Detecting Conditional Healthiness of Food Items from Natural Language Text

Michael Wiegand · Dietrich Klakow

Abstract In this article, we explore the feasibility of extracting suitable and unsuitable food items for particular health conditions from natural language text. We refer to this task as *conditional healthiness classification*. For that purpose, we annotate a corpus extracted from forum entries of a food-related website. We identify different relation types that hold between food items and health conditions going beyond a binary distinction of suitability and unsuitability and devise various supervised classifiers using different types of features. We examine the impact of different task-specific resources, such as a *healthiness lexicon* that lists the healthiness status of a food item and a sentiment lexicon. Moreover, we also consider task-specific linguistic features that disambiguate a context in which mentions of a food item and a health condition co-occur and compare them with standard features using bag of words, part-of-speech information and syntactic parses. We also investigate in how far individual food items and health conditions correlate with specific relation types and try to harness this information for classification.

1 Introduction

Food plays a substantial part in each of our lives. This is not only due to the fact that we need it to survive but also since it has social and cultural significance. With the growing health awareness in many parts of the population, there is consequently

Michael Wiegand
Spoken Language Systems, Saarland University, Building C7.1, D-66123 Saarbrücken
Tel.: +49-681-30258133
Fax: +49-681-30258124
E-mail: michael.wiegand@lsv.uni-saarland.de

Dietrich Klakow
Spoken Language Systems, Saarland University, Building C7.1, D-66123 Saarbrücken

a high demand for the knowledge about the tolerability of different food items for specific health conditions. In view of the variety of both different types of food and health conditions it does not come as a surprise that there does not exist a common repository summarizing all 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.

In this article, we examine the task of identifying mentions that a food item is suitable (1) or unsuitable (2) given a particular health condition. We will also refer to this type of classification as *conditional healthiness classification*. Therefore, Sentence (1) is considered an example of *conditional healthiness* while Sentence (2) is considered an example of *conditional unhealthiness*.

(1) People suffering from gout may eat *tomatoes*.
(2) During pregnancy women should not consume any *alcohol*.

Instead of just considering the binary distinction between suitability and unsuitability, we will consider typical subclasses. For example, a suitable food item may also alleviate the symptoms of the health condition (3) or even prevent the disease from breaking out (4).

(3) *Ginger* is very good for your stomach – it helped me a lot against my heartburn.
(4) Iron deficiency can usually be prevented by consuming *meat* on a regular basis.

To the best of our knowledge, this is the first work that addresses this type of classification. Therefore, we need to answer some basic questions: We want to know whether this type of information can be extracted from natural language text at all, what relation types occur sufficiently often so that automatic extraction is possible, and what types of features for supervised classifiers should be taken into consideration. Apart from standard bag-of-words features, we consider a set of linguistic features. In addition, we will also test some manually compiled keywords collected from our dataset. Thus, we have an upper bound estimate in how far lexical information would help to distinguish between the different classes.

Moreover, we want to investigate in how far natural language processing is beneficial in the presence of simple heuristics. A straightforward baseline would be, for instance, to always classify a food item according to general healthiness (which is usually derived from general nutrient content) and thus completely ignore specific health conditions. For example, since apples are generally considered healthy, they would also be predicted as *conditional healthy* food items. In addition, we are interested in how far certain health conditions correlate with particular class labels. If there is generally a strong correlation between a particular health condition and a particular class label, e.g. diarrhoea and *causation*[1], then classifying food items for a particular health condition from natural language text could be reduced to just extracting food items that frequently co-occur with mentions of that condition. (The class label would always be the one that mostly co-occurs with the respective health condition.) Further contextual disambiguation would be superfluous.

---

[1] In our dataset, people mostly discuss food items that cause the outbreak of that disease rather than food items that protect them against it or alleviate the symptoms if they contract it.

Note that we are exclusively interested in the classification of individual utterances rather than an aggregate assessment of a set of utterances. However, it should be noted that the former is a pre-requisite of the latter: it is not possible to produce an aggregate assessment without a correct assessment of (most) individual utterances.

In terms of applications, the task we examine in this article could be seen as a further step towards a more intelligent search engine, that is, a search engine being able to retrieve content not only on the basis of keyword matching but also on the basis of some limited form of text understanding. Websites that host large user forums on food and/or health-related issues, as the one from which we extracted our gold standard (see also Section 2), could incorporate such technology in their search inferface allowing their users a more focused search.

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 of this article, all German examples will be accompanied by English translations.

The major contributions of this article are (i) a new annotation scheme for conditional healthiness classification, (ii) a comparison of diverse features for supervised classification and (iii) a detailed error analysis in which we try to uncover the reasons why certain features are not effective or why they only produce low classification performance for certain classes.

2 The Dataset

In order to generate a dataset for our experiments, we used a crawl of *chefkoch.de*[2] (Wiegand et al, 2012b) 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. In Wiegand et al (2012a, 2014) it was found that for domain-specific relation extraction in the food domain, such a corpus largely outperforms alternative open-domain data collections, such as *Wikipedia*.

The advantage of using a domain-specific corpus is that we find much more content for the target relations than in other standard corpora. For instance, we found that our domain-specific corpus contains almost *5 times* as many co-occurrences of a food item and a health condition (in one sentence)[3] as the German *Wikipedia* dump created at (roughly) the same time our domain-specific food corpus was created. Compared with the German subset of the *Web 1T 5gram* (LDC2009T25)[4], our domain-specific corpus contains even *6 times* as many co-occurrences.[5]

---

[2] www.chefkoch.de

[3] Note that co-occurrences of a food item and a health condition are only an approximation of genuine food-health relations. In other words, not every of these co-occurrences necessarily conveys a proper food-health relationship. However, without manually annotating each of these co-occurrences, this is the best approximation we can produce.

[4] For this corpus, we can only observe food-health co-occurrences within 5grams and not entire sentences.

[5] The low coverage on the 5grams can be partially explained by the fact that this corpus only contains ngrams observed at least 40 times. We assume that there are considerably more co-occurrences of food items and health conditions among web-based 5grams at lower frequencies.

Table 1 List of health conditions (*English translation/frequency*) for which suitable/unsuitable food items are to be extracted.

Health Conditions
Abwehrkräfte (healthy immune system/2), Akne (acne/15), Asthma (asthma/20), Blähungen (flatulence/1), Blasenentzündung (cystitis/32), Bluthochdruck (hypertension/32), Kalziummangel (acalcinosis/5), Depression (depression/19), Diabetes (diabetes/134), Durchfall (diarrhoea/255), Eisenmangel (iron deficiency/29), Eiweißallergie (protein intolerance/6), Fieber (fever/127), Gicht (gout/34), Grippe (flu/86), Halsschmerzen (sore throat/183), Heiserkeit (hoarseness/14), Heuschnupfen (hay fever/8), Herzkrankheit (heart disease/24), Husten (cough/191), Karies (caries/25), Kondition (body condition/4), Kopfschmerzen (headache/280), Magengeschwür (stomach ulcer/9), Mandelentzündung (tonsillitis/11), Müdigkeit (lassitude/44), Mumps (mumps/1), Neurodermitis (dermatitis/194), Nierensteine (kidney stones/16), Reizdarm (irritable bowel syndrome/4), Rheuma (rheumatism/29), Salmonellenvergiftung (salmonella/17), Schnupfen (coryza/86), Schuppenflechte (psoriasis/9), Sodbrennen (heartburn/232), Übelkeit (nausea/103), Übergewicht (overweight/79), Untergewicht (underweight/13), Verstopfung (constipation/75), Vitamin C Mangel (lack of vitamin C/1), Zahnschmerzen (toothache/39)

While we are aware of the fact that the conditional healthiness of food items is also discussed in scientific (medical) texts, these types of text (with a size similar to our web corpus) are not available to us. (In fact, we believe that it is fairly difficult to acquire a corpus from the medical domain with the same diversity of co-occurrences of health conditions and food items. This is due to the fact that individual medical texts are more thematically focused. That is, they typically just deal with a particular pair of health condition and food item, e.g. *eggs* and *heart disease*.) However, we do not only examine this classification task on a web corpus just because other text types are not available to us. We think that the text analysis on web data serves its own purpose. Scientific medical texts are due to their complexity hardly accessible to people without solid background knowledge in medical science. On the other hand, conditional healthiness attracts a large part of the general population. The forum entries we extracted employ a language that is much easier to follow. Thus, we believe that text excerpts from forum entries are more suitable to satisfy the information need of that part of the population.

Not only do medical texts differ from social media in the type of language, they also differ in the type of content they provide. 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. In this work, we do not make any attempts to separate genuine facts from claims. The majority of relations extracted are just unsubstantiated claims made by the public.

2.1 The Choice of Health Conditions

Table 1 lists the health conditions we consider for the creation of our dataset. We focused on those health conditions where at least a mild relationship between the condition and food (in general) is known. This is ensured by the fact that the list is a subset of conditions for which suitable and unsuitable food items have been listed in the gold standard introduced in Wiegand et al (2012c). Another side-effect of

PRECEDING CONTEXT:	Ich trinke auch keinen Kaffee , mag auch nicht so gern Tee.
	(I don't drink any coffee and I don't like tea very much, either.)
PRECEDING CONTEXT:	Mal welchen, wenn ich krank bin.
	(I only drink it when I'm ill.)
TARGET SENTENCE:	Aber ich bekomme vom Tee_{food_item} normalerweise $Sodbrennen_{health\_condition}$.
	(But, I usually get $heartburn_{health\_condition}$ from drinking tea_{food_item}.)
SUCCEEDING CONTEXT:	Am liebsten mag ich ganz normales Wasser.
	(I prefer drinking mineral water.)
SUCCEEDING CONTEXT:	Das ist auch immer so, wenn ich woanders eingeladen bin.
	(That's also what I choose when I'm invited somewhere.)

Fig. 1 Illustration of a data instance; target food item: *tea*, target health condition: *heartburn*, relation type: *CAUSE* (Section 2.3.6).

restricting ourselves to entries from this gold standard is that these health conditions occur fairly frequently (for an average health condition we find 64 mentions in our dataset).

2.2 Annotation

Our final dataset consists of 2604 instances, where each instance represents a sentence in which a food item co-occurs with one of the health conditions. While we restrict the health conditions to the ones listed in Table 1, we make no restriction on the target food items we consider. Similar to Wiegand et al (2012a) and Wiegand et al (2012b), food items are detected with the help of GermaNet (Hamp and Feldweg, 1997), the German version of WordNet (Miller et al, 1990).

Figure 1 illustrates a typical instance. Note that for one instance, we only specify one target food item and one health condition. This means that in case a sentence contains more than one of these expressions, it may appear several times as a target sentence but each time we will assign a different target food item or health condition.

Each instance of the extracted corpus was manually assessed. Two native speakers of German were employed as annotators. We initially selected three health conditions for which we extracted all sentences in which there is a co-occurrence with some food item. Each of these instances was annotated by both annotators in order to measure interannotation agreement. (The remaining instances were only annotated by one annotator each.) This resulted in 276 instances. For each instance, the annotators were to choose exactly one category label. We measured an inter-annotation agreement of Cohen's $\kappa = 0.7651$. This agreement can be interpreted as *substantial* (Landis and Koch, 1977) and should be sufficiently high for our experiments. The dataset, including the annotation guidelines we devised, will be made publicly available for research purposes.

Table 2 Statistics of the different category labels.

Type	Abbreviation	Frequency	Percentage
No relation	NOREL	539	20.70
Beneficial	BENEF	502	19.28
Causation	CAUSE	482	18.51
Suitable	SUIT	428	16.44
Embedded relation	EMBREL	302	11.60
Unsuitable	UNSUIT	247	9.49
Prevention	PREVENT	74	2.84
Worsening	WORSEN	30	1.15

2.3 The Annotation Scheme

We now describe the different category labels. Their distribution is depicted in Table 2.[6]

2.3.1 Suitable (SUIT)

SUIT encompasses all those statements in which the consumption of the target food item is suitable for people affected by a particular health condition (5). 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, even though it is not impossible. The way how the suitability is expressed does not explicitly address positive effects.

(5) Ich hatte auch Neurodermitis, daher verwendete meine Mutter nur *Dinkelmehl* [anstatt Weizenmehl]; man merkt fast keinen Unterschied.
(I also had dermatitis which is why my mother used *spelt flour* [instead of wheat flour]; you can taste almost no difference.)

2.3.2 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 (6).

(6) Wenn ich Halsschmerzen bekomme, dann trinke ich immer *Milch*, das lindert die Schmerzen.
(Usually, a glass of *milk* helps me when I got a sore throat.)

2.3.3 Prevention (PREVENT)

PREVENT presents an even stronger positive effect than the relation type *BENEF*. It claims that the consumption of the target food item can prevent the outbreak of a particular disease (7).

---

[6] In the following, we may also refer to category labels as *classes* or *relation types*.

(7) Der Bildung von <u>Nierensteinen</u> kann durch *Zitronensäure* vorgebeugt werden.
(*Citric acid* reduces the chances of <u>kidney stones</u> developing.)

2.3.4 Unsuitable (UNSUIT)

UNSUIT can be considered the negative counterpart of *SUIT*. It describes cases in which the consumption of the target food item is deemed unsuitable. Unsuitability means that one expects a negative effect, that is, a deterioration of the health situation on the part of the person who suffers from a particular health condition (8). However, there is no explicit mention that there is some deterioration or how the deterioration manifests itself. Typically, the speaker just advises against the consumption of the target food item given a particular health condition.

(8) *Eier* sollten während der <u>Schwangerschaft</u> nicht gegessen werden.
(*Eggs* should not be eaten during <u>pregnancy</u>.)

2.3.5 Worsening (WORSEN)

WORSEN is the negative counterpart of *BENEF*. This relation type explicitly says that the consumption of the target food item results in a deterioration (9). Thus, utterances of that type are perceived to be more intense than utterances of type *UNSUIT*. The type of deterioration may also be mentioned.

(9) Da bringt *Rotwein* bei mir den <u>Heuschnupfen</u> so richtig zur Geltung.
(If I now drink *red wine*, this will heavily increase the symptoms of my <u>hay fever</u>.)

2.3.6 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) Es ist allgemein bekannt, dass *Cola* <u>Karies</u> verursacht.
(It's a common fact that the regular consumption of *coke* causes <u>caries</u>.)

2.3.7 Embedded Relations (EMBREL)

EMBREL describes cases in which one of the previous six relation types are embedded in a context so that one cannot conclude that this relation type holds. Typical embeddings are questions (11), irrealis (12) or irony (13).

(11) Weiß jemand, wie das mit *Tofu* und <u>Schwangerschaft</u> ist?
(Does anyone know whether I can eat *tofu* during my <u>pregnancy</u>?)
(12) Wenn man von *Schokolade* <u>Akne</u> bekäme, würde ich aufhören, sie zu essen.
(If *chocolate* caused <u>acne</u>, I would stop eating it.)
(13) Vom *Pute* essen wird man schon keine <u>Kopfschmerzen</u> kriegen, da sie mit Schmerzmitteln vollgepumpt werden.
(Eating *turkey hen* will prevent you from getting a <u>headache</u>, as they drugged them up to the eyeballs.)

Negated relations are only labeled as *EMBREL* if they cannot be resolved as another unnegated relation type. For instance, if a sentence expresses a negated relation of type *SUIT* (14), then this can in most cases be translated as an unnegated case of *UNSUIT* (15).

(14) Ich finde, Du solltest mit Deiner <u>Diabetes</u> keinen *Kuchen* essen.
     (I don't think that you should eat *cake* with your <u>diabetes</u>.)
(15) Ich finde, dass *Kuchen* für Dich mit Deiner <u>Diabetes</u> ungeeignet ist.
     (I think that *cake* is unsuitable for you with your <u>diabetes</u>.)

However, a negated case of *CAUSE* (16), for example, cannot be associated with any other existing category.

(16) Du wirst von *Eiern* schon kein <u>Durchfall</u> bekommen (allerdings wenn Du ihn bereits hast, würde ich Dir vom Verzehr abraten).
     (You won't get <u>diarrhoea</u> from eating *eggs* (but if you contracted it I wouldn't recommend eating them either).)

If the interrelation between target food item and health condition is expressed with some degree of uncertainty (17); the relation involves an additional restriction, for example with regard to the quantity (18); or the relation is reported by the speaker but there is no definite indication that they share that view (19), then this is not considered as a type of embedding. (In other words, Sentences (17) and (18) would still be labeled *CAUSE* while Sentence (19) would be labeled *UNSUIT*.) This is due to the fact that the majority of interrelations are fairly weak. Usually, they are only observed with a smaller number of people. As a consequence of that, it is quite rare that a speaker presents the interrelation between a food item and a particular health condition as a definite fact (that always holds).

(17) Von *rotem Fleisch* **kann** man <u>Herzkrankheiten</u> bekommen.
     (Eating *red meat* **may** cause <u>heart diseases</u>.)
(18) Wenn man **viel** *Alkohol* trinkt, verursacht das <u>Bluthochdruck</u>.
     (Drinking **a lot of** *alcohol* causes <u>hypertension</u>.)
(19) **Irgendwo habe ich mal gelesen**, dass *Obst* bei <u>Diabetes</u> nicht empfohlen wird.
     (**I've read somewhere** that *fruit* is not recommended for people suffering from <u>diabetes</u>.)

2.3.8 No Relation (NOREL)

While in all previously discussed cases the target food item and health condition are somehow related, there are cases in which the co-occurrence is merely co-incidental (20). On our dataset, this is actually the most frequent category (Table 2).

(20) Sein Problem ist gar nicht so sehr die <u>Diabetes</u>, ich mache mir eher Sorgen um das *Fett* (das er zu sich nimmt).
     (It's not his *diabetes* I'm concerned about but the enormous amounts of <u>fat</u> (that he consumes).)

Table 3 Some example relations (*food item*, *health condition*) from the gold standard illustrating the wide knowledge contained in our text corpus.

Relation Type	Instances
SUIT	(Mandeln/almonds, Neurodermitis/dermatitis); (Lamm/lamb, Neurodermitis/dermatitis); (Frischkornbrei/fresh grain porridge, Schwangerschaft/pregnancy); (Stevia/stevia, Diabetes/diabetes); (Kartoffelpüree/mashed potatoes, Halsschmerzen/sore throat)
BENEF	(Kaugummi/chewing gum, Sodbrennen/heartburn); (Kokosflocken/coconut flakes, Sodbrennen/heartburn); (Grünkohl/green cabbage, Kalziummangel/acalcinosis); (Ingwer/ginger, Übelkeit/nausea); (Pflaumensaft/prune juice, Verstopfung/constipation)
PREVENT	(Zitronensäure/citric acid, Nierensteine/kidney stones); (Grapefruit/grape fruit, Diabetes/diabetes); (Fenchel/fennel, Calciummangel/acalcinosis); (Vollkornprodukte/whole grain products, Karies/caries)
UNSUIT	(Mohn/poppy seeds, Schwangerschaft/pregnancy); (Lauch/leek, Schwangerschaft/pregnancy); (Rhababer/rhubarb, Rheuma/rheumatism); (Ingwer/ginger, Fieber/fever); (Dinkel/spelt, Durchfall/diarrhoea)
WORSEN	(Milch/milk, Husten/cough); (Kaffee/coffee, Blasenentzündung/cystitis); (Schokolade/chocolate, Neurodermitis/dermatitis); (Rotwein/red wine, Schnupfen/coryza); (Kaffee/coffee, Akne/acne)
CAUSE	(Hefe/yeast, Sodbrennen/heartburn); (Safran/saffron, Kopfschmerzen/headache); (Honig/honey, Asthma/asthma); (Holunderbeere/elderberry, Durchfall/diarrhoea); (Früchtetee/fruit-infused tea, Sodbrennen/heartburn); (Flohsamen/psyllium, Verstopfung/constipation)

Table 4 Division of coarse-grained classes into fine-grained classes.

Coarse-grained Class	Fine-grained Classes
Conditional Healthy	SUIT, BENEF, PREVENT
Conditional Unhealthy	UNSUIT, WORSEN, CAUSE

Finally, Table 3 displays a set of relations our dataset contains. We deliberately also include less common relations in order to support our claim that the textual source we chose for this task is suitable for knowledge acquisition. If our dataset only contained typical relations that are common knowledge – by that we mean relations, such as *CAUSE(sugar, caries)*, *BENEF(chicken broth, flu)* or *UNSUIT(alcohol, pregnancy)* – this would be hardly convincing.

2.4 Coarse-grained Classification

In the beginning of this article, we introduced the two main categories, *conditional healthiness* and *conditional unhealthiness*. Table 4 lists the fine-grained (sub)classes for each of these categories that have just been defined in the previous sections. In this article, we will focus on extracting instances of the fine-grained classes. We think that the different fine-grained categories belonging to the same main categories possess quite different properties that require different kinds of features. By focusing on the main categories these different properties would be less obvious.

3 Feature Design

In the following, we will discuss the different features we employ in this work. The features are divided in the following groups: word-based features (Section 3.1), task-specific linguistic features (Section 3.2), generic linguistic features (Section 3.3), sentiment features (Section 3.4), features derived from a healthiness lexicon (Section 3.5), food and health condition priors (Section 3.6) and manually extracted keywords (Section 3.7).

3.1 Word-based Features

We use simple bag-of-words features. The words that we encode are the words between the mention of the target food item and the target health condition, and the words immediately preceding and following these two expressions.

3.2 Task-specific Linguistic Features

The task-specific linguistic features we use are listed in Table 5. The initial linguistic features in that table are configurational cues that are mostly designed to indicate whether there is some relationship between target food item and health condition. The co-occurrence within the same clause is usually a good predictor. There are three features to establish this property. While *scope* operates on the syntactic parse output[7], *boundary* and *otherFood* are determined on the token level. The quality of parsing is likely to be affected by the heavy noise in our data (forum entries often contain spelling and grammar mistakes and tend to be fragmented), so some back-off features (i.e. *boundary* and *otherFood*) may be needed. With the feature *prom* we want to investigate whether the target food item is more likely to be involved in a relation if it is in a prominent sentence position, e.g. the beginning of the sentence. A negative counterpart of *prom* is *side*. Possibly, side dishes are less within the focus, so they may be less likely to be involved in a relation. For health conditions, we did not observe similar contexts that would make prominence features (similar to *prom* and *side*) necessary. Therefore, we do not include such features for health conditions. *foodBefCond* takes into consideration the order in which the target food item and health condition appear. This textual order may also reflect the temporal order of events, that is, first a food item is consumed, and then there is some impact on the health condition, e.g. some illness may break out. We do not assume that there is a general correspondence between textual and temporal order, but for some relation types there may be some tendency for a particular textual order.

Table 5 also contains features to detect contextual phenomena that discard the validity of a relation that is embedded. We focus on those types of contextual embeddings that have been presented in Section 2.3.7. For the detection of the feature *question*, we rely on typical surface cues, i.e. *?* and the various interrogative pronouns. In order to detect *irrealis* constructions we scan the target sentence for typical

---

[7] We adopt the definition of *semantic scope* from Wiegand and Klakow (2010).

cue words (e.g. *wenn (if)* or *ob (whether)*). For negation modeling, i.e. *negFood* and *negCond*, we mainly adapted to German the lists of negation words and the scope modeling from Wilson et al (2005). From a semantic point of view *againstCond* is used in similar cases as *negCond*. We insert a pattern with this special construction as it is not captured by the negation words from Wilson et al (2005).[8] We assume that the reason for its omission is that this type of construction is very domain or task specific. On our dataset, it is fairly frequent. (It occurs in 180 target sentences.) In order to detect irony, we scan sentences for *weird* words typically containing sequences of reduplicated characters (e.g. *sooooo* or *seeeeeeehr*). This is admittedly only a very shallow feature, but, on the other hand, we only encountered few cases of irony on our dataset.

The features of the third group have in common that they count the frequency of certain types of words. (Unlike the features of the previous group that also made use of look-up lists, these features do not address contextual embedding.) Each of these features comes in two versions. One considers the corresponding cues in the target sentence only (those features carry the suffix *-TS*) while another considers the entire context (those features carry the suffix *-EC*), that is, in addition to the target sentence it takes into account the two sentences preceding and following the target sentence (Figure 1). By this distinction we want to examine which contextual window is most informative for this classification task. *synoHlth(EC|TS)* and *synoUnh(EC|TS)* count the number of near synonyms of the word *healthy* and the word *unhealthy*. These lexicons were taken from Wiegand and Klakow (2013a). *synoHlth(EC|TS)* uses 65 cue words while *synoUnh(EC|TS)* uses 33 cue words. As there have been previous attempts to extract causative relations (Girju and Moldovan, 2002; Girju, 2003; Beamer and Girju, 2009; Kozareva, 2012), we could also employ previously introduced lexical resources in order to detect our class *CAUSE*. We translated the cues from Girju (2003) to German. The resulting lexicon contains 49 expressions.[9] With regard to the remaining relation types, there do not exist comparable lexical resources we could make use of. Presumably, this is due to the fact that, unlike *CAUSE*, these are fairly task-specific relation types. Finally, we also count the mentions of diseases (health conditions) in the context of the target food item co-occurring with the target health condition. Possibly, the mention of several diseases in the context of such co-occurrences is indicative of a particular relation type. For the detection of mentions of diseases, we semi-automatically compiled a look-up list mainly relying on the web. In order to have a sufficient coverage on our web forum entries, we made sure that for (very) technical terms, we also have informal expressions (e.g. *Grippe (flu)* for *Influenza (influenza)*) that are more likely to be used in every-day language (as can be found in our corpus). The final list of diseases contains 411 expressions.

---

[8] Negation will also be discussed in our error analysis, particularly in the context of sentiment features (Section 5.3) and in the context of the class *UNSUIT* (Section 5.6) (for *UNSUIT* negation is an important predictor).

[9] As there is not a one-to-one mapping between the English keywords from Girju (2003) and their German counterparts, the size of the English and the German lists varies slightly.

Table 5 Description of the linguistic feature set.

Feature	Illustration/Further Information
scope: Are target food item and health condition within the same semantic scope?	*Obwohl sie sehr nahrhaft sind, können Erdbeeren doch eine negative Wirkung bei Menschen mit Neurodermitis hervorrufen.*
	(Strawberries, though they are very nutritious, can have negative effects on people with dermatitis.)
boundary: Is there punctuation mark between target food item and health condition?	*Zum Frühstück habe ich eine Banane gegen meinen Durchfall gegessen, aber ich hätte lieber weiße Bohnen$_{target}$ gegessen.*
	(For breakfast I had a banana against my diarrhoea, but I would have preferred baked beans$_{target}$.)
otherFood: Is there another food item between target food item and health condition?	*Ich war im Supermarkt, um frisches Obst$_{target}$, Gemüse, Eis gegen meine Halsschmerzen und einen Braten für das Mittagessen am Sonntag zu kaufen.*
	(I went to the supermarket to buy some fresh fruits$_{target}$, vegetables, some ice cream against my sore throat, and also some roasted meat for our Sunday dinner.)
prom: Is target food item in a prominent position?	Prominent positions: i.e. beginning/end of a sentence/subclause
side: Is target food item used as a side dish?	*Fischstäbchen mit Kartoffelbrei$_{target}$ sind vielleicht nicht das Richtige für Dich mit Deinem Sodbrennen.*
	(Fish fingers with mashed potatoes$_{target}$ may not be the right dish for you, if you suffer from heartburn.)
foodBefCond: Does target food item occur before target health condition?	*Gestern hatte ich Tiramisu zum Nachtisch, und seit heute Morgen leide ich unter furchtbarem Durchfall.*
	(Yesterday, I had tiramisu for dessert, and this morning I got some terrible diarrhoea.)
question: Is target sentence a (direct) question?	*Denkst Du wirklich, dass ich diese Banane nicht mit meinem Magengeschwür essen sollte?*
	(Do you really think that I shouldn't eat this banana with my stomach ulcer?)
irrealis: Is (assumed) relation embedded in irrealis context?	*Wenn Bohnen nicht gut bei Diabetes wären... (If beans were unsuitable for people with diabetes ...); Sie fragen sich, ob Bohnen bei Diabetes erlaubt sind. (They wonder, whether beans are suitable for people with diabetes.)*
negFood: Is target food item negated?	*Du solltest keine Vollkornprodukte essen, wenn Du einen Reizdarm hast.*
	(You should not eat any wholemeal products if you suffer from irritable bowel syndrome.)
negCond: Is target health condition negated?	*Seit ich morgens regelmäßig Flohsamen esse, habe ich keine Verstopfung mehr.*
	(Since I have regularly been taking psyllium in the morning, I have had no more constipation.)
againstCond: Is target health condition preceded by *against*?	*Ich rate Dir zu eine Hühnerbrühe gegen Dein Fieber.*
	(I recommend that you have a hot chicken broth against your fever.)
weird: Is there an occurrence of a *weird* word?	*Sicher doch, Schokolade ist das aaaaallerbeste bei Deinem Kampf gegen Übergewicht.*
	(Sure, chocolate is the veeeeery best you can eat if you struggle with overweight.)
synoHlthEC/synoHlthTS: Number of near synonyms of *healthy* in entire context/in target sentence only	examples: *vitaminreich (high in vitamin)*, *heilsam (healing)*, *gesundheitsstärkend (tonic)*, etc.
synoUnhEC/synoUnhTS: Number of near synonyms of *unhealthy* in entire context/in target sentence only	examples: *krebserregend (carcinogenic)*, *schädlich (harmful)*, etc.
causeEC/causeTS: Number of causation cues in entire context/in target sentence only	lexical cues from Girju (2003) (translated into German): *führen zu (give rise to)*, *hervorrufen (induce)*, *erzeugen (produce)*, *bewirken (bring about)* etc.
diseaseEC/diseaseTS: Number of diseases in entire context/in target sentence only	usage of a look-up list of diseases mainly created with the help of the web

Table 6 Description of the generic linguistic features (illustration for example sentence: *Empfindliche Menschen bekommen Durchfall vom Verzehr von Holunderbeeren. (Some sensitive people get diarrhoea from eating elderberries.)*)

Subgroup	Feature	Illustration
pos	part-of-speech sequence between target food item and health condition	APPRART_NN_APPR
	part-of-speech tag preceding target food item	$APPR_{beforeFood}$
	part-of-speech tag following target food item	$\$._{afterFood}$
	part-of-speech tag preceding target health condition	$VVFIN_{beforeCond}$
	part-of-speech tag following target health condition	$APPRART_{afterCond}$
path	path on the syntactic parse tree from target food item to target health condition	↑NP_↑PP_↑NP_↑PP_↑VP_↓NP

3.3 Generic Linguistic Features

Apart from the task-specific linguistic features, we will also consider common linguistic features that are often employed in relation extraction tasks. They are listed in Table 6. On the one hand, we consider part-of-speech information (Bunescu and Mooney, 2005; Kessler and Nicolov, 2009; Mintz et al, 2009). More precisely, we consider the part-of-speech sequence between the target food item and the target health condition. In addition, we also include the part-of-speech tags of the words immediately preceding and following the target expressions. (We do not include the part-of-speech tags of the target food item and the target health condition since, being common nouns, they are always the same. Therefore, nothing can be learnt from that information.) On the other hand, we consider the path on the syntactic parse tree from the target food item to the target health condition (Gildea and Jurafsky, 2002; Kessler and Nicolov, 2009; Mintz et al, 2009).

3.4 Features Derived from a Sentiment Lexicon

The two main category labels, i.e. conditional healthiness and unhealthiness, resemble very much the labels from polarity classification (Wilson et al, 2005). In this subtask of sentiment analysis one needs to distinguish between positive and negative opinions. Both problems (i.e. conditional healthiness and polarity classification) can be considered a binary classification problem comprising two opposing (polar) classes. Due to this similarity, we want to investigate whether lexical cues from polarity classification can be harnessed as features for our task. The cues we consider in this article are so-called *polar expressions*, i.e. words conveying either positive polarity, such as *gut (good), angenehm (pleasant)* or *herrlich (superb)*, or negative polarity, such as *schlecht (bad), unangenehm (awkward)* or *schrecklich (horrible)*. They are obtained from the sentiment lexicon underlying the *PolArt* system (Klenner et al, 2009). This lexicon has been manually created and contains 7854 polar expressions. The lexicon also assigns an intensity score to each polar expression varying from 0.5 to 1.0 where the latter denotes very strong intensity. For our features we specially mark the strong polar expressions, that is, the expressions that have been assigned a score of 1.0.

Table 7 Features derived from the sentiment lexicon (Klenner et al, 2009).

Subgroup	Feature	Abbreviation
polarCont	Number of positive polar expressions in target sentence only	posTS
	Number of negative polar expressions in target sentence only	negTS
	Number of strong positive polar expressions in target sentence only	sPosTS
	Number of strong negative polar expressions in target sentence only	sNegTS
	Number of positive polar expressions in entire context	posEC
	Number of negative polar expressions in entire context	negEC
	Number of strong positive polar expressions in entire context	sPosEC
	Number of strong negative polar expressions in entire context	sNegEC
polarSynt	Is target food item syntactically related to positive polar expression?	posFoodSynt
	Is target food item syntactically related to negative polar expression?	negFoodSynt
	Is target food item syntactically related to strong positive polar expression?	sPosFoodSynt
	Is target food item syntactically related to strong negative polar expression?	sNegFoodSynt
	Is target health condition syntactically related to positive polar expression?	posCondSynt
	Is target health condition syntactically related to negative polar expression?	negCondSynt
	Is target health condition syntactically related to strong positive polar expression?	sPosCondSynt
	Is target health condition syntactically related to strong negative polar expression?	sNegCondSynt

The entire set of sentiment features we use is displayed in Table 7. Overall, these features can be divided into two main groups: *polarCont* comprises features that count the occurrences of different types of polar expressions in the entire context or just the target sentence, while *polarSynt* is a set of features that considers polar expressions that are (directly) syntactically related to either the target food item or the target health condition. If, for example, the target health condition is a prepositional object of a polar expression, as *Zahnschmerzen (toothache)* is of *Hausrezept (remedy)* in Sentence (21), then one or more features from the group *polarSynt* fire. We add *polarSynt* in order to find out whether sentiment information combined with syntactic knowledge is actually necessary. Ideally, the combination is more precise than the plain sentiment features from the group *polarCont*.

(21) *Nelken* sind ein altes **Hausrezept**[+] gegen Zahnschmerzen.
      (*Cloves* are an old **home remedy**[+] against toothache.)

The sentiment features we employ in this article also resemble some of our task-specific linguistic features (Section 3.2), namely *synoHlth(EC|TS)* and *synoUnh(EC|TS)*. However, there is some crucial difference between these two types of features: Our sentiment features are derived from a domain-independent lexical resource while the other features are very domain specific. Therefore, the sentiment lexicon with its 7854 entries is much larger than the look-up lists for the task-specific linguistic features (they comprise 97 entries in total). Despite the difference in size, the domain-specific look-up lists contain only 40% of the expressions that can also be found in the domain-independent sentiment lexicon. Apparently, there are some important domain-specific expressions missing in the sentiment lexicon (e.g. *krebserregend (carcinogenic)* or *heilkräftig (medicative)*). So, even though the domain-specific features include smaller look-up lists, they may still be effective for our task.

3.5 Features Derived from a Healthiness Lexicon

A further group of features incorporates the knowledge of healthiness of food items. This knowledge is obtained from a lexicon introduced in Wiegand et al (2012c) which covers 2803 food items. Each food item is specified as being either *rather* or *definitely* (un)healthy. The healthiness judgment has been carried out based on the general nutrient content of each food item. It has been created totally independently of the annotation of our text corpus. (For a more detailed description of the annotation scheme, we refer the reader to Wiegand et al (2012c).) Each food item has been annotated by two annotators. Along their individual annotation, a third (adjudicated) annotation has been produced. For our experiments, we use the latter. We consider a food item *(un)healthy* if it is either classified as *rather* or *definitely* (un)healthy. In addition, we also have additional features for *definitely* (un)healthy food items.

Even though the healthiness labels that can be found in the healthiness lexicon bear some resemblance to the two main categories, i.e. conditional healthiness and unhealthiness (Section 2.4), they are not the same. The healthiness lexicon does not encode *conditional* healthiness, i.e. healthiness with regard to a particular health condition, but *prior* healthiness, i.e. healthiness of a food item per se. To further show that these two concepts are not the same, Table 8 displays the 20 food items that most strongly correlate with the conditional healthiness class (i.e. *SUIT*, *BENEF* and *PREVENT*) and the conditional unhealthiness class (i.e. *UNSUIT*, *WORSEN*, *CAUSE*). Correlation was determined with the help of *Pointwise Mutual Information* (Church and Hanks, 1990). We denote the healthiness status according to the healthiness lexicon by $+$ for *healthy* and $-$ for *unhealthy*, respectively. While we observe some notable agreement between prior and conditional healthiness, there is some heavy disagreement between prior and conditional unhealthiness. *Milchprodukte (dairy products)*, for example, are mostly considered healthy due to their nutrient content. However, for people with specific health conditions, e.g. *Neurodermitis (dermatitis)*, they can actually be harmful. Similarly, *Muscheln (mussels)* and *Pilze (mushrooms)* are nutritious but for people with a sensitive stomach, they can cause nausea or bowel complaints.

The specific features derived from that healthiness lexicon are listed in Table 9. They are divided into two groups. *healthTarget* describes the healthiness of the target food item. These features are designed to check in how far conditional healthiness correlates with the prior healthiness of the target food item. The second group of features, *healthCont*, encompasses the prior healthiness status of *neighbouring* food items in the given context of an instance. Thus, we want to find out in how far the prior healthiness knowledge about neighbouring food items helps to predict the conditional healthiness of the target food item. While there is definitely not a perfect correlation between prior and conditional healthiness when individual food items are considered (as shown in Table 8), it may still be the case that there is a stronger correlation between conditional healthiness and groups of healthy expressions co-occurring with each other rather than just a mention of a single healthy expression in a sentence. Moreover, for unknown target food items (i.e. food items that are not contained in the healthiness lexicon) these features might also be helpful (provided that there are mentions of known food items within the same instance) while the features from

Table 8 Top 20 food items that most strongly correlate with conditional healthiness and unhealthiness (+ and − denote the prior healthiness according to the healthiness lexicon (Wiegand et al, 2012c); **bold type font** denotes mismatch between prior and conditional healthiness).

Rank	Conditional Healthy	Conditional Unhealthy
1	Hühnersuppe (chicken broth) +	Alkohol (alcohol) −
2	Flohsamen (psyllium) +	Süßstoff (aspartame) −
3	Zwiebelsaft (onion juice) +	**Rhabarber (rhubarb)** +
4	Natron (baking soda) *N/A*	**Kuhmilch (cow milk)** +
5	Salbei (salvia) +	Chips (crisps) −
6	Salbeitee (salvia tea) +	**Knoblauch (garlic)** +
7	Kräuter (herbs) +	**Pilze (mushrooms)** +
8	Ingwertee (ginger tea) +	Schweinefleisch (pork) −
9	Tee (tea) +	**Milchprodukt (dairy product)** +
10	Thymiantee (thyme tea) +	**Ei (egg)** +
11	Fenchelhonig (fennel honey) +	Fertiggericht (instant meal) −
12	Ingwer (ginger) +	Geschmacksverstärker (flavour enhancer) −
13	Saft (juice) +	Cocktail (cocktail) −
14	Heilerde (healing earth) *N/A*	**Spargel (asparagus)** +
15	Fenchel (fennel) +	**Quark (curd cheese)** +
16	Suppe (soup) *N/A*	**Muscheln (mussels)** +
17	Mineralwasser (mineral water) +	Weißmehl (white flour) −
18	Paprika (bell pepper) +	**Stevia (stevia)** +
19	**Sirup (syrup)** −	**Roggen (rye)** +
20	Radieschen (radish) +	Sekt (sparkling wine) −

Table 9 Description of the healthiness features; all features are derived from the *healthiness lexicon* (Wiegand et al, 2012c); for the contextual features *healthCont*, the healthiness status of the target food item is not included.

Subgroup	Feature	Abbreviation
healthTarget	Is target food item *a priori* healthy?	targetHlth
	Is target food item *a priori* definitely healthy?	targetDefHlth
	Is target food item *a priori* unhealthy?	targetUnh
	Is target food item *a priori* definitely unhealthy?	targetDefUnh
healthCont	Number of food items that are *a priori* healthy in target sentence only	hlthTS
	Number of food items that are *a priori* definitely healthy in target sentence only	defHlthTS
	Number of food items that are *a priori* unhealthy in target sentence only	unhTS
	Number of food items that are *a priori* definitely unhealthy in target sentence only	defUnhTS
	Number of food items that are *a priori* healthy in entire context	hlthEC
	Number of food items that are *a priori* definitely healthy in entire context	defHlthEC
	Number of food items that are *a priori* unhealthy in entire context	unhEC
	Number of food items that are *a priori* definitely unhealthy in entire context	defUnhEC

healthTarget could not make any prediction. Even though the healthiness lexicon is fairly large, it does not cover very exotic food items, such as *Natron (baking soda)* or *Heilerde (healing earth)*, which are not unlikely to appear as target food items in our dataset.

3.6 Food and Health Condition Priors

We also want to examine in how far there is a correlation between health conditions and our class labels, on the one hand, and a correlation between food items and our

class labels, on the other hand. The purpose of this investigation is twofold. Firstly, we use these correlations as a baseline. For example, if there is a strong correlation between certain health conditions and a particular class label irrespective of a particular target food item, we may wonder whether there is any point in a sophisticated textual analysis. (The same can, of course, be said if there is a strong correlation between certain food items and a particular class label.) In that case, one may equally well extract relations by just extracting food items frequently co-occurring with a particular health condition and then assigning the relation type with which the health condition mostly co-occurs. Secondly, we also want to examine whether the knowledge of these correlations can be usefully combined with the other information learnt from texts.

In order to further motivate feature engineering taking into consideration these types of correlations, we will have a look at the class distributions among the 10 most frequent health conditions and food items, as shown in Tables 10 and 11, respectively. With regard to the relation between health conditions and class labels (Table 10), we find that the class distribution varies quite a lot between the different conditions. In most cases, there is a tendency towards one group of classes, in other words, either the majority of instances belong to the subclasses of conditional healthiness (i.e. *SUIT*, *BENEF* and *PREVENT*) as it is most notably the case for *Husten (cough)* or *Halsschmerzen (sore throat)*, or the majority of instances belong to the subclasses of conditional unhealthiness (i.e. *UNSUIT*, *WORSEN* and *CAUSE*), as it is the case for *Neurodermitis (dermatitis)*. However, we also find that in many cases a significant amount of instances also belong to the other group of class labels. In the case of *Übelkeit (nausea)* there is almost a balance between instances belonging to the subclasses of conditional healthiness and unhealthiness. For a few health conditions, some of the class labels have not been observed at all. This mostly affects the class labels *PREVENT* and *WORSEN*. This can be ascribed to the general sparsity of these two class labels (Table 2). Beyond those two class labels, we only find that for *Schwangerschaft (pregnancy)* there is not a single instance labeled as *CAUSE* which is also absolutely plausible.

As far as the relation between food items and class labels are concerned (Table 11), we find similar results. On the one hand, there are food items which clearly display a tendency to appear with subclasses of conditional healthiness (e.g. *Tee (tea)*, *Honig (honey)* or *Hühnersuppe (chicken broth)*) or subclasses of conditional unhealthiness (e.g. *Alkohol (alcohol)* or *Fett (fat)*), but there are also food items which are almost uniformly distributed throughout the different class labels (most notably *Kaffee (coffee)*, *Fleisch (meat)* and *Milch (milk)*). It is also interesting to see that among some of the food items that have a tendency towards some specific class labels, these class labels need not be consistent with the respective prior healthiness of the food item (Section 3.5). This is most striking for the target food item *Cola (coke)* which strongly co-occurs with conditional healthiness, i.e. *SUIT* and *BENEF*. For example, it is often considered helpful for curing diarrhoea. However, *Cola (coke)* is generally considered to be unhealthy as far as its nutrient content is concerned. As a consequence, using the correlations between food items and our relation types based on observations from our corpus may