Towards the Detection of Reliable Food-Health Relationships

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

Abstract

We investigate the task of detecting reliable statements about food-health relationships from natural language texts. For that purpose, we created a specially annotated web corpus from forum entries discussing the healthiness of certain food items. We examine a set of task-specific features (mostly) based on linguistic insights that are instrumental in finding utterances that are commonly perceived as reliable. These features are incorporated in a supervised classifier and compared against standard features that are widely used for various tasks in natural language processing, such as bag of words, part-of-speech and syntactic parse information.

1 Introduction

In this paper, we explore some linguistic high-level features to detect food-health relationships in natural language texts that are perceived reliable. By food-health relationships we mean relations that claim that a food item is suitable (1) or unsuitable (2) for some particular health condition.

(1) *Baking soda* is an approved remedy against heartburn.
(2) During pregnancy women should not consume any *alcohol*.

The same health claim may be uttered in different ways (3)-(5) and, as a consequence, may be perceived and judged differently. For the automatic extraction of health claims, we believe that statements that are perceived as *reliable* (4)-(5) are the most important to retrieve.

(3) *Eggs* do not have a negative impact on people suffering from heart diseases.
(4) **According to a leading medical scientist**, the consumption of *eggs* does not have a negative impact on people suffering from heart diseases.

(5) I'm suffering from a heart disease and **all my life** I've been eating many *eggs*; it never had any impact on my well-being.

In this work, we will mine a web corpus of forum entries for such relations. Social media are a promising source of such knowledge as, firstly, the language employed is not very technical and thus, unlike medical texts, accessible to the general public. Secondly, social media can be considered as an exclusive repository of *popular wisdom*. With regard to the health conditions, we can find, for example, home remedies. Despite the fact that many of them are not scientifically proven, there is still a great interest in that type of knowledge. However, even though such content is usually not subject to any scientific review, users would ideally appreciate an automatic assessment of the quality of each relation expressed. In this work, we attempt a first step towards this endeavour by automatically classifying these utterances with regard to *reliability*.

The features we examine will be based on linguistic insights that are instrumental in finding utterances that are commonly perceived as reliable. These features are incorporated in a supervised classifier and compared against standard features that are widely used for various taks in natural language processing, such as bag of words, part-of-speech and

syntactic parse information.

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

2 Related Work

As far as the extraction of health relations from social media are concerned, the prediction of epidemics (Fisichella et al., 2011; Torii et al., 2011; Diaz-Aviles et al., 2012; Munro et al., 2012) has recently attracted the attention of the research community.

Relation extraction involving food items has also been explored in the context of ontology alignment (van Hage et al., 2005; van Hage et al., 2006; van Hage et al., 2010) and also as a means of knowledge acquisition for virtual customer advice in a supermarket (Wiegand et al., 2012a).

The works most closely related to this paper are Yang et al. (2011) and Miao et al. (2012). Both of these works address the extraction of food-health relationships. Unlike this work, they extract relations from scientific biomedical texts rather than social media. Yang et al. (2011) also cover the task of *strength analysis* which bears some resemblance to the task of finding reliable utterances to some extent. However, the features applied to that classification task are only standard features, such as bag of words.

3 Data & Annotation

As a corpus 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 web portal for food-related issues in the German language. From this dataset, sentences in which some food item co-occurred with some health condition (e.g. *pregnancy*, *diarrhoea* or *flu*) were extracted. (In the following, we will also refer to these entities as *target food item* and *target health condition*.) The food items were identified with the help of GermaNet (Hamp and Feldweg, 1997), the German version of WordNet (Miller et al., 1990), and the health conditions were used from Wiegand et al. (2012b). In total, 2604 sentences were thus obtained.

For the manual annotation, each target sentence (i.e. a sentence with a co-occurrence of target food item and health condition) was presented in combination with the two sentences immediately preceding and following it. Each target sentence was manually assigned two labels, one specifying the type of suitability (§3.1) and another specifying whether the relation expressed is considered reliable or not (§3.2).

