conflicts.

Table I provides concrete examples of different types of conflicts. To begin with, case 1 presents a pair of advice statements that demonstrate conflict due to opposite actions, i.e., eating vs. not eating green leafy vegetables. Case 2 demonstrates conflict due to opposite effects, i.e., causing listeriosis vs. fighting anemia. Sometimes conflicts occur due to a physiological, temporal or contextual condition. For instance, in case 3, the conflict in performing stretching exercise is due to pregnancy. This example also demonstrates how the conflict detection should be aware of the physiological contexts (e.g., pregnancy, disease, medical history) of a user. For the pair of advice statements in case 5, a conflict occurs only for skimmed milk. Thus, it is a sub-typical conflict. Case 6 demonstrates another case of conflict that arise due to quantitative differences. This is also an example of conditional conflict. Case 7 illustrates a cumulative conflict. In this case, running and taking Salmeterol both increase heart rate temporarily. Although performing only one of the actions is usually safe, doing both of them simultaneously can increase heart rate beyond the range of normal heart rate. Thus, we provide a **comprehensive** and **interpretable** taxonomy of conflicts which can play an important role in conflict resolution.

III. Scope of Study

This research focuses on health and wellness advice targeted for general audiences that are found in online health sites and mobile health apps. The choice of our operational scenario for digital health advice is guided by three factors: (i) most popular general health topics (i.e., exercise, diet, and weight loss) [5], (ii) most common conditions and diseases for which people use these digital resources (i.e., pregnancy, diabetes) [19], [20], and (iii) potential interactions between different health topics. Hence, we choose 8 health topics as presented in Table II. Although anemia and digestive health (e.g., food allergy/intolerance) do not belong to the most popular health topics, we include them in our study as (i) a significant portion of population world wide suffer from these [21], [22] and (ii) they demonstrate interactions with other topics, i.e., their advice are sometimes conflicting with advice of other selected topics (e.g., diabetes, weight loss).

In the fourth column of Table II we have listed 8 different health apps. Among these 4 are Android apps (Effective Weight Loss Guide, Healthy Nutrition Guide, Health and Nutrition Guide, and Anemia Help) and the other 4 are iOS apps. As the sources of online health advice, we have used WebMD, Yahoo! Health, MayoClinic, and HealthLine, all of which belong to top ten most popular health sites as of 2016 [23]. We have collected advice statements from these sources that primarily relate to food, exercise, life style (sleeping, drinking), and some over-the-counter drugs. Some of these

978-1-5090-4327-9/17/$31.00 ©2017 IEEE

Health Topic	Number of Advice		Mobile App Name
	Health websites	Mobile Apps	
Anemia	39	6	Anemia Help
Diabetes	81	18	Health & Nutrition Guide
Digestive health	30	35	Food Shopping Essential
Diet	85	256	Healthy Nutrition Guide Health & Nutrition Guide Effective Weight Loss Guide
Exercise	51	60	Health & Nutrition Guide Effective Weight Loss Guide
Pregnancy	48	92	Pregnancy Pregnancy Foods to Avoid
Weight loss	32	323	QuickWeight Effective Weight Loss Guide
Total	366	790	1156

TABLE II: Numbers of advice collected from 8 health topics from authentic health websites (column 2) and mobile health apps (column 3). The rightmost column contains the name of the mobile apps. Advice statements on a single topic were collected from several websites, so the website names are not presented here for the sake of brevity.

advice statements are aimed at certain context, e.g., advice for lactose intolerant people. These contexts are encoded in the apps as *metadata*.

IV. Solution

Algorithm 1: ***ConflictDetect***(L, A_j)

Input : L, list of advice since time T;
Input : A_j, new incoming advice;
Output : conflictFlag, typeOfConflict;

1.1 H: HashMap of each advice in L and corresponding token ;
1.2 conflictFlag $\leftarrow$ false ;
1.3 $A'_j \leftarrow$ Preprocess(A_j) ;
1.4 $S_{A'_j} \leftarrow$ ExtractSemanticClauses(A'_j) ;
1.5 $S_{A_j} \leftarrow$ ExtractSemanticTokens$(S_{A'_j})$;
1.6 **for** *each advice statement* A_i *in* L **do**
1.7 $S_{A_i} \leftarrow H$.getValue(A_i);
1.8 $SO_i \leftarrow$ set of objects from S_{A_i};
1.9 $SO_j \leftarrow$ set of objects from S_{A_j};
1.10 **if** $SO_i \cap SO_j \neq \phi$ **then**
// to check context metadata
1.11 $Sub_i \leftarrow$ set of subjects from S_{A_i};
1.12 $Sub_j \leftarrow$ set of subjects from S_{A_j};
1.13 **if** *compatible*(Sub_i, Sub_j) ==*true* **then**
1.14 $CO \leftarrow SO_i \cap SO_j$;
1.15 **for** *each object o in* CO **do**
1.16 conflictFlag $\leftarrow$ false ;
1.17 $P_i \leftarrow$ AssignPolarity(S_{A_i});
1.18 $P_j \leftarrow$ AssignPolarity(S_{A_j});
1.19 **if** $P_i \neq P_j$ **then**
1.20 conflictFlag $\leftarrow$ true ;
1.21 DetectRefinedConflict(S_{A_i}, S_{A_j});
1.22 **else**
1.23 DetectQuantitativeConflict(S_{A_i}, S_{A_j});

Existing textual contradiction detection systems are based on statistical learning [15], [24], [16]. It is not feasible in this case, as statistical learning methods require a significant amount of labeled training data to avoid sparsity of feature space. But, there is no available dataset on conflicting health advice. Also, labelling health advice for potential conflict is intellectually more demanding than labelling potential pair of contradictory sentences.

Hence, we develop ***Preclude***, a novel system consisting of a collection of semantic rules and a conflict detection algorithm (**Algorithm 1**) that detects conflicting pairs of advice statements and types of conflicts by analyzing the semantics of advice statements. Unlike statistical learning based contradiction detection systems, *Preclude* (i) detects conflicts in a context aware manner while utilizing relatively small amounts of training data and (ii) informs users about potential types of conflicts (e.g., temporal, quantitative) that can aid users' decision making process to resolve the conflict. Our assumption is *Preclude* runs as a **watchdog** application in personal devices and intercepts health advice to detect conflicts and thus preclude safety risks.

At first we present our solution for a **single sentence** advice. Later, we describe how *Preclude* handles **multiple** sentence advice statements (Section IV-E).

Preclude uses a collection of novel semantic parsing rules to extract different semantics of an advice (Sections IV-A, IV-B). These rules are empirically learned from training data and are guided by linguistic inference, e.g., the structure of sentences, co-located words and their Parts Of Speech (POS) tags [25], and grammatical relationships of the words. *Preclude* keeps track of all previous advice a user received using a list L. Whenever the user receives a new advice (from an app/website), the advice text is parsed and a typed dependency representation of the advice is generated using the Stanford CoreNLP pipeline [26]. Next, potential conflicts between this advice and any previous advice are detected using the semantic rules and the **4-phase** solution (**Algorithm 1**) as follows.

A. Phase 1: Semantic Clause Extraction

In this phase, an advice statement is divided into four types of semantic clauses. Although there are generic clause extraction tools in NLP to extract noun and temporal clauses, we are the first to extract action, effect, and conditional clauses. We develop semantic clause extraction rules by utilizing dependency relationships found in the advice statements from training data and linguistic patterns of standard English language [27].

Action Clause: It contains action verb(s), object(s), and quantitative tokens of each object. It is further decomposed to

978-1-5090-4327-9/17/$31.00 ©2017 IEEE

extract these tokens (Section IV-B).

Temporal Clause, t_i: It denotes temporal conditions or suggested point of time of an action. They are contained in Prepositional Phrases (PP). Some sample indicators of temporal expressions are: *after, before, as soon as, till, until, when, whenever, while,* and *during*. We create a lexicon of potential temporal expressions by combining lexicons from English grammar [27] and regular expressions from SUTime [28].

Effect/cause clause, e_i: It indicates the purpose of an action in an advice. In case of imperative sentences, the action and object clauses are followed by effect clauses. Here, the effect clause is denoted by prepositions of cause, including *to, as, so, because of, on account of, for, from, out of, due to,* and *in order to*. In addition, by analyzing the training data we find other phrasal verbs that indicate purpose, e.g., *lead to, make, help in*. We create a lexicon of potential effect/cause indicators from training data and grammatical resources [27]. In addition, a set of rules is created to filter false positives in effect extraction. For example, filtering "to" when it does not indicate effect, e.g., *used to, seem to, have to, according to*.

Conditional clause, c_i: It restricts the action under some specific conditions. Conditional clauses are indicated by subordinate clauses or phrases starting with preposition, such as, *if, when, before, after, without* or verbal phrase like, *make sure*.

B. Phase 2: Semantic Tokenization of Action Clause

In second phase, an action clause is further decomposed into the following tokens: action, object, subject, and quantity. Here instead of describing the rules verbatim, we present the intuition and overview of the rules for the sake of clarity.

Action, a_i: This token is present in an advice sentence if (i) the sentence is imperative or (ii) the sentence is declarative and starts with a Verb Phrase (VP), such as,

Imperative: *Include* peanut butter in your daily diet.

Declarative starting with VP: *Adding* peanut butter for cooking helps to fight anemia.

It should be noted that in case of the second advice presented above, *Preclude* tags only *adding* as action token while in case of Parts of Speech (POS) tagging both *adding* and *cooking* are tagged as verb. *Preclude* includes action verbs (e.g., drink, eat), phrasal verbs (e.g., stick to), and negated verbs (e.g., don't take, avoid eating) as action tokens.

Object, o_i: Extracting objects are crucial for conflict detection, as conceptual overlap (i.e., having a common object) is the precondition of conflict between a pair of advice. Objects are noun or noun phrase.[4] The key challenges in this stage are:

(i) Differentiating objects and other noun phrases (i.e., ignoring noun phrases that are not objects).

(ii) Maintaining object hierarchy: Often one advice refers to a sub-type of an object of another advice, e.g., one advice suggests avoiding dairy and another suggests eating cheese. In this case, object extraction should be aware of that cheese is a sub-type of dairy.

(iii) Finding compound objects: Often objects are compound words or phrases. For finding semantic overlap, *Preclude* includes both simple and compound objects, e.g., mapping *apple juice* to <apple, juice, apple juice >. This is crucial due to safety critical nature of the problem (i.e., if someone is prescribed to avoid apples due to fructose intolerance, she should avoid apple juice as well).

These challenges are addressed by utilizing external knowledge base and semantic rules. Firstly, for filtering objects from non-object noun phrases we use *MetaMap*, a knowledge base to discover Metathesaurus concepts referred in text [17]. Specifically, MetaMap is customized based on training data to filter only relevant types of objects, e..g, foods, drinks, activities, diseases, and syndromes. Secondly, to maintain object hierarchy, multiple external knowledge bases are used. The topics requiring object hierarchy include, seafood [29], vegetables [30], grains [31], etc. Finally, compound objects are extracted using semantic parsing rules, e.g., if component words in a compound object are nouns, then consider all of them as candidate objects.

Often objects are associated with modifiers indicating subtype and quantity. These lead to sub-typical and quantitative conflicts. Sub-typical tokens are mapped in objects. Quantitative tokens are described later.

Subject: This token stores context metadata from apps and advice (or null in case there is no metadata). Context metadata refers to the subject to whom an advice is targeted. Subject can be specified as a header to advice or can appear in advice text. Such as, for the advice: *Men should have 30 to 38 grams fiber a day and women (aged between 18-50) should have 25 grams fiber a day.*, the two subjects are *men* and *women (aged between 18-50)*.

Quantitative, q_i: Quantitative clauses or phrases indicate a suggested amount of an object, frequency, or duration of suggested action. Quantitative tokens can be specific (i.e., contain numeric) or indefinite (i.e., contain only quantifier like, *few, more, plenty*). Although the coreNLP maps the quantitative tokens as adverbial Quantifier Phrases (QP) and Cardinal Numbers (CD), more level of detail is required for inferring the semantics of text. Such as, range, minimum, maximum, duration, and frequency. *Preclude* addresses these cases.

It should be noted that the semantic rules described in sections IV-A and IV-B are **novel**, **customized** to health advice statements, and guided by linguistic patterns. They enable *Preclude* to extract the semantics of advice statements and accurately detect the conflicts as demonstrated in Section V.

C. Phase 3: Assigning Polarity to Action and Effect Tokens

Polarity of an action/effect in an advice indicates whether the set of objects is encouraged or discouraged. Polarity of actions is assigned by building a customized lexicon of verbs from the training data and extending it by using verb synset from WordNet [18]. The initial positive and negative lists developed from the training data contain 18 and 21 verbs, respectively. After extending the lists using WordNet, the positive and negative lists contain 152 and 153 verbs,

[4]In case of intransitive verbs (e.g., run, exercise), often there is no object in the sentence. Then we use verbs to detect conceptual overlap.

978-1-5090-4327-9/17/$31.00 ©2017 IEEE

Precondition	Rule	Resulting Conflict
(1-4) Same object	$a_i \neq a_j$	Direct Conflict i.e., Opposite Polarity
	$a_j = null, a_i \neq e_j$	
	$a_i = null, e_i \neq a_j$	
	$a_i = a_j = null, e_i \neq e_j$	
(5-7) Direct Conflict	$c_i \neq null, c_j \neq null,$ c_i and c_j are not mutually exclusive	Conditional Conflict
	$c_i = null, c_j \neq null$	
	$c_i \neq null, c_j = null$	
(8-10) Direct Conflict	$t_i \neq null, t_j \neq null$ t_i and t_j are not mutually exclusive	Temporal Conflict
	$t_i = null, t_j \neq null$	
	$t_i \neq null, t_j = null$	
(11-13) Direct Conflict	$s_i \neq null, s_j = null$	Sub-typical Conflict
	$s_i = null, s_j \neq null$	
	$s_i \neq null, s_j \neq null, s_i \neq s_j$	
(14) Same Polarity	$f_i \neq f_j$, $s_i = null, s_j = null$	Quantitative Conflict
(15) Same Polarity	unit(n_i) = unit(n_j), $n_i \neq n_j$	Quantitative Conflict

TABLE III: Rules for detecting conflicts between advice A_i $<s_i^m$, o_i^m, a_i^m, q_i^m, c_i^m, t_i^m, $e_i^m>$and advice A_j $<s_j^n$, o_j^n, a_j^n, q_j^n, c_j^n, t_j^n, $e_j^n>$. The superscripts are dropped for the sake of simplicity.

respectively. Then, for each action found in the test data, it is labeled as *positive* or *negative* based on it's appearance in the positive verb list or negative verb list. If the action does not appear in any list, then the polarity is labeled as *null*. In that case, polarity is assigned to the corresponding effect clause.

The default polarity of an effect is positive. A negative effect is denoted by two patterns in the *effect* clauses. Firstly and more commonly, a negative effect is denoted by <Verb Phrase (VP), Noun Phrase (NP)> tuple, where NP is a disease, syndrome, or an unhealthy content (e.g., high calorie, trans fat, salt) and VP is a verb phrase that causes that NP. These specific <VP, NP> tuples are denoted as negative markers. Customized lexicons are built from training data and MetaMap to identify presence of negative markers. Secondly, a negative effect is also denoted by negation of verb/adjective phrases (e.g., not safe). Similar to assigning polarity to action, we build a customized lexicon of verbs and adjectives from the training data and extend it using synsets from WordNet. If any of the two aforementioned patterns is found in an effect clause, then it's polarity is *negative*.

It should be noted that although there are several existing lexicons of positive and negative words, using them results in performance deterioration in our case. Because, empirical observation confirms that this problem demands **domain specific lexicons** (Section V-D). For example, VPs such as *cause, lead to* are found to have *negative* polarity in our training data, while in traditional settings they are *neutral*.

D. Phase 4: Conflict Detection Among Pairs of Advice

After assigning polarity to the semantic tokens of an advice, the problem is reduced to mapping the token sets to the potential cases of contradiction presented in Section II. A set of rules is developed corresponding to each case as presented in Table III. Upon detecting conceptual overlap from the semantic tokens and assigning polarity, these rules are executed. The temporal order of rule execution is as follows.

Firstly, it is checked whether the polarity of the two advice statements are opposite (rules 1-4). If they are opposite, then it is a direct conflict.

Secondly, upon detecting a direct conflict further rules are executed to check whether this conflict can be refined (lines 1.21-1.23 of **Algorithm 1**). Thus, rules for conditional conflict (rules 5-7), temporal conflict (rules 8-10), and sub-typical conflict (rules 11-13) are executed in parallel. It should be noted that a conflict can satisfy multiple rules simultaneously, e.g., a conflict can be conditional as well as sub-typical.

Thirdly, if the polarity of the two advice are the same (i.e., none of the rules 1-4 holds), then the quantitative tokens of the overlapping object(s) are checked for quantitative conflicts. These conflicts occur when two advice statements have the same polarity about a common object but differ in terms of quantity of the common object. The difference in quantity can be caused by adverbial quantifiers (e.g., few, more) with opposite polarity (e.g., one advice suggests to *eat more kale* while the other suggests to *take less kale to mitigate side effects of a medication*). Such cases are handled by rule 14. In addition, the difference in quantity can also be caused by numerical mismatch (e.g., case 6 of Table I). A pair of advice with the same polarity can be numerically conflicting if the following two conditions hold: (i) both of their quantitative tokens of the common object are numerical with the same unit and (ii) the values of the quantitative tokens are not compatible (i.e., unequal or have different ranges) (rule 15). Currently, *Preclude* does not handle the case of numerical quantitative tokens with different units.

E. Handling Multiple Sentences

One distinguishing factor between detecting conflicting advice and detecting textual contradiction is the length of the sentences considered. In traditional contradiction detection literature, the pair of text under consideration have the same length, i.e., each of the texts contains a single sentence. But, a pair of potentially conflicting advice statements often have different number of sentences. Different sentences in an advice statement can convey different information as presented below.

(i) A pair of consecutive sentences often contain an action-effect tuple where one sentence contains a suggested action and the other contains the resulting effect(s) of the action. (ii) An action suggested in one sentence is often explained in further detail in subsequent sentence(s). (iii) An action discouraged in one sentence is often followed by one or more sentences containing alternate action(s). (iv) Consecutive sentences often suggest different actions with no common objects.

For the first two cases the semantic tokens are merged, as the sentences suggest the same action. In other cases, each sentence results in a separate tuple of semantic tokens. Thus, *Preclude* handles multiple sentences by following linguistic intuition derived from the textual health advice domain.

V. Evaluation

In this section, at first we demonstrate how the evaluation data is annotated for ground truth. Then, the performance of different components of *Preclude* is evaluated. Next, *Preclude* is compared with a baseline method. The effect of context awareness in *Preclude* is demonstrated in Section V-E. In addition, results from a survey is reported in Section V-F to present how people perceive conflicts in health advice.

978-1-5090-4327-9/17/$31.00 ©2017 IEEE

A. Ground Truth Annotation

For evaluation purposes, we split the dataset of 1156 advice statements into training and testing sets with 380 and 776 advice statements, respectively. We empirically develop the semantic decomposition rules and the conflict detection rules from the training set and evaluate the effectiveness of these rules on the test set. Potential candidate pairs in training and test sets are about (380^2=) 144K and (776^2=) 602K. Among these pairs, conflicts from cases 1-6 of Table I occur if there is at least one common topic/object between the pair of advice statements. Conflicts of case 7 of Table I may occur even if there is no common object. It also requires the advice statements to be temporally co-located (i.e., advice should be provided within a certain time lapse). We are not detecting the conflicts from case 7 in this work, as it requires additional physiological data of a user and accurate modeling of the effects of different health interventions that are currently unavailable. So, for efficient ground truth annotation, we filter advice pairs that do not have any common object as follows.

validated pairs	1294
pairs with gold label	1266
% of pairs with gold label	97.8
Number of conflicts (combining test and training set)	
Direct conflict	364
Refined conflict	624
Not conflict	306
Fleiss κ	
Direct conflict	0.977
Refined conflict	0.982
Not conflict	0.970
Overall	0.977

TABLE IV: Statistics for the validated pairs. A gold label reflects a consensus of three votes from the three annotators.

For labeling ground truth, objects are manually extracted from each advice by 3 human annotators. Each object of a sentence is labeled as one of 3 classes: positive, negative, and neutral. Conflicts of cases 1-5 of Table I occur if the polarity of the two advice statements with respect to the object is opposite. The other case is quantitative conflict, which occurs when there is at least one common object with the same polarity and the quantitative tokens are incompatible. 336 and 830 pairs of potentially conflicting advice statements are found from the training set and test set, respectively. Each of these pairs has at least one common object with opposite polarity. An additional 128 pairs of advice statements are found as potential candidates for quantitative conflicts. Each of these pairs has at least one common object with the same polarity and each advice of the pair contains a numerical/adverbial quantifier corresponding to that object.

Finally, the filtered (336+830+128)=1294 pairs are annotated by 3 human annotators. The statistics of this annotation is presented in Table IV. Here the *refined* class refers to the temporal, quantitative, sub-typical, and conditional conflicts. Among the 1294 pairs of advice, 364 have direct conflicts, 624 have refined conflicts, and 306 are not conflicting. Here, out of 1294 validated pairs, 1266 pairs obtain gold label (i.e., all three annotators agree on the label). The agreement among annotators are calculated using Fleiss κ statistics [32]. κ is scaled between 0-1 and a higher value of κ indicates higher inter-annotator agreement.

	Accuracy
Temporal Clause Extraction	90%
Conditional Clause Extraction	95%
Effect Clause Extraction	88%
Object Extraction	97%
Action Extraction	87%
Quantitative Token Extraction	85%
Polarity Assignment	94%

TABLE V: Accuracy of different components of *Preclude* .

B. Performance of Different Components of Preclude

In this section we measure the performance of *Preclude* on the test data. At first we measure the performance of different components of *Preclude* as presented in Table V. The ground truth for each token/clause is manually annotated. We measure the performance of token/clause extraction in terms of accuracy. At first, we measure the accuracy of detecting semantic clauses. Accuracy of temporal clause extraction is 90%. Accuracy of detecting conditional clauses is 95%, as most of the indicators of conditional clauses in the test set are present in the training set as well. The structure of effect clause varies widely as discussed in Section IV. The accuracy of effect clause extraction is 88%.

The accuracy of object token extraction includes the accuracy of sub-type token extraction, as a sub-type token is part of an object. Object extraction achieved an accuracy of 97%. This is because (i) MetaMap is customized to filter irrelevant objects, and (ii) the training set was a balanced representation of the test set in terms of object relation patterns. Accuracy of action token extraction is 87%. Although action extraction has fewer challenges than object extraction, more error is introduced here from parsing (System error) as the parser generated wrong labels for some of the verbs in the test set. As mentioned earlier, quantitative tokens include numerical as well as adverbial and adjective quantifiers. We find the overall accuracy of detecting different types of quantity tokens is 85%. In this case, the lexicon collected from training data was extended by adding synonyms and antonyms. However, our approach missed some unit tokens and thus resulted in comparatively lower accuracy in token extraction. Finally, the overall accuracy of polarity assignment is 94%.

Conflict types	Total Number of actual conflicts	Number of detected conflicts	Recall
Direct Conflict	254	228	0.90
Conditional Conflict	182	173	0.95
Temporal Conflict	97	78	0.80
Sub-typical Conflict	239	227	0.94
Quantitative Conflict	39	29	0.74
Numerical Conflict	19	15	0.79

TABLE VI: Total number of different types of conflicts and recall of detecting those conflicts in the test set. It should be noted a pair of advice can have multiple conflicts.

C. Performance of Conflict Detection

We present the performance of *Preclude* across different classes of conflicts in Table VI. Direct conflicts (i.e., conflicts corresponding to rules (1-4) in Table III) are detected with 0.9

978-1-5090-4327-9/17/$31.00 ©2017 IEEE

recall. The rest are the refinements of direct conflicts (rules (5-15) in Table III). As extracting conditional clauses receives high accuracy, conditional conflicts are detected with 0.95 recall. Recall of detecting temporal and sub-typical conflicts are 0.80 and 0.94, respectively. In the last two rows of Table VI, recall of detecting two types of quantitative conflicts are shown. Recall of detecting conflicts due to adverbial quantifier and numerical quantifier mismatches are 0.74 and 0.79, respectively. As our approach is a pipeline approach, error from the quantitative clause extraction is propagated to the later phase (i.e., quantitative conflict detection).

D. Comparison with a Baseline

Considering the health application domain, the most relevant work is presented in [7] by Alamari et al., where they focus on finding contradictory claims from abstracts of medical research papers. They group the research claims together based on the topic of the claims. Claims within a group are labeled as YES or NO to denote the polarity of the proposition of a claim. Thus they reduce the problem to binary classification of claims from the same group as YES or NO, where claims from different classes in a group indicate a conflict. For classification they used unigram, bigram, sentiment, directionality (e.g., increase vs. decrease) and negation features. Unlike us, they took a statistical approach to learning features.

A binary classification is performed for baseline compatibility. As the authors in [7] consider direct contradictions only, we use only the direct conflicts from our dataset to compare *Preclude* with the baseline method. In their evaluation they used linear Support Vector Machine (SVM) for classification. We try both linear and polynomial SVM (while varying the cost parameter 20 times ranging from 0.1 to 10) and report the best results only. For both baselines linear SVM outperforms polynomial SVM. Three standard performance metrics of classification are used here, namely, precision, recall, and F1 score (i.e., harmonic mean of precision and recall) [33]. The results are shown in Table VII.

Also in [7], they create directionality, sentiment and negation lexicon sets from the training data and use them as features for classification. Two versions of the baseline method are compared against *Preclude*. In *Baseline1*, the original negation, sentiment, and directionality lexicon sets used in [7] are used. It results in very low recall. In *Baseline2*, additional negation, sentiment, and directionality lexicon sets constructed from our training data are combined with their original lexicon set. This results in significant increase in recall and F1 scores from Baseline1 to Baseline2. This implies the significance of using a **domain adapted lexicon set**.

Method	Accuracy	Precision	Recall	F1
Baseline1	58%	0.52	0.10	0.17
Baseline2	60%	0.63	0.21	0.31
Preclude	90%	0.85	0.93	0.89

TABLE VII: Comparing our proposed solution with baseline methods: *Preclude* increase accuracy and F1 of Baseline2 by about 1.5 times and 3 times, respectively.

Preclude increases the accuracy, recall, and F1 score of *Baseline2* by 1.5 times, 4.5 times, and 3 times, respectively. This is because finding conflicts in health advice requires linguistic semantics that are not used in the baseline method. *Preclude* captures these semantics through semantic decomposition and heuristics developed from the training data. Also, the statistical method requires a larger amount of training data to reduce the sparsity of feature space.

E. Effect of Context Awareness

So far we have reported potential conflicts that can occur in any condition. In this experiment, we demonstrate the effectiveness of **context awareness** by using metadata collected from health apps. Specifically, we compare the number of conflicts between *Food Shopping Essential* (an app with context metadata) and a set of 4 other apps covering 4 health topics. These 4 apps are Effective Weightloss Guide (for weight loss), Pregnancy, Health & Nutrition Guide, and Anemia Help (Table II). We consider 3 different contexts covered in the *Food Shopping Essential* app and show how the number of potential conflicts changes based on the contexts in Table VIII. It contains the number of actual conflicts and the number of detected conflicts under different contexts.

For example, when considering interactions between *Food Shopping Essential* and *Effective Weightloss Guide* app (row 1, Table VIII), number of actual conflicts are 17, 12, and 6 for lactose, histamine, and gluten intolerance, respectively. If none of these contexts are considered, the number of conflicts between *Food Shopping Essential* and *QuickWeight* app are 49, almost 8 times as high as the number of conflicts under the context of gluten intolerance. *Preclude* can detect 14, 8, and 6 conflicts for lactose, histamine, and gluten intolerance, respectively. In the following three rows, variation in potential number of conflicts based on context awareness is shown for *Pregnancy, Health & Nutrition Guide* and *Anemia* apps. These results demonstrate how the number of conflicts can vary widely based the physiological context. Also, some contexts are more prone to conflicts, e.g., lactose intolerance is more conflicting with other apps than gluten intolerance.

	Food Shopping Essential app			
	With Context			Without Context
App Topics	Lactose intolerance	Histamin intolerance	Gluten intolerance	
Effective Weightloss Guide	14/17	8/12	6/6	**41/49**
Pregnancy	7/9	2/2	2/3	**15/18**
Health & Nutrition Guide	13/19	3/5	3/3	**25/33**
Anemia Help	4/5	1/1	1/1	**8/9**

TABLE VIII: Effect of context awareness in *Preclude* : number of conflicts vary considerably based on the context. Each cell contains the number of detected conflicts and the number of actual conflicts. Such as, in case of lactose intolerance, the numbers of detected and actual conflicts between *effective weightloss guide* and *food shopping essential* apps are 14 and 17, respectively.

F. Survey

A controlled survey is conducted with 24 participants (16 males and 8 females) to assess whether people who are not engaged in medical professions can detect conflicting health advice. The participants are selected using convenience sampling. The mean age of the participants is 32. 22 out of

978-1-5090-4327-9/17/$31.00 ©2017 IEEE

24 participants are engineering graduate students. Half of the participants identified health web sites as their primary source of health information. One third of the participants reported to have multiple health apps installed in their phones that they use daily or weekly. Although the survey participants do not fully represent the potential target population who may come across conflicting health advice from health apps and websites (e.g., chronic disease patients, elderly people), the goal of the study was to find whether healthy adults who are not in medical profession and posses reasonable cognitive ability can detect pairs of conflicting health advice statements.

The survey is composed of ten sections where each section presents a pair of advice and three propositions corresponding to the pair of advice. The propositions are: (i) *For the above pair of advice it was clear to determine what each advice suggested*, (ii) *The pair of advice presented above are conflicting*, and (iii) *For this set of advice it was easy to determine whether there was any conflict*. A 5-class Likert scale is used for collecting responses on the propositions, where the available options are: *strongly disagree, disagree, neutral, agree*, and *strongly agree.* The ten pairs of advice presented in the survey are selected from our annotated dataset. These pairs achieved gold label annotations from the three annotators. Among the ten pairs, nine are labeled as conflicting. The participants were allowed to search online for further information to understand an advice or detect a conflict (although only 6 participants reported to do so). Some key insights derived from the survey are presented below:

Ability to understand advice and detect conflict: On average, only 7.5% of the time people reported not understanding a pair of advice while the average error rate is 38%. This implies people are often unaware of the mistakes they make in interpreting advice, i.e., although they reported the pair of advice to be understandable, they actually missed subtle clues (e.g., quantitative difference, sub-type of an object) that indicate conflicts. Thus they could not detect the conflict between a pair of advice. While the participating population was healthy and mostly aged between 30-35, elderly people or people with less mental agility may be more susceptible to such errors [34], [35].

Projecting to real scenario: In the survey, for the sake of convenience of the participants a conflicting pair of advice is presented at a time, i.e., they only had to handle only 2 advice statements at a time. Although this is a controlled setting, there was 38% error. In reality, advice statements may come intermittently over time and in a non-consecutive manner. Also, the number of potential candidates of conflicting advice can be dozens or more. In such cases, detecting conflicts will require humans to recall all previous advice statements, have high attention span, and possess high level of domain knowledge. As these are scarce resources, specially for elderly people or people suffering from one or more chronic diseases [8], [34], [36], [35], the error rate in conflict detection by humans can increase in real scenarios. These results emphasize the importance of automatic conflict detection.

G. Interpretability of Preclude

In addition to conflict detection, *Preclude* also identifies the potential cases and conditions that cause conflicts. The goal is to make people aware of the nuances of conflicts so that they can make informed decision for conflict resolution. For example, in case of the second conflicting pair in Table I, *Preclude* points out the temporal condition (i.e., during pregnancy) and sub-type (i.e., Pate made of meat) as the causes of the conflicts. At this time *Preclude* reports potential conflicts and the reasons of the conflicts. In the future it may be possible to assess when conflicts are not important enough to be presented, thereby reducing the amount of conflicting information to only when necessary. For example, the severity of a conflict can sometimes be inferred using the effect clause, its polarity, and the user context.

VI. Related Works

Although we are the first to define and solve the problem of detecting conflicting health advice, there are some relevant research in Natural Language Processing (NLP) and human-in-the-loop Cyber Physical Systems (CPSs).

A. Textual Contradiction Detection

A relevant research topic in NLP is *textual entailment*, where the goal is to determine whether a given text fragment follows from another given text fragment. In existing NLP research, textual contradiction is usually defined as negative entailment. Given two pieces of text, they can be either textually entailed, contradictory, or neutral. However, most of the existing works formulate the textual contradiction detection as a binary classification task to distinguish the contradictory pairs from the non-contradictory pairs [15], [24], [16].

De Marneffe et al. provide a comprehensive taxonomy of contradiction in text that can be detected from linguistic evidence (e.g. negation, antonym, and structural or lexical disagreements) [15]. To solve the problem, they adapt a feature extraction based supervised technique where the features represent different linguistic notions of contradictions, e.g., negation, antonym, numerical mismatch, opposite polarity, etc.. In another work, the authors compare the structural similarity of two sentences and detect contradiction based on minimum alignment cost between the pair [24]. A limitation of these works is they do not utilize any external knowledge which plays a vital role in detecting true contradictions as pointed by the authors in [16], [37].

In contrast to previous works, Ritter et al. perform contradiction detection on automatically extracted web data [16] and demonstrate the importance of using external knowledge base for accurate contradiction detection. They take a functional relation based approach where they extract a relation between the subject and the object of a sentence. However, they overlook contradictions caused by negation, numerical mismatch. Also, the accuracy of detecting contradictions largely depends on the accuracy of detecting whether a phrase is functional. But in reality, contradictory phrases are not always functional. A recent relevant research is identifying potentially contradictory claims from medical research papers [7] as presented in Section V-D.

Although the existing methods utilize several important linguistic features for detecting textual contradiction, they have limited applicability in conflict detection from textual health advice. Because, **none** of the existing works provide

978-1-5090-4327-9/17/$31.00 ©2017 IEEE

appropriate taxonomy of conflicts that can arise while running multiple medical apps. For example, none of the existing works defines the conflicts caused by the **cases 2-5, 7** of Table I as conflicts.

B. Conflict Detection in Human Centric CPS Apps

Munir et al. focus on detecting dependencies across interventions generated by different human-in-the-loop apps (e.g., health apps, safety app) [38]. Unlike our work, they use simulated apps and structured metadata from each app. Metadata contain (i) interventions performed by each app and (ii) corresponding potential physiological parameters that might be affected by each intervention. They rely on HumMod, a physiological simulator [39], to approximate the potential effects of an intervention. HumMod uses over 7800 variables to capture cardiovascular, respiratory, renal, neural, endocrine, skeletal muscle, and metabolic physiology. But HumMod can simulate the effects of only a small set of interventions, such as, effects of only 4 drugs, effects of 2 types of exercises, and effects of taking basic nutrients (e.g., carbohydrates, protein, etc.). Also, currently it estimates the potential effects of an intervention only for a 37 year old healthy male whose weight and height are 159 lbs and 70.1 inches, respectively. It can not be personalized to any other age, gender, height, weight. Also, it does not consider the user's context (e.g., disease, physiological condition). Thus the current capability of the simulator is limited. On the other hand, *Preclude* focuses on detecting conflicts in textual health advice using external knowledge bases and linguistic features. Although *Preclude* currently does not detect cumulative effect conflicts (case 7 of Table I), it can be extended to detect such conflicts using advanced version of HumMod or similar sophisticated simulators that can model effects of more health interventions in a personalized manner.

VII. Discussion

Although, in this work we evaluate the effectiveness of our solution using only health advice data on 8 topics, the solution approach is **generalizable** to advice/interventions from other health topics and advice from other domains, including, but not limited to, intervention from smart home/smart city, advice from doctor's prescription, advice from drug description, etc. Because, the underlying linguistics rules are generalizable to advice data that has similar action-effect semantics.

Being a rule based system, the functionality of *Preclude* is limited by the *effectiveness* of training data and richness of the knowledge bases used. But, the alternative feature based solution does not perform well in the health domain (demonstrated by the baseline method) as they require large amount of annotated data which is currently unavailable. In addition, the context awareness of *Preclude* is based on explicit metadata from apps and advice. With the rapid growth of smart platforms to track physiological parameters (e.g., FitBit, Smart watch), in the near future more context and personalization data can be incorporated.

VIII. Conclusion

In this work, we develop *Preclude*, a semantic rule based system to detect conflicts in health advice in a comprehensive and context aware manner using linguistic inference and external knowledge bases. Detecting conflicts in health advice poses syntactic as well as semantic challenges. The syntactic challenges include large variation in structure and length of a pair of advice, while the semantic challenges include detecting conceptual overlap between a pair of advice statements and inferring the meaning of an advice. To address the syntactic challenges, *Preclude* decomposes a given advice statement using a set of linguistic rules that are learnt empirically based on linguistic references. To address the semantic challenges, it utilizes semantic decomposition of advice and multiple rich external knowledge bases, e.g., MetaMap, WordNet, Wikipedia. We create and release an annotated advice dataset containing 1156 advice statements on 8 health topics from several real health apps and websites. Our thorough evaluation using this dataset demonstrates the effectiveness of *Preclude* in detecting potential conflicts in textual health advice. In addition to detecting conflicts with 90% accuracy, *Preclude* also provides refined information about the potential causes of the conflicts to aid informed decision making for conflict resolution.

IX. Acknowledgement

This paper was supported, in part, by NSF CNS-1527563 and CNS-1646470. We would like to thank our reviewers for their insightful comments.

References

[1] B. Stiller, T. Bocek, F. Hecht, G. Machado, P. Racz, and M. Waldburger, "StartUp Health Insights: Digital Health Funding Rankings Q3 2014," Tech. Rep., 2014. [Online]. Available: http://www.startuphealth.com/content/insights-2014q3

[2] D. Witters and S. Agrawal, "How mobile technology can improve employees' well-being," November 2014. [Online]. Available: http://www.gallup.com/businessjournal/179111/mobile-technology-improve-employees.aspx

[3] K. Hämeen-Anttila, H. Nordeng, E. Kokki, J. Jyrkkä, A. Lupattelli, K. Vainio, and H. Enlund, "Multiple information sources and consequences of conflicting information about medicine use during pregnancy: a multinational internet-based survey," *Journal of medical Internet research*, vol. 16, no. 2, p. e60, 2014.

[4] "Health fact sheet— pew research center," 2013. [Online]. Available: http://www.pewinternet.org/fact-sheets/health-fact-sheet/

[5] P. Krebs and D. T. Duncan, "Health app use among us mobile phone owners: a national survey," *JMIR mHealth and uHealth*, vol. 3, no. 4, 2015.

[6] D. Kienhues, M. Stadtler, and R. Bromme, "Dealing with conflicting or consistent medical information on the web: When expert information breeds laypersons' doubts about experts," *Learning and Instruction*, vol. 21, no. 2, pp. 193–204, 2011.

[7] A. Alamri and M. Stevensony, "Automatic identification of potentially contradictory claims to support systematic reviews," in *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on*. IEEE, 2015, pp. 930–937.

[8] D. Garlan, D. P. Siewiorek, A. Smailagic, and P. Steenkiste, "Project aura: Toward distraction-free pervasive computing," *IEEE Pervasive computing*, vol. 1, no. 2, pp. 22–31, 2002.

[9] T. Okoshi, J. Ramos, H. Nozaki, J. Nakazawa, A. K. Dey, and H. Tokuda, "Attelia: Reducing user's cognitive load due to interruptive notifications on smart phones," in *Pervasive Computing and Communications (PerCom), 2015 IEEE International Conference on*. IEEE, 2015, pp. 96–104.

[10] A. Lau and E. Coiera, "Impact of web searching and social feedback on consumer decision making: a prospective online experiment," *Journal of medical Internet research*, vol. 10, no. 1, p. e2, 2008.

[11] J. A. Diaz, R. A. Griffith, J. J. Ng, S. E. Reinert, P. D. Friedmann, and A. W. Moulton, "Patients' use of the internet for medical information," *Journal of general internal medicine*, vol. 17, no. 3, pp. 180–185, 2002.

[12] P. Bagade, A. Banerjee, and S. K. Gupta, "Rapid evidence-based development of mobile medical iot apps," in *2016 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*. IEEE, 2016, pp. 1–6.

[13] A. Banerjee and S. K. Gupta, "Your mobility can be injurious to your health: Analyzing pervasive health monitoring systems under dynamic context changes," in *Pervasive Computing and Communications (PerCom), 2012 IEEE International Conference on*. IEEE, 2012, pp. 39–47.

[14] M. Rabbi, M. H. Aung, M. Zhang, and T. Choudhury, "Mybehavior: automatic personalized health feedback from user behaviors and preferences using smartphones," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 2015, pp. 707–718.

[15] M.-C. De Marneffe, A. N. Rafferty, and C. D. Manning, "Finding contradictions in text." in *ACL*, vol. 8, 2008, pp. 1039–1047.

[16] A. Ritter, D. Downey, S. Soderland, and O. Etzioni, "It's a contradiction—no, it's not: a case study using functional relations," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2008, pp. 11–20.

[17] A. R. Aronson and F.-M. Lang, "An overview of metamap: historical perspective and recent advances," *Journal of the American Medical Informatics Association*, vol. 17, no. 3, pp. 229–236, 2010.

[18] C. Fellbaum, "A semantic network of english verbs," *WordNet: An electronic lexical database*, vol. 3, pp. 153–178, 1998.

[19] H. J. Seabrook, J. N. Stromer, C. Shevkenek, A. Bharwani, J. de Grood, and W. A. Ghali, "Medical applications: a database and characterization of apps in apple ios and android platforms," *BMC research notes*, vol. 7, no. 1, p. 573, 2014.

[20] V. Obiodu and E. Obiodu, "An empirical review of the top 500 medical apps in a european android market," *Journal of Mobile Technology in Medicine*, vol. 1, no. 4, pp. 22–37, 2012.

[21] P. Lam, "Anemia: Causes, symptoms and treatments," 2015. [Online]. Available: http://www.medicalnewstoday.com/articles/158800.php

[22] "Nutrition digest: Digestive issues," 2015. [Online]. Available: http://americannutritionassociation.org/newsletter/digestive-issues

[23] "Top 15 most popular health websites — june 2016," 2016. [Online]. Available: http://www.ebizmba.com/articles/health-websites

[24] D. Andrade, M. Tsuchida, T. Onishi, and K. Ishikawa, "Detecting contradiction in text by using lexical mismatch and structural similarity," in *Proceedings of the 10th NTCIR Conference*, 2013.

[25] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*. Association for Computational Linguistics, 2003, pp. 173–180.

[26] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky, "The stanford corenlp natural language processing toolkit." in *ACL (System Demonstrations)*, 2014, pp. 55–60.

[27] J. Eastwood, *Oxford Guide to English Grammar*. Oxford University Press, 2003.

[28] A. X. Chang and C. D. Manning, "Sutime: A library for recognizing and normalizing time expressions." in *LREC*, 2012, pp. 3735–3740.

[29] "List of types of seafood," en.wikipedia.org/wiki/List_of_types_of_seafood, accessed: 2016-03-15.

[30] "List of vegetables," simple.wikipedia.org/wiki/List_of_vegetables, accessed: 2016-03-15.

[31] "List of grains," vegetablesfruitsgrains.com/list-of-grains/, accessed: 2016-03-15.

[32] J. L. Fleiss, "Measuring nominal scale agreement among many raters." *Psychological bulletin*, vol. 76, no. 5, p. 378, 1971.

[33] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1289–1305, 2003.

[34] M. L. Adams, L. A. Deokar, Angela J.and Anderson, and V. J. Edwards, "Self-reported increased confusion or memory loss and associated functional difficulties among adults aged 60 years 21 states," Center for Disease Control and Prevention, Tech. Rep., 05 2013.

[35] S. E. McDowell, H. S. Ferner, and R. E. Ferner, "The pathophysiology of medication errors: how and where they arise," *British journal of clinical pharmacology*, vol. 67, no. 6, pp. 605–613, 2009.

[36] J. A. Gazmararian, M. V. Williams, J. Peel, and D. W. Baker, "Health literacy and knowledge of chronic disease," *Patient education and counseling*, vol. 51, no. 3, pp. 267–275, 2003.

[37] C. Shih, C. Lee, R. T. Tsai, and W. Hsu, "Validating contradiction in texts using online co-mention pattern checking," *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 11, no. 4, p. 17, 2012.

[38] S. Munir, M. Y. Ahmed, and J. A. Stankovic, "Eyephy: Detecting dependencies in cyber-physical system apps due to human-in-the-loop."

[39] R. L. Hester, A. J. Brown, L. Husband, R. Iliescu, D. Pruett, R. Summers, and T. G. Coleman, "Hummod: a modeling environment for the simulation of integrative human physiology," *Frontiers in physiology*, vol. 2, 2011.

978-1-5090-4327-9/17/$31.00 ©2017 IEEE