Towards Contextual Healthiness Classification of Food Items - A Linguistic Approach

Michael Wiegand and **Dietrich Klakow**
Spoken Language Systems
Saarland University
D-66123 Saarbrücken, Germany
{Michael.Wiegand|Dietrich.Klakow}@lsv.uni-saarland.de

Abstract

We explore the feasibility of contextual healthiness classification of food items. We present a detailed analysis of the linguistic phenomena that need to be taken into consideration for this task based on a specially annotated corpus extracted from web forum entries. For automatic classification, we compare a supervised classifier and rule-based classification. Beyond linguistically motivated features that include sentiment information we also consider the prior healthiness of food items.

1 Introduction

Food plays a substantial part in each of our lives. With the growing health awareness in many parts of the population, there is consequently a high demand for the knowledge about healthiness of food. In view of the variety of both different types of food and nutritional aspects it does not come as a surprise that there is no comprehensive repository of that knowledge. Since, however, much of this information is preserved in natural language text, we assume that it is possible to acquire some of this knowledge automatically with the help of natural language processing (NLP).

In this paper, we take a first step towards this endeavour. We try to identify mentions that a food item is healthy (1) or unhealthy (2).

(1) There is not a healthy diet without a lot of fruits, vegetables and salads.

(2) The day already began unhealthy: I had a piece of cake for breakfast.

This task is a pre-requisite of more complex tasks, such as finding food items that are suitable for certain groups of people with a particular health condition (3) or identifying reasons for the healthiness or unhealthiness of particular food items (4).

(3) Vegetables are healthy, in particular, if you suffer from diabetes.
(4) Potatoes are healthy since they are actually low in calories.

The major problem of identifying some *Is-Healthy* or *Is-Unhealthy* relation is that the simple co-occurrence of a food item and the word *healthy* or *unhealthy* is not sufficiently predictive as shown in (5)-(7).

(5) Chocolate is *not* healthy.
(6) *The industry says* chocolate is healthy, but I guess this is just a marketing strategy.
(7) *If* chocolate is healthy, then I will run for the next presidential election.

We describe the contextual phenomena that underlie these cases and provide detailed statistics as to how often they occur in a typical text collection. From this analysis we derive features to be incorporated into a classifier.

Our experiments are carried out on German data. We believe, however, that our findings carry over to other languages since the aspects addressed in this work are (mostly) language universal. For the sake of general accessibility, all examples will be given as English translations.

To the best of our knowledge, this is the first work that addresses the classification of healthiness of food items using NLP.

2 Related Work

In the food domain, the most prominent research addresses ontology or thesaurus alignment (van Hage et al., 2010), a task in which concepts from different sources are related to each other. In this context, hyponymy relations (van Hage et al., 2005) and part-whole relations (van Hage et al., 2006) have been explored. More recently, Wiegand et al. (2012a) examined extraction methods for relations involved in customer advice in a supermarket. In Chahuneau et al. (2012), sentiment information has been related to food prices with the help of a large corpus consisting of restaurant menus and reviews.

In the health/medical domain, the majority of research focus on domain-specific relations involving entities, such as genes, proteins and

International Joint Conference on Natural Language Processing, pages 19–27,
Nagoya, Japan, 14-18 October 2013.

drugs (Cohen and Hersh, 2005). More recently, the prediction of epidemics (Fisichella et al., 2011; Torii et al., 2011; Diaz-Aviles et al., 2012; Munro et al., 2012) has attracted the attention of the research community. In addition, there has been research on processing healthcare claims (Popowich, 2005) and detecting sentiment in health-related texts (Sokolova and Bobicev, 2011).

3 The Dataset

In order to generate a dataset for our experiments, we used a crawl of *chefkoch.de*[1] (Wiegand et al., 2012a) consisting of $418,558$ webpages of food-related forum entries. *chefkoch.de* is the largest German web portal for food-related issues.

While we are aware of the fact that the healthiness of food items is also discussed in scientific texts we think that the text analysis on social media serves its own purpose. The language in social media is much more accessible to the general population. Moreover, social media can be considered as an exclusive repository of *popular wisdom* containing, for example, home remedies.

3.1 Healthiness Markers & Food Items

As it is impractical for us to manually label the entire web corpus with healthiness information, we extracted for annotation sentences in which there is a healthiness marker and a mention of a food item. By healthiness marker, we understand an expression that conveys the property of being healthy. Apart from the word *healthy* itself, we came up with 17 further common expressions (e.g. *nutritious*, *healthful* or *in good health*). Since the word *healthy* covers more than 95% of the mentions of healthiness markers in our entire corpus, however, we decided to restrict our healthiness marker exclusively to mentions of that expression. Thus, our main focus in this classification task is the contextual disambiguation, i.e. the task to decide whether a specific co-occurrence of the expression *healthy* and some food item denotes a genuine *Is-(Un)Healthy* relation.

The food items for which we extract co-occurrences with the healthiness marker *healthy* (Table 7) will henceforth be referred to as *target food items*. In order to obtain a suitable list of items for our experiments, we manually compiled a list of frequently occurring types of food.

3.2 “Unhealthy” vs. “Not Healthy”

In order to obtain instances that express an *Is-Unhealthy* relation, we exclusively consider negated instances of the *Is-Healthy* relation (8). We also experimented with a dataset with mentions of the word *unhealthy* (paired with our target food items) to extract instances such as (9).

(8) I am convinced that cake is *not healthy*.
(9) I am convinced that cake is *unhealthy*.

Using the same target food items, the *unhealthy*-dataset is, however, less than 14% of the size of the *healthy*-dataset. We also found that instances of the *Is-Unhealthy*-relation are not easier to detect on the *unhealthy*-dataset, since the *unhealthy*-dataset produced much poorer classifiers for detecting *Is-Unhealthy* relations than the *healthy*-dataset using negations as a proxy.

4 Annotation

Our final dataset comprises $2,440$ instances, where each **instance** consists of a sentence with the co-occurrence of some food item and the word *healthy* accompanied by the two sentences immediately preceding and the two sentences immediately following it.

The dataset was manually annotated by two German native speakers. On 4 target food items (this corresponds to 574 target sentences)[2] we measured an inter-annotation agreement of Cohen’s $\kappa = 0.7374$ (Landis and Koch, 1977) which should be sufficiently high for our experiments.

The annotators had to choose from a rich set of category labels that particularly divide the negative examples (i.e. those cases in which the co-occurrence of the target food item and *healthy* neither expresses an *Is-Healthy* nor an *Is-Unhealthy* relation) into different categories.

In the following, we describe the different category labels. Their distribution is shown in Table 1.

4.1 Is-Healthy Relation (HLTH)

This class describes instances in which there holds an *Is-Healthy* relation between the mention of *healthy* and the target food item (10).

(10) Potatoes are incredibly healthy, versatile in the kitchen and very tasty.

Table 1 shows that less than 20% of the co-occurrences of the target food item and *healthy* express this relation. This may already indicate that its extraction is difficult.

[1] www.chefkoch.de

[2] This is the only part of the dataset which was annotated by both annotators in parallel.

Type	Abbrev.	Frequency	Percentage
Is-Healthy	HLTH	488	20.00
Is-Unhealthy	UNHLTH	171	7.01
OTHER:			
No Relation	NOREL	788	32.30
Restricted Relation	RESTR	312	12.79
Unspecified Intersection	INTERS	198	8.11
Embedding	EMB	157	6.43
Comparison Relation	COMP	121	4.96
Unsupported Claim	CLAIM	87	3.57
Other Sense	SENSE	77	3.16
Irony	IRO	25	1.02
Question	Q	16	0.66

Table 1: Statistics of the different (linguistic) phenomena.

4.2 Is-Unhealthy Relation (UNHLTH)

We already stated in §3.2 that we consider negated instances (11) as instances for the *Is-Unhealthy* relation. We have a fairly broad notion of negation, e.g. (12) and (13) will also be assigned to this category. These *partial* negations are at least as frequent as *full* negations (11). However, we assume that the latter are often employed only as a means of being polite even though the speaker's intention is that of a full negation. The fact that we also observed fewer mentions of *unhealthy* co-occurring with a target food item than negated mentions of *healthy* would be in line with this theory (*unhealthy* is usually perceived to be more intense/blunter than *not healthy*).

(11) Chocolate is not healthy.
(12) Chocolate is not very healthy.
(13) Chocolate is hardly healthy.

4.3 Other Relations

Apart from the two target relations, we observe the following other relationships:

4.3.1 Restricted Relation (RESTR)

This category describes cases in which the *Is-Healthy* relation holds provided some additional condition is fulfilled. Typical conditions address a special kind of preparing the target food item (14) or make quantitative restrictions as to the amount of the target food item to be consumed (15). As such, one cannot infer from restricted relations to general properties of food items.

(14) Steamed vegetables are extremely healthy.
(15) A teaspoon of honey each day has been proven to be quite healthy.

4.3.2 Unspecified Intersection (INTERS)

In relation extraction, syntactic relatedness between the candidate entities of a relation is usually considered an important cue (Zhou et al., 2005; Mintz et al., 2009). In particular, the specific *type* of syntactic relation needs to be considered. If in our task *healthy* is an attributive adjective of the target food item (16), this is not an indication of a genuine *Is-Healthy* relation that we are looking for. With this construction, one usually refers to all those entities that share the two properties (*intersection*) of being the target food item and being healthy. This case is different from both *HLTH* (17) and *RESTR* (18).

(16) I usually buy the healthy fat.
(17) Fat is healthy.
(18) I usually buy the healthy fat, the one that contains a high degree of unsaturated fatty acids.

HLTH, typically realized as a predicative adjective (17), requires that this intersection of properties includes the *entire* set of entities representing the target food item. For both *RESTR* and *INTERS*, on the other hand, this intersection only includes a proper subset of the target food item. In addition, *RESTR* provides some (vital) additional information about this subset that allows it to be (easily) identified (e.g. the property of containing a high degree of unsaturated fatty acids in (18)). However, for *INTERS*, no further properties are specified in order to identify it – the information of being healthy is not telling as we actually want to find out how to detect healthy food. As a consequence, instances of type *INTERS* are hardly informative when it comes to answering whether a particular food item is healthy or not. We do not even know how large the proportion of the intersection with regard to the overall amount of the target food item is. It may well be extremely small. That is why in this work, instances of *INTERS* will neither be used as evidence for the healthiness nor the unhealthiness of a particular food item.

4.3.3 Comparison Relation (COMP)

If the target food item is compared with another food item with regard to their healthiness status (19) & (20), one cannot conclude anything regarding the *absolute* healthiness of the target food item. This is due to the fact that a comparison assumes healthiness as a (continuous) scale rather than a binary (discrete) property. It determines the positions of the two food items relative to each other on that particular scale.

(19) Honey is healthier than chocolate. (target food item: *honey*)
(20) Honey is as healthy as chocolate. (target food item: *honey*)

4.3.4 Unsupported Claim (CLAIM)

In our initial data analysis, we found frequent cases in which the author of a forum entry reports a (controversial) statement regarding the healthiness status of a particular food item. These claims are often used as a means of starting a discussion about that issue (21).

(21) Some people claim that chocolate is healthy. What do you make of it?

If it is not possible to infer from such reported statement that the reported view is shared by the author (and we found that this is true for many reported statements), we tag it as *CLAIM*.

4.3.5 Question (Q)

There may also be cases in which the *Is-(Un)Healthy* relation is embedded in a question (22).

(22) Is chocolate healthy?

4.3.6 Irony (IRO)

Irony (23) is a figure of speech that can frequently be observed in user-generated text (Tsur et al., 2010). With a proportion of less than 1%, this, however, does not apply for the forum entries that comprise our data collection.

(23) Everyone knows that sweets are healthy, in particular, chocolate with its many calories even makes you lose weight.

4.3.7 Embedding (EMB)

In addition to the previous categories *CLAIM* and *IRO*, there exist other ways of embedding the healthiness relation into a context so that the general validity of it is discarded. We introduce a common label for all those other remaining types that include, for instance, *modal embedding* (24) or *irrealis construction* (25).

(24) Honey could be healthy.
(25) If chocolate were healthy, people eating it wouldn't put on so much weight.

4.3.8 Other Sense (SENSE)

Both the target food item and the German healthiness cue *gesund* are (potentially) ambiguous expressions. For instance, *gesund* can be part of several multiword expressions, such as *gesunder Menschenverstand* (engl. *common sense*).

4.3.9 No Relation (NOREL)

While in all previously discussed cases the target food item and *healthy* are somehow related, there are cases in which the co-occurrence is merely coincidental (26).

(26) Tomatoes are very healthy and they can be ideally served on bread. (target food item: *bread*)

On our dataset, this is the most frequent label.

5 Feature Design

All features we use are summarized in Table 2 along examples. Apart from bag of words (*word*), we use following features:

5.1 Linguistic Features

The linguistic features are mainly derived from our quantitative data analysis in §4. Given the limited space of this paper, we will only point out some special properties.

The first group of (linguistic) features (Table 2) is designed to detect some relationship between target food item and *healthy*. The co-occurrence within the same clause is usually a good predictor. There are three features to establish this property: *clause*, *boundary* and *otherFood*.

We already pointed out in §4.3.2 that not only syntactic relatedness between *healthy* and the target food item as such but also the specific syntactic relation plays a decisive role for this task. The two most common relations are that *healthy* is a predicative adjective (of the target food item), which is usually indicative of *HLTH*, and that *healthy* is an attributive adjective (of the target food item), which is usually indicative of *INTERS* (on our dataset in more than 90% of the instances labeled with *INTERS* this is the case). This is reflected by the two features *predRel* and *attrRel* (and the back-off features *pred* and *attr*). An additional feature *attrFood* captures a special construction in which *healthy* as an attributive adjective actually denotes *HLTH* instead of *INTERS*.

For the conditional healthiness *RESTR* (§ 4.3.1), we found two predominant subcategories of restrictions: restrictions with regard to the quantity with which the target food item should be consumed (*quant*) and references to a specific subtype of the target food item, which we want to capture with a few precise surface patterns (*spec*) and a feature that checks whether the target food item precedes an attributive adjective (*attrNoH*).

Table 2 also contains features to detect various contextual embeddings (*opHolder*, *question*, *irrealis*, *modal* and *irony*). *opHolder* is to detect cases of *CLAIM*. We assume once some opinion holder other than the author of the forum post (i.e. 1st person pronoun) is identified, there is a *CLAIM*.

We also investigate whether *healthiness* correlates with *sentiment*. For instance, if the author promotes the healthiness of some food item, does this also coincide with positive sentiment (e.g.

tasty, *good* etc.)? Our features *positive/negative polar* check for the presence of polar expressions.

5.2 Knowledge-based Features using a Healthiness Lexicon

We also incorporate features referring to the prior knowledge of healthiness of food items. We use a lexicon introduced in Wiegand et al. (2012b) which covers approximately 3000 food items, and we refer to it as ***healthiness lexicon***. Each food item is specified as being either healthy or unhealthy in that lexicon. The healthiness judgment has been carried out based on the general nutrient content of each food item. A detailed description of the annotation scheme and annotation agreement can be found in Wiegand et al. (2012b).

The specific features derived from that lexical resource are listed in Table 2. They are divided into two groups. *prior* describes the prior healthiness of the target food item. Since our task is to determine the *contextual* healthiness, the usage of such a feature is legitimate. The *contextual* healthiness need not to coincide with the *prior* healthiness. For instance, in (27), *chocolate* is described as a healthy food item even though it is a priori considered unhealthy.

(27) Chocolate is healthy as it's high in magnesium and provides vitamin E.

We use this knowledge as a baseline. If we cannot exceed the classification performance of *prior* (alone), then acquiring the knowledge of healthiness with the help of NLP is hardly effective.

priorCont describes the prior healthiness status of *neighbouring food items* in the given context.

6 Rule-based Classification

We also examine rule-based classifiers since they can be built without any training data. Each classifier is defined by a (large) conjunction of linguistic features. Features indicating a class other than the target class are used as negated features in that conjunction. The rule-based classifiers only consider features where a positive or negative correlation towards the target class is (more or less) obvious. Table 3 shows the rule-based classifiers for each of our classes. For *HLTH*, it basically states that *healthy* has to be a predicative adjective of the target food item (*predRel*), and the target food item and *healthy* have to appear within the same clause (or there is no boundary sign between them). After that, a long list of negated features follows: *quant*, *spec* and *attrNoH*, for example, are negated because they are typical cues for *RESTR*. The remaining features are negated since they are either indicative of *UNHTLTH*, *COMP*, *EMB*, *CLAIM*, *SENSE*, *IRO* or *Q*. The classifier for *UNHLTH* only differs from *HLTH* in that either of the negation cues, i.e. *negTarget* or *negHealth*, has to be present.

HLTH	predRel ∧ (clause ∨ ¬boundary) ∧ ¬quant ∧ ¬spec ∧ ¬attrNoH ∧ ¬negTarget ∧ ¬negHealth ∧ ¬comp ∧ ¬opHolder ∧ ¬modal ∧ ¬irrealis ∧ ¬question ∧ ¬sense ∧ ¬weird
UNHLTH	predRel ∧ (clause ∨ ¬boundary) ∧ ¬quant ∧ ¬spec ∧ ¬attrNoH ∧ (negTarget ∨ negHealth) ∧ ¬comp ∧ ¬opHolder ∧ ¬modal ∧ ¬irrealis ∧ ¬question ∧ ¬sense ∧ ¬weird

Table 3: Rule-based classifiers based on linguistic features (Table 2).

7 Experiments

In this section we present the results on automatic classification.

7.1 Classification of Individual Utterances

In this subsection, we evaluate the performance of the different feature sets on sentence-level classification using supervised learning and rule-based classification. We investigate the detection of the two classes *HLTH* (§4.1) and *UNHLTH* (§4.2). Each instance to be classified is a sentence in which there is a co-occurrence of a target food item and a mention of *healthy* along its respective context sentences. The dataset was parsed using the Stanford Parser (Rafferty and Manning, 2008). We carry out a 5-fold cross-validation on our manually labeled dataset. As a supervised classifier, we use Support Vector Machines (*SVM*^{light} (Joachims, 1999) with a linear kernel). For each class, we train a binary classifier where positive instances represent the class to be extracted while negative instances are the remaining instances of the entire dataset (§4).

7.1.1 Comparison of Various Feature Sets

Table 4 lists the results for various feature sets that we experimented with. ***take-all*** is an unsupervised baseline that considers all instances of our dataset as positive instances (of the class which is examined, i.e. *HLTH* or *UNHLTH*). In other words, this baseline indicates how well the mere co-occurrence of *healthy* and the target food item predicts either of our two classes.[3] Our second

[3] Restricting the co-occurrence to a certain window size did not improve the F-Score of *take-all*.

Word-based Features		
Feature	**Abbrev.**	**Illustration/Further Information**
bag of words between the mention of *healthy* and target food item, and the additional words that precede or follow *healthy* and target food item	word	N/A
Linguistic Features		
Feature	**Abbrev.**	**Illustration/Further Information**
Are target food item and *healthy* within the same clause?	clause	*I like chocolate$_{target}$, even though I consider fruits the healthy option for snacks.* Feature operates on parse output.
Is there a punctuation mark between target food item and *healthy*?	boundary	*I know that vegetables are extremely healthy**;** but I prefer chocolate$_{target}$.* Token-level back-off feature to *clause*.
Is there another food item between target food item and *healthy*?	otherFood	*We always had healthy meals with lots of vegetables and salad, but this does not mean that we were not allowed to eat chocolate$_{target}$.* Token-level back-off feature to *clause*.
Is target food item in a prominent position?	prom	Prominent positions: e.g. beginning/end of a sentence/subclause.
Is target food item used as a side dish?	side	*Broccoli with potatoes$_{target}$ is a healthy dish.* Patterns from relation type *Served-with* used in Wiegand et al. (2012a).
Is *healthy* a predicative adjective relating to target food item?	predRel	*Vegetables are healthy.*
Is *healthy* an attributive adjective relating to target food item?	attrRel	*I would recommend buying some healthy fat.*
Is *healthy* a predicative adjective?	pred	*I really like bananas$_{target}$ and they are healthy, too.*
Is *healthy* an attributive adjective?	attr	*For that we need to use some kind of fat$_{target}$; I particularly favour the healthy ones.*
Does *healthy* precede target food item?	precede	If *healthy* precedes the target food item, then this often indicates *attributive* usage.
Is *healthy* an attributive adjective of a general food expression (i.e. *meal*, *dish*, *food*, etc.) that is not target food item?	attrFood	*Salad is a healthy dish.*
Is there some quantification?	quant	*100g per day*; *in moderation*; *a teaspoon of*; a list of 75 quantifying expressions was collected from the web (`rezepte.nit.at/kuechenmasse.html` and `de.wikibooks.org/wiki/Kochbuch/_Maßangaben`).
Is target food item modified by an attributive adjective other than *healthy*?	attrNoH	*steamed vegetables; fried potatoes*
Is target food item further specified?	spec	*bread$_{target}$ made of whole grains*; *cake$_{target}$ with low-fat ingredients*; Complementary feature to *attrNoH* (feature detects specifications in the form of contact clauses or prepositional phrases immediately attached to the target food item).
Is there a cue indicating an opinion holder other than the author?	opHolder	*Some people claim that chocolate is healthy.* This feature relies on a set of predicates indicating the presence of an opinion holder (Wiegand and Klakow, 2011).
Is target sentence a (direct) question?	question	*Is chocolate healthy?*
Is *healthy* embedded in some *irrealis* context?	irrealis	*If honey were healthy*; *I wonder, whether honey is healthy*. Translation of the cues used in hedge classification (Morante and Daelemans, 2009).
Is *healthy* modified by a modal verb?	modal	*Honey might be healthy.*
Is target food item negated?	negTarget	*No cake is healthy.* We adapted to German the negation word lists and the scope modeling from Wilson et al. (2005).
Is *healthy* negated?	negHealth	*Chocolate is not healthy.* We adapted to German the negation word lists and the scope modeling from Wilson et al. (2005).
Is there any occurrence of a *weird* word?	weird	*Sure, chocolate is veeeeery healthy.* Regular expression detecting suspicious reduplications of characters in order to detect irony.
Does the context suggest that *healthy* is part of a comparison?	comp	We check for typical inflectional word forms (i.e. *healthier* and *healthiest*) and constructions, such as *as healthy as*.
Does the context of *healthy* suggest another sense of the word?	sense	Contexts in which *healthy* has a different meaning (using online dictionaries, such as `www.duden.de/rechtschreibung/gesund` and `de.wiktionary.org/wiki/gesund`).
Number of positive/negative polar expressions (excluding mentions of *healthy*)	polar*	Usage of the German *PolArt* sentiment lexicon (Klenner et al., 2009).
Number of near synonyms of *(un)healthy*	syno*	Examples for healthy: *high in vitamin*, *tonic*, etc.; examples for unhealthy: *carcinogenic*, *harmful*, etc. (manually compiled list of 99 synonyms by an annotator **not** involved in feature engineering).
Number of diseases	disease*	411 entries, created with the help of the web (`bildung.wikia.com/wiki/Alphabetische_Liste_der_Krankheiten`).
Task-specific Knowledge-based Features using a Healthiness Lexicon		
Feature	**Abbrev.**	**Illustration/Further Information**
Is target food item *a priori* healthy? Is target food item *a priori* unhealthy?	prior*	Feature employs the healthiness lexicon from Wiegand et al. (2012b).
Number of food items (excluding target food item) that are *a priori* healthy Number of food items (excluding target food item) that are *a priori* unhealthy	priorCont*	Feature employs the healthiness lexicon from Wiegand et al. (2012b).

*: there exist two features which differ in the context they consider: (a) only target sentence (indicated by suffix *-TS*) (b) entire context (indicated by suffix *-EC*)

Table 2: Description of the feature set; the set contains several cue word lists, in order to **avoid overfitting**, we either translated existing resources from English or used diverse web-resources that are **not** related to our dataset.

Features	HLTH			UNHLTH		
	Pre	Rec	F1	Pre	Rec	F1
take-all (*baseline 1*)	20.3	**100.0**	33.7	6.9	**100.0**	13.0
prior (*baseline 2*)	28.0	87.3	42.3	29.7	44.0	35.3
priorCont	21.2	96.9	34.7	14.3	34.8	20.3
prior+priorCont	28.0	86.9	42.3	29.7	44.0	35.3
word	35.9	66.5	46.6	39.7	42.5	41.0
linguistic	38.3	66.1	48.3	35.9	43.5	39.1
word+linguistic	40.2	63.6	49.1^{*}	40.9	47.1	43.4^{*}
word+prior	38.1	70.1	49.2°	46.7	43.3	44.7
word+priorCont	35.0	65.3	45.5	40.0	42.9	41.0
word+prior+priorCont	37.4	70.8	48.8°	**46.8**	42.8	44.4
word+linguistic+priorCont	41.4	64.3	50.2	42.8	42.1	41.7
word+linguistic+prior	44.1	68.3	$53.3^{\circ\dagger\ddagger}$	44.8	60.5	$\mathbf{51.1}^{\circ\dagger\ddagger}$
all features	44.5	69.3	$\mathbf{53.9}^{\circ\dagger\ddagger}$	42.9	63.5	$51.0^{\circ\dagger\ddagger}$
rule-based	**53.4**	17.9	26.8	45.0	11.0	17.7

significantly better than *word* * at $p < 0.1/^{\circ}$ at $p < 0.05$; better than *word+linguistic* † at $p < 0.05$; better than *word+prior* ‡ at $p < 0.05$ (paired t-test)

Table 4: Comparison of different feature sets.

baseline is *prior* (see §5.2 for motivation).

take-all has optimal recall but a very poor precision. The second baseline *prior* is notably better. *prior* may help to distinguish between *HLTH* and *UNHLTH* but it does not contribute to distinguishing these classes from the rest of the relation types (Table 1).

If we turn to the features that largely exploit contextual information, i.e. *word* and *linguistic* (§5.1), we find that both features are better than the previous features. This is an indication that learning from text is effective. The same can be said about *word+linguistic* and *word+prior*, which also outperform *word*. *word+linguistic+prior* is the best feature set outperforming both *word+linguistic* and *word+prior*. We conclude that all of the three groups of features we presented in §5 are relevant for this task.

In terms of recall and F-score the supervised classifier always outperforms the rule-based classifier. This does not come as a surprise as the supervised classifier learns from labeled training data while the rule-based classifier is unsupervised. On the other hand, we also find that the precision of the rule-based classifier largely outperforms our best supervised classifier on *HLTH*.

The fact that the best overall F-score achieved is not higher may be ascribed to the heavy noise (spelling/grammar mistakes) contained in our web-data. However, we believe that even with those data we can show the relative effectiveness of the different feature types which is the most relevant aspect in our *proof-of-concept* investigation.

Class	Features
HLTH	prom, attrNoH, predRel, comp, negHealth, *negative* polarEC, sense, opHolder, irrealis
UNHLTH	negHealth, negTarget, attrRel, comp, diseaseTS, *negative* polarEC

Table 5: List of the best subset of linguistic features (Table 2) for each individual class.

7.1.2 Inspection of Linguistic Features

Table 5 shows the best performing feature subset using a best-first forward selection as implemented in *Weka* (Witten and Frank, 2005). The table shows that diverse features are important including features to detect restricted relations (§4.3.1) (i.e. *attrNoH*) or comparisons (i.e. *comp*), features to distinguish predicative from attributive adjectives for the detection of unspecified intersection (§4.3.2) (i.e. *predRel* and *attrRel*), various features to determine contextual embedding (i.e. *opHolder*, *irrealis* and *negHealth*) and sentiment information (i.e. *negative polarEC*).

7.1.3 Detecting Anti-Prior Healthiness

We now take a closer look at *anti-prior* instances which are utterances in which the relation expressed is opposite to the relation that one would *a priori* assume, e.g. *chocolate is healthy* instead of *chocolate is unhealthy*. In our gold standard, we identified these instances with the help of the actual (manually assigned) label and our healthiness lexicon (§5.2).[4] Such instances may be very interesting to extract, even though they are rare (15% on *HLTH* and *UNHTLH*). Previously, supervised classifiers with *word+prior* produced similar performance as classifiers with *word+linguistic* (Table 4). Since linguistic features are fairly expensive to produce, the prior knowledge of healthiness seems an attractive alternative. But this is misleading. Table 6 displays the recall (by supervised classification) on only anti-prior instances and shows that the usage of *prior* which, in isolation, would detect none of these instances, gives a much lower recall than *linguistic* when added to *word*. Therefore, *word+linguistic* would be the preferable feature set if one had to choose between *word+prior* and *word+linguistic*.

[4] Whenever HLTH co-occurs with prior unhealthiness (according to the healthiness lexicon) or UNHLTH co-occurs with prior healthiness, there is an anti-prior instance.

Feature Set	word+prior	word+linguistic
Recall	17.2	54.6

Table 6: Recall on *anti-prior* instances.

7.2 Aggregate Classification

Finally, we automatically rank food items according to healthiness based on the aggregate of text mentions. Ideally, the ranking should separate healthy from unhealthy food items. We want to know whether with our text corpus and contextual classification, one can actually approximate a correct prior healthiness. Aggregate classification means that we make a healthiness prediction for a specific food item based on *all* text mentions of that food item co-occurring with the word *healthy*. It may be easier to achieve a robust aggregate classification than a robust individual classification. This is because in aggregate-based tasks, there is a certain degree of redundancy contained in the data, as instances of a group of utterances (belonging to the same food item) may often comprise similar information. For such classifiers, one should focus on a higher precision since a reasonable recall is enabled by the redundancy in the data.

Our baseline ***RAW*** is completely unsupervised and does not include any linguistic processing. We use the *Pointwise Mutual Information (PMI)* which is estimated on our large web corpus (§3).[5]

$$PMI(food\,item, healthy) = log\frac{P(food\,item, healthy)}{P(food\,item)P(healthy)} \quad (1)$$

For the automatic classification, we consider ***LEARN*** which uses the output of the supervised classifier comprising the features *word+linguistic* (we ***must*** exclude the feature *prior* as this would include the knowledge we want to predict automatically in this experiment)[6] while ***RB*** is the output of the rule-based classifier we presented in §6 (which does not contain *prior* as a feature either).

In order to convert the classifications of individual utterances for a target food item (by *LEARN* and *RB*) to one ranking score (according to which we rank all the target food items), we simply compute the ratio between instances predicted to be healthy and those predicted to be unhealthy:

$$score_{LEARN/RB}(food\,item) = \frac{\#HLTH_{predicted}(food\,item)}{\#UNHLTH_{predicted}(food\,item)} \quad (2)$$

RAW	wholemeal product ≻ fat ≻ colza oil ≻ vegetables ≻ tea ≻ protein ≻ olive oil ≻ honey ≻ meat ≻ sugar ≻ salad ≻ bread ≻ chocolate ≻ potato ≻ rice ≻ banana ≻ cake ≻ water ≻ egg
LEARN	banana ≻ olive oil ≻ wholemeal product ≻ tea ≻ colza oil ≻ salad ≻ vegetables ≻ protein ≻ potato ≻ chocolate ≻ meat ≻ bread ≻ rice ≻ water ≻ sugar ≻ cake ≻ egg ≻ fat ≻ honey
RB	potato ≻ protein ≻ wholemeal product ≻ banana ≻ olive oil ≻ vegetables ≻ bread ≻ salad ≻ water ≻ tea ≻ colza oil ≻ rice ≻ honey ≻ egg ≻ chocolate ≻ fat ≻ meat ≻ sugar ≻ cake

Table 7: Aggregate ranking; green denotes (actual) healthy items, red (actual) unhealthy items.

where $\#HLTH_{predicted}(food\ item)$ are the number of instances the classifier predicts the label *HLTH* for the target food item while $UNHLTH_{predicted}(food\ item)$ are the number of instances labeled as *UNHLTH*, respectively.

Table 7 shows the results of the three rankings. The actual labels are derived from the healthiness lexicon (§5.2). The table clearly shows that the ranking produced by *RAW* contains most errors. *fat* is the second most highly ranked food item. This can be explained by the high proportion of *INTERS* (§4.3.2) among the co-occurrences of *fat* and *healthy* (almost 50%). *LEARN* and *RB* produce a better ranking, thus proving that a contextual (linguistic) analysis is helpful for this task. *RB* also outperforms *LEARN* presumably because of its much higher precision (as measured for individual classification in Table 4: 53.4% vs. 40.2% for *HLTH* and 45.0% vs. 40.9% for *UNHLTH*).

8 Conclusion

We presented a first step towards contextual healthiness classification of food items. For this task, we introduced a new annotation scheme. Our annotation revealed that many different linguistic phenomena are involved. Thus, this problem can be considered an interesting task for NLP. We demonstrated that a linguistic analysis is not only necessary for classifying individual utterances but also for ranking food items based on an aggregate of text mentions.

Acknowledgements

This work was performed in the context of the Software-Cluster project EMERGENT. Michael Wiegand was funded by the German Federal Ministry of Education and Research (BMBF) under grant no. "01IC10S01". The authors would like to thank Stephanie Köser and Eva Lasarcyk for annotating the dataset presented in this paper. We would also like to thank Benjamin Roth for interesting discussions.

[5] For $P(food\ item, healthy)$, we consider all *sentences* in which the target food item and *healthy* co-occur.

[6] We train for each target food item a classifier using only the instances with the other target food items as training data.

References

Victor Chahuneau, Kevin Gimpel, Bryan R. Routledge, Lily Scherlis, and Noah A. Smith. 2012. Word Salad: Relating Food Prices and Descriptions. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP/CoNLL)*, pages 1357–1367, Jeju Island, Korea.

Aaron M. Cohen and William R. Hersh. 2005. A survey of current work in biomedical text mining. *Briefings in Bioinformatics*, 6:57 – 71.

Ernesto Diaz-Aviles, Avar Stewart, Edward Velasco, Kerstin Denecke, and Wolfgang Nejdl. 2012. Epidemic Intelligence for the Crowd, by the Crowd. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*, Dublin, Ireland.

Marco Fisichella, Avar Stewart, Alfredo Cuzzocrea, and Kerstin Denecke. 2011. Detecting Health Events on the Social Web to Enable Epidemic Intelligence. In *Proceedings of the International Symposium on String Processing and Information Retrieval (SPIRE)*, pages 87–103, Pisa, Italy.

Thorsten Joachims. 1999. Making Large-Scale SVM Learning Practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, pages 169–184. MIT Press.

Manfred Klenner, Stefanos Petrakis, and Angela Fahrni. 2009. Robust Compositional Polarity Classification. In *Proceedings of Recent Advances in Natural Language Processing (RANLP)*, pages 180–184, Borovets, Bulgaria.

J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant Supervision for Relation Extraction without Labeled Data. In *Proceedings of the Joint Conference of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL/IJCNLP)*, pages 1003–1011, Singapore.

Roser Morante and Walter Daelemans. 2009. Learning the Scope of Hedge Cues in Biomedical Texts. In *Proceedings of the BioNLP Workshop*, pages 28–36, Boulder, CO, USA.

Robert Munro, Lucky Gunasekara, Stephanie Nevins, Lalith Polepeddi, and Evan Rosen. 2012. Tracking Epidemics with Natural Language Processing and Crowdsourcing. In *Proceedings of the Spring Symposium for Association for the Advancement of Artificial Intelligence (AAAI)*, pages 52–58, Toronto, Canada.

Fred Popowich. 2005. Using Text Mining and Natural Language Processing for Health Care Claims Processing. *SIGKDD Explorations*, 7(1):59–66.

Anna Rafferty and Christopher D. Manning. 2008. Parsing Three German Treebanks: Lexicalized and Unlexicalized Baselines. In *Proceedings of the ACL Workshop on Parsing German (PaGe)*, pages 40–46, Columbus, OH, USA.

Marina Sokolova and Victoria Bobicev. 2011. Sentiments and Opinions in Health-related Web Messages. In *Proceedings of Recent Advances in Natural Language Processing (RANLP)*, pages 132–139, Hissar, Bulgaria.

Manabu Torii, Lanlan Yin, Thang Nguyen, Chand T. Mazumdar, Hongfang Liu, David M. Hartley, and Noele P. Nelson. 2011. An exploratory study of a text classification framework for internet-based surveillance of emerging epidemics. *International Journal of Medical Informatics*, 80(1):56–66.

Oren Tsur, Dmitry Davidov, and Ari Rappoport. 2010. ICWSM - A Great Catchy Name: Semi-Supervised Recognition of Sarcastic Sentences in Online Product Reviews. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*, Washington, DC, USA.

Willem Robert van Hage, Sophia Katrenko, and Guus Schreiber. 2005. A Method to Combine Linguistic Ontology-Mapping Techniques. In *Proceedings of International Semantic Web Conference (ISWC)*, pages 732 – 744, Galway, Ireland. Springer.

Willem Robert van Hage, Hap Kolb, and Guus Schreiber. 2006. A Method for Learning Part-Whole Relations. In *Proceedings of International Semantic Web Conference (ISWC)*, pages 723 – 735, Athens, GA, USA. Springer.

Willem Robert van Hage, Margherita Sini, Lori Finch, Hap Kolb, and Guus Schreiber. 2010. The OAEI food task: an analysis of a thesaurus alignment task. *Applied Ontology*, 5(1):1 – 28.

Michael Wiegand and Dietrich Klakow. 2011. The Role of Predicates in Opinion Holder Extraction. In *Proceedings of the RANLP Workshop on Information Extraction and Knowledge Acquisition (IEKA)*, pages 13–20, Hissar, Bulgaria.

Michael Wiegand, Benjamin Roth, and Dietrich Klakow. 2012a. Web-based Relation Extraction for the Food Domain. In *Proceedings of the International Conference on Applications of Natural Language Processing to Information Systems (NLDB)*, pages 222–227, Groningen, the Netherlands. Springer.

Michael Wiegand, Benjamin Roth, Eva Lasarcyk, Stephanie Köser, and Dietrich Klakow. 2012b. A Gold Standard for Relation Extraction in the Food Domain. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, pages 507–514, Istanbul, Turkey.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing Contextual Polarity in Phrase-level Sentiment Analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pages 347–354, Vancouver, BC, Canada.

Ian Witten and Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers, San Francisco, US.

GuoDong Zhou, Jian Su, Jie Zhang, and Min Zhang. 2005. Exploring Various Knowledge in Relation Extraction. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 427–434, Ann Arbor, MI, USA.