3.1 Types of Suitability

The suitability-label indicates whether a polar relationship holds between the target food item and the target health condition, and if so, which. Rather than just focusing on positive polarity, i.e. suitability, and negative polarity, i.e. unsuitability, we consider more fine-grained classes. As such, the suitabilitylabel does not provide any explicit information about the reliability of the utterance. In principle, every polar relationship between target food item and health condition expressed in a text could also be formulated in such a way that it is perceived reliable. In this work, we will consider the suitability-label as given. We use it as a feature in order to measure the correlation between suitability and reliability. The usage of fine-grained labels is to investigate whether subclasses of suitability or unsuitability have a tendency to co-occur with reliability. (In other words: we may assume differences among labels with the same polarity type.) We define the following set of fine-grained suitability-labels:

3.1.1 Suitable (SUIT)

SUIT encompasses all those statements in which the consumption of the target food item is claimed to be suitable for people affected by a particular health condition (6). By *suitable*, we mean that there will not be a negative effect on the health of a person once he or she consumes the target food item. However, this relation type does not state that the consumption is likely to improve the condition of the person either.

(6) I also got dermatitis which is why my mother used *spelt flour* [instead of wheat flour]; you don't taste a difference.

[1] www.chefkoch.de

positive labels	BENEF, SUIT, PREVENT
negative labels	UNSUIT, CAUSE

Table 1: Categorization of suitability-labels.

3.1.2 Beneficial (BENEF)

While SUIT only states that the consumption of the target food item is suitable for people with a particular health condition, BENEF actually states that the consumption alleviates the symptoms of the condition or even cures it (7). While both SUIT and BENEF have a positive polarity, SUIT is much more neutral than BENEF.

(7) Usually, a glass of *milk* helps me when I got a sore throat.

3.1.3 Prevention (PREVENT)

An even stronger positive effect than the relation type BENEF presents PREVENT which claims that the consumption of the target food item can prevent the outbreak of a particular disease (8).

(8) *Citric acid* largely reduces the chances of kidney stones to develop.

3.1.4 Unsuitable (UNSUIT)

UNSUIT describes cases in which the consumption of the target food item is deemed unsuitable (9). Unsuitability means that one expects a negative effect (but it need not be mentioned explicitly), that is, a deterioration of the health situation on the part of the person who is affected by a particular health condition.

(9) *Raw milk cheese* should not be eaten during pregnancy.

3.1.5 Causation (CAUSE)

CAUSE is the negative counterpart of PREVENT. It states that the consumption of the target food item can actually cause a particular health condition (10).

(10) It's a common fact that the regular consumption of *coke* causes caries.

The suitability-labels can also be further separated into two polar classes (i.e. positive and negative labels) as displayed in Table 1.

3.2 Reliability

Each utterance was additionally labeled as to whether it was considered reliable (4)-(5) or not (3). It is this label that we try to predict in this work. By *reliable*, we understand utterances in which the relations expressed are convincing in the sense that a reputable source is cited, some explanation or empirical evidence for the relation is given, or the relation itself is emphasized by the speaker. In this work, we are exclusively interested in detecting utterances which are *perceived* reliable by the reader. We leave aside whether the statements from our text corpus are actually correct. Our aim is to identify linguistic cues that evoke the impression of *reliability* on behalf of the reader.

3.3 Class Distributions and Annotation Agreement

Table 2 depicts the distribution of the reliability-labels on our corpus while Table 3 lists the class distribution of the suitability-labels including the proportion of the reliable instances among each category. The proportion of reliable instances varies quite a lot among the different suitability-labels, which indicates that the suitability may be some effective feature.

Note that the class OTHER in Table 3 comprises all instances in which the co-occurrence of a health condition and a food item was co-incidental (11) or there was some embedding that discarded the validity of the respective suitability-relation, as it is the case in questions (12).

(11) It's not his diabetes I'm concerned about but the enormous amounts of *fat* that he consumes.

(12) Does anyone know whether I can eat *tofu* during my pregnancy?

In order to measure interannotation agreement, we collected for three health conditions their co-occurrences with any food item. For the suitability-labels we computed Cohen's $\kappa = 0.76$ and for the reliability-labels $\kappa = 0.61$. The agreement for reliability is lower than for suitability. We assume that the reason for that lies in the highly subjective notion of reliability. Still, both agreements can be interpreted as *substantial* (Landis and Koch, 1977) and should be sufficiently high for our experiments.

Type	Frequency	Percentage
Reliable	480	18.43
Not Reliable	2124	81.57

Table 2: Distribution of the reliability-labels.

Type	Frequency	Perc.	Perc. Reliable
BENEF	502	19.28	33.39
CAUSE	482	18.51	22.57
SUIT	428	16.44	17.91
UNSUIT	277	10.64	34.05
PREVENT	74	2.84	14.04
OTHER	841	32.30	0.00

Table 3: Distribution of the suitability-labels.

4 Feature Design

4.1 Task-specific High-level Feature Types

We now describe the different task-specific high-level feature types. We call them *high-level* feature types since they model concepts that typically generalize over sets of individual words (i.e. low-level features).

4.1.1 Explanatory Statements (EXPL)

The most obvious type of reliability is a suitability-relation that is also accompanied by some explanatory statement. That is, some reason for the relation expressed is given (13). We detect reasons by scanning a sentence for typical discourse cues (more precisely: conjunctions) that anchor such remarks, e.g. *which is why* or *because*.

(13) *Honey* has an antiseptic effect **which is why** it is an ideal additive to milk in order to cure a sore throat.

4.1.2 Frequent Observation (FREQ)

If a speaker claims to have witnessed a certain relation very frequently or even at all times, then there is a high likelihood that this relation actually holds (14). We use a set of adverbs (18 expressions) that express high frequency (e.g. *often*, *frequently* etc.) or constancy (e.g. *always*, *at all times* etc.).

(14) What **always** helps me when I have the flu is a hot *chicken broth*.

4.1.3 Intensifiers (INTENS)

Some utterances may also be perceived reliable if their speaker adds some emphasis to them. One way of doing so is by adding intensifiers to a remark (15).

(15) You can treat nausea with *ginger* **very effectively**.

The intensifiers, we use are a translation of the lexicon introduced in Wilson et al. (2005). For the detection, we divide that list into two groups:

The first group INTENS_{simple} are unambiguous adverbs that always function as intensifiers no matter in which context they appear (e.g. *very* or *extremely*).

The second group includes more ambiguous expressions, such as adjectives that only function as an intensifier if they modify a polar expression (e.g. *horrible pain* or *terribly nice*) otherwise they function as typical polar expressions (e.g. *you are horrible*$^-$ or *he sang terribly*$^-$). We employ two methods to detect these ambiguous expressions. INTENS_{polar} requires a polar expression of a polarity lexicon to be modified by the intensifier, while INTENS_{adj} requires an adjective to be modified. In order to identify polar expressions we use the polarity lexicon underlying the *PolArt* system (Klenner et al., 2009). We also consider adjectives since we must assume that our polarity lexicon does not cover all possible polar expressions. We chose adjectives as a complement criterion as this part of speech is known to contain many polar expressions (Hatzivassiloglou and McKeown, 1997; Hatzivassiloglou and Wiebe, 2000).

4.1.4 Strong Polar Expressions (STROPO)

Instead of adding intensifiers in order to put more emphasis to a remark (§4.1.3), one may also use polar expressions that convey a high polar intensity (16). For instance, *nice* and *excellent* refer to the same scale and convey positive polarity but *excellent* has a much higher intensity than *nice*. Taboada et al. (2011) introduced an English polarity lexicon *SO-CAL* in which polar expressions were also assigned an intensity label. As our German polarity lexicon (§4.1.3) does not contain comparable intensity labels, we used a German translation of *SO-CAL*. We identified polar expressions with a high intensity score (i.e. ± 4 or ± 5) as *strong polar expressions*. It includes 221 highly positive and 344 highly negative polar expressions. We also distinguish the polarity type (i.e. STROPO^+ refers to positive and STROPO^- refers to negative polarity).

(16) *Baking soda* is an **excellent** remedy against heartburn.

4.1.5 Superlatives (SUPER)

Another way of expressing high polar intensity is by applying superlatives (17). Superlatives can only be formed from gradable adjectives. At the same time, the greatest amount of such adjectives are also subjective expressions (Hatzivassiloglou and Wiebe, 2000). As a consequence, the detection of this grammatical category does not depend on a subjectivity/polarity lexicon but on simple morphological suffixes (e.g. *-est* in *strongest*)[2] or combinations with certain modifiers (e.g. *most* in *most terrific*).

(17) *Baking soda* is the **most effective** remedy against heartburn.

4.1.6 Statements Made By Authorities (AUTH)

If a statement is quoted from an authority, then it is usually perceived more reliable than other statements (4). Authorities in our domain are mostly scientists and medical doctors. Not only does a mention of those types of professions indicate an authority but also the citation of their work. Therefore, for this feature we also scan for expressions, such as *journal*, *report*, *survey* etc. Our final look-up list of cues comprises 53 expressions.

We also considered using the knowledge of user profiles in order to identify speakers whose profession fall under our defined set of authorities. Unfortunately, the overwhelming majority of users who actually specified their profession cannot be considered as authorities (for the relations that we are interested in) by mere consideration of their profession. Most users of *chefkoch.de* are either office employees, housewifes, students or chefs. Less than 1% are authorities according to our definition. Due to the severe sparsity of authorities, we refrained from using the professions as they are specified in the user profiles.

[2]We could not use part-of-speech tagging for the detection of superlatives since unlike the standard English part-of-speech tag set (i.e. the Penn Treebank Tag Set (Marcus et al., 1993)), information regarding gradation (i.e. comparative and superlative) is not reflected in the standard German tag set (i.e. Stuttgart Tübinger Tag Set (Schiller et al., 1995)).

4.1.7 Doctors' Prescriptions (PRESC)

Some of our food-health relations are also mentioned in the context of doctors' prescriptions (5). That is, a doctor may prescribe a patient to consume a particular food item since it is considered suitable for their health condition, or he/she may forbid a food item in case it is considered unsuitable. As already pointed out in §4.1.6, doctors usually present an authority with regard to food-health relations. That is why, their remarks should be considered reliable.

In order to detect doctors' prescriptions, we mainly look for (modal) verbs in a sentence that express obligations or prohibition. We found that, on our dataset, people rarely mention their doctor explicitly if they refer to a particular prescription. Instead, they just mention that they must or must not consume a particular food item. From the context, however, it is obvious that they refer to their doctor's prescription (18).

(18) Due to my diabetes I **must** not eat any *sweets*.

4.1.8 Hedge Cues (HEDGE)

While all previous features were designed to identify cases of reliable statements, we also include features that indicate the opposite. The most obvious type of utterances that are commonly considered unreliable are so-called *hedges* (Lakoff, 1973) or speculations (19).

(19) *Coke* **may** cause cancer.

For this feature, we use a German translation of English cue words that have been found useful in previous work (Morante and Daelemans, 2009) which results in 47 different expressions.

4.1.9 Types of Suitability-Relations (REL)

Finally, we also incorporate the information about what type of suitability-relations the statement was labeled with. The suitability-labels were already presented and motivated in §3.1. The concrete features are: REL_{SUIT} (§3.1.1), REL_{BENEF} (§3.1.2), $\text{REL}_{PREVENT}$ (§3.1.3), REL_{UNSUIT} (§3.1.4), REL_{CAUSE} (§3.1.5).

Suffix	Description
-WND$_{food}$	context window around food item
-WND$_{cond}$	context window around health condition
-TS	target sentence only
-EC	entire (instance) context

Table 4: Variants for the individual feature types.

4.2 Variants of Feature Types

For our feature types we examine several variants that differ in the size of context/scope. We distinguish between the target sentence and the entire context of an instance, i.e. the target sentence plus the two preceding and following sentences (§3). If only the target sentence is considered, we can also confine the occurrence of a cue word to a fixed window (comprising 5 words) either around the target food item or the target health condition rather than considering the entire sentence.

Small contexts usually offer a good precision. For example, if a feature type occurs nearby a mention of the target food item or health condition, the feature type and the target expression are likely to be related to each other. The downside of such narrow contexts is that they may be too sparse. Wide contexts may be better suited to situations in which a high recall is desirable. However, ambiguous feature types may perform poorly with these contexts as their co-occurrence with a target expression at a large distance is likely to be co-incidental.

Table 4 lists all the variants that we use. These variants are applied to all feature types except the types of suitability (§4.1.9) as this label has only been assigned to an entire target sentence.

4.3 Other Features

Table 5 lists the entire set of features that we examine in this work. The simplest classifier that we can construct for our task is a trivial classifier that predicts all statements as reliable statements. The remaining features comprise bag of words, part-of-speech and syntactic parse information. For the latter two features, we employ the output of the Stanford Parser for German (Rafferty and Manning, 2008).

Features	Description
all	trivial classifier that always predicts a reliable statement
bow	bag-of-words features: all words between the target food item and target health condition and the words immediately preceding and following each of them
pos	part-of-speech features: part-of-speech sequence between target food item and health condition and tags of the words immediately preceding and following each of the target expressions
synt	path from syntactic parse tree from target food item to target health condition
task	all task-specific high-level feature types from §4.1 with their respective variants (§4.2)

Table 5: Description of all feature sets.

5 Experiments

Each instance to be classified is a sentence in which there is a co-occurrence of a target food item and a target health condition along its respective context sentences (Section 3). We only consider sentences in which the co-occurrence expresses an actual suitability relationship between the target food item and the target health condition, that is, we ignore instances labeled with the suitability-label OTHER (§3.3). We make this restriction as the instances labeled as OTHER are not eligible for being reliable statements (Table 3). In this work, we take the suitability-labels for granted (this allows us to easily exclude the instances labeled as OTHER). The automatic detection of suitability-labels would require a different classifier with a different set of features whose appropriate discussion would be beyond the scope of this paper.

5.1 Comparison of the Different Task-specific High-level Features

In our first experiment, we want to find out how the different task-specific high-level features that we have proposed in this work compare to each other. More specifically, we want to find out how the individual features correlate with the utterances that have been manually marked as reliable. For that purpose, Table 6 shows the top 20 features according to Chi-square feature selection computed with *WEKA* (Witten and Frank, 2005). More information regarding the computation of Chi-square statistics in the context of text classification can be found Yang and Pederson (1997). Note that we apply feature selection only as a means of feature comparison. For

Rank	Feature	Score
1	**FREQ-WND$_{food}$**	105.1
2	FREQ-TS	102.8
3	FREQ-WND$_{cond}$	75.9
4	FREQ-EC	29.2
5	**AUTH-EC**	23.7
6	**STROPO^{+}-WND$_{cond}$**	20.5
7	**REL$_{BENEF}$**	20.2
8	REL$_{SUIT}$	16.8
9	**INTENS$_{simple}$-WND$_{cond}$**	16.4
10	AUTH-TS	15.4
11	STROPO^{+}-TS	15.0
12	INTENS$_{simple}$-EC	14.1
13	STROPO^{+}-WND$_{food}$	13.7
14	INTENS$_{adj}$-WND$_{food}$	13.2
15	INTENS$_{simple}$-WND$_{food}$	12.1
16	INTENS$_{simple}$-TS	11.6
17	**PRESC-WND$_{food}$**	11.0
18	INTENS$_{adj}$-WND$_{cond}$	9.7
19	INTENS$_{polar}$-EC	9.0
20	AUTH-WND$_{food}$	7.9

Table 6: Top 20 features according to Chi-square feature ranking (for each feature type the most highly ranked variant is highlighted).

classification (§5.2), we will use the entire feature set.

5.1.1 What are the most effective features?

There are basically five feature types that dominate the highest ranks. They are FREQ, AUTH, STROPO, REL and INTENS. This already indicates that several features presented in this work are effective. It is interesting to see that two types of suitability-labels, i.e. REL$_{BENEF}$ and REL$_{SUIT}$, are among the highest ranked features which suggests that suitability and reliability are somehow connected.

Table 7 shows both precision and recall for each of the most highly ranked variant of the feature types that appear on the top 20 ranks according to Chi-square ranking (Table 6). Thus, we can have an idea in how far the high performing different feature types differ. We only display one feature per feature type due to the limited space. The table shows that for most of these features precision largely outperforms recall. REL$_{BENEF}$ is the only notable exception (its recall actually outperforms precision).

5.1.2 Positive Orientation and Reliability

By closer inspection of the highly ranked features, we found quite a few features with positive orientation, i.e. STROPO^{+}-WND$_{cond}$, REL$_{BENEF}$, REL$_{SUIT}$, STROPO^{+}-WND$_{cond}$, while their negative counterparts are absent. This raises the question whether there is a bias for positive orientation for the detection of reliability.

Feature	Prec	Rec
FREQ-WND$_{food}$	71.13	14.38
AUTH-EC	41.81	15.42
STROPO^{+}-WND$_{cond}$	63.38	3.54
REL$_{BENEF}$	33.39	39.17
INTENS$_{simple}$-WND$_{cond}$	41.73	11.04
PRESC-WND$_{food}$	45.00	5.63

Table 7: Precision and recall of different features (we list the most highly ranked variants of the feature types from Table 6).

We assume that there are different reasons why the positive suitability-labels (REL$_{BENEF}$ and REL$_{SUIT}$) and strong positive polarity (STROPO^{+}) are highly ranked features:

As far as polarity features are concerned, it is known from sentiment analysis that positive polarity is usually easier to detect than negative polarity (Wiegand et al., 2013). This can largely be ascribed to social conventions to be less blunt with communicating negative sentiment. For that reason, for example, one often applies negated positive polar expressions (e.g. *not okay*) or irony to express a negative sentiment rather than using an explicit negative polar expression. Of course, such implicit types of negative polarity are much more difficult to detect automatically.

The highly ranked suitability-labels may be labels with the same orientation (i.e. they both describe relationships that a food item is suitable rather than unsuitable for a particular health condition), yet they have quite different properties.[3] While REL$_{BENEF}$ is a feature positively correlating with reliable utterances, the opposite is true of REL$_{SUIT}$, that is, there is a correlation but this correlation is negative. Table 8 compares their respective precision and also includes the trivial (reference) classifier *all* that always predicts a reliable statement. The table clearly shows that REL$_{BENEF}$ is above the triv-

[3]It is not the case that the proportion of reliable utterances is larger among the entire set of instances tagged with positive suitability-labels than among the negative instances tagged with negative suitability-labels (Table 1). In both cases, they are at approx. 26%.

ial feature while REL$_{SUIT}$ is clearly below. (One may wonder why the gap in precision between those different features is not larger. These features are also high-recall features – we have shown this for REL$_{BENEF}$ in Table 7 – so the smaller gaps may already have a significant impact.) In plain, this result means that a statement conveying that some food item alleviates the symptoms of a particular disease or even cures it (REL$_{BENEF}$) is more likely to contain utterances that are perceived reliable rather than statements in which the speaker merely states that the food item is suitable given a particular health condition (REL$_{SUIT}$). Presumably, the latter type of suitability-relations are mostly uttered parenthetically (not emphatically) or they are remarks in which the relation is inferred, so that they are unlikely to provide further background information. In Sentence (20), for example, the suitability of *wholemeal products* is inferred as the speaker's father eats these types of food due to his *diabetes*. The focus of this remark, however, is the psychic well-being of the speaker's father. That entire utterance does not present any especially reliable or otherwise helpful information regarding the relationship between *diabetes* and *wholemeal products*.

(20) My father suffers from diabetes and is fed up with eating all these *wholemeal products*. We are worried that he is going to fall into a depression.

Having explained that the two (frequently occurring) positive suitability-labels are highly ranked features because they separate reliable from less reliable statements, one may wonder why we do not find a similar behaviour on the negative suitability-labels. The answer to this lies in the fact that there is no similar distinction between REL$_{BENEF}$ and REL$_{SUIT}$ among utterances expressing unsuitability. There is no neutral negative suitability-label similar to REL$_{SUIT}$. The relation REL$_{UNSUIT}$ expresses unsuitability which is usually connected with some deterioration in health.

5.1.3 How important are explanatory statements for this task?

We were very surprised that the feature type to indicate explanatory statements EXPL (§4.1.1) performed very poorly (none of its variants is listed in Table 6) since we consider it as one of the more interesting types of utterances to extract. In order to find a reason for this, we manually annotated all reliable utterances as to whether they can be regarded as an explanatory statement (EXPL$_{all}$) and, if so, whether (in principle) there are lexical cues (such as our set of conjunctions) to identify them (EXPL$_{cue}$). Table 9 shows the proportion of these two categories among the reliable utterances. With more than 20% being labeled as this subtype, explanatory statements are clearly not a fringe phenomenon. However, lexical cues could only be observed in approximately $1/3$ of those instances. The majority of cases, such as Sentence (21), do not contain any lexical cues and are thus extremely difficult to detect.

Feature	REL$_{SUIT}$	*all*	REL$_{BENEF}$
Prec	17.81	26.46	33.39

Table 8: The precision of different REL-features compared to the trivial classifier *all* that always predicts a reliable utterance.

Type	EXPL$_{all}$	EXPL$_{cue}$
Percentage	22.59	8.30

Table 9: Proportion of explanatory statements among reliable utterances (EXPL$_{all}$: all reliable instances that are explanatory statements; EXPL$_{cue}$: subset of explanatory statements that also contain a lexical cue).

(21) *Citrus fruits* are bad for dermatitis. They increase the itch. Such fruits are rich in acids that irritate your skin.

In addition, all variants of our feature type EXPL have a poor precision (between $20 - 25\%$). This means that the underlying lexical cues are too ambiguous.

5.1.4 How important are the different contextual scopes?

Table 6 clearly shows that the contextual scope of a feature type matters. For example, for the feature type FREQ, the most effective scope achieves a Chi-square score of 105.1 while the worst variant only achieves a score of 29.2. However, there is no unique contextual scope which always outperforms the other variants. This is mostly due to the

Feature Set	Prec	Rec	F1
all	26.46	**100.00**	41.85
bow	37.14	62.44	46.45
bow+pos	36.85	57.64	44.88
bow+synt	39.05	58.01	46.58
task	35.16	72.89	47.21
bow+task	**42.54**	66.01	**51.56***

Table 10: Comparison of different feature sets (summary of features is displayed in Table 5); * significantly better than *bow* at $p < 0.05$ (based on paired t-test).

fact the different feature types have different properties. On the one hand, there are unambiguous feature types, such as AUTH, which work fine with a wide scope. But we also have ambiguous feature types that require a fairly narrow context. A typical example are strong (positive) polar expressions (STROPO$^+$). (Polar expressions are known to be very ambiguous (Wiebe and Mihalcea, 2006; Akkaya et al., 2009).)

5.2 Classification

Table 10 compares the different feature sets with regard to extraction performance. We carry out a 5-fold cross-validation on our manually labeled dataset. As a classifier, we chose Support Vector Machines (Joachims, 1999). As a toolkit, we use *SVMLight*[4] with a linear kernel.

Table 10 clearly shows the strength of the high-level features that we proposed. They do not only represent a strong feature set on their own but they can also usefully be combined with bag-of-words features. Apparently, neither part-of-speech nor parse information are predictive for this task.

5.3 Impact of Training Data

Figure 1 compares bag-of-words features and our task-specific high-level features on a learning curve. The curve shows that the inclusion of our task-specific features improves performance. Interestingly, with *task* we obtain a good performance on smaller amounts of data. However, this classifier is already saturated with 40% of the training data. From then onwards, it is more effective to use the combination *bow+task*. Our high-level features generalize well which is particularly important for situations in which only few training data available.

[4] http://svmlight.joachims.org

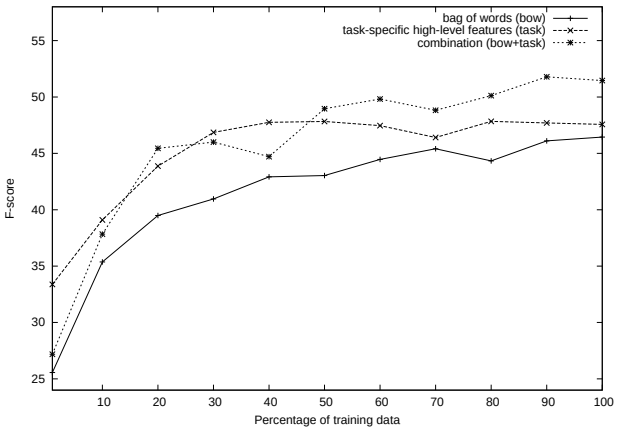

Figure 1: Learning curve of the different feature sets.

However, in situations in which large training sets are available, we additionally need bag of words that are able to harness more sparse but specific information.

6 Conclusion

In this paper, we examined a set of task-specific high-level features in order to detect food-health relations that are perceived reliable. We found that, in principle, a subset of these features that include strong polar expressions, intensifiers and adverbials expressing frequent observations are fairly predictive and complement bag-of-words information. Moreover, the effectiveness of the different features depends very much on the context to which they are applied.

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. The authors would also like to thank Prof. Dr. Wolfgang Menzel for providing the German version of the SO-CAL polarity lexicon that has been developed at his department.

References

Cem Akkaya, Janyce Wiebe, and Rada Mihalcea. 2009. Subjectivity Word Sense Disambiguation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 190–199, Singapore.

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.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet - a Lexical-Semantic Net for German. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15, Madrid, Spain.

Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1997. Predicting the Semantic Orientation of Adjectives. In *Proceedings of the Conference on European Chapter of the Association for Computational Linguistics (EACL)*, pages 174–181, Madrid, Spain.

Vasileios Hatzivassiloglou and Janyce Wiebe. 2000. Effects of Adjective Orientation and Gradability on Sentence Subjectivity. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 299–305, Saarbrücken, Germany.

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*. 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.

George Lakoff. 1973. Hedging: A Study in Media Criteria and the Logic of Fuzzy Concepts. *Journal of Philosophical Logic*, 2:458 – 508.

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

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, June. Special Issue on Using Large Corpora.

Qingliang Miao, Shu Zhang, Bo Zhang, Yao Meng, and Hao Yu. 2012. Extracting and Visualizing Semantic Relationships from Chinese Biomedical Text. In *Proceedings of the Pacific Asia Conference on Language, Information and Compuation (PACLIC)*, pages 99–107, Bali, Indonesia.

George Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. 1990. Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*, 3:235–244.

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 th Spring Symposium for Association for the Advancement of Artificial Intelligence (AAAI)*, pages 52–58, Toronto, Canada.

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.

Anne Schiller, Simone Teufel, Christine Stöckert, and Christine Thielen. 1995. Vorläufige Guidelines für das Tagging deutscher Textcorpora mit STTS. Technical report, Universität Stuttgart, Insitut für maschinelle Sprachverarbeitung, and Seminar für Sprachwissenschaft, Universität Tübingen.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, 37(2):267 – 307.

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. *Internal Journal of Medical Informatics*, 80(1):56–66.

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.

Janyce Wiebe and Rada Mihalcea. 2006. Word Sense and Subjectivity. In *Proceedings of the International Conference on Computational Linguistics and Annual Meeting of the Association for Computational Linguistics (COLING/ACL)*, pages 1065–1072, Syndney, Australia.

Michael Wiegand, Benjamin Roth, and Dietrich Klakow. 2012a. Web-based Relation Extraction for the Food Domain. In *Proceeding 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.

Michael Wiegand, Manfred Klenner, and Dietrich Klakow. 2013. Bootstrapping polarity classifiers with rule-based classification. *Language Resources and Evaluation*, Online First:1–40.

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.

Yiming Yang and Jan Pederson. 1997. A Comparative Study on Feature Selection in Text Categorization. In *Proceedings the International Conference on Machine Learning (ICML)*, pages 412–420, Nashville, US.

Hui Yang, Rajesh Swaminathan, Abhishek Sharma, Vilas Ketkar, and Jason D'Silva, 2011. *Learning Structure and Schemas from Documents*, volume 375 of *Studies in Computational Intelligence*, chapter Mining Biomedical Text Towards Building a Quantitative Food-disease-gene Network, pages 205–225. Springer Berlin Heidelberg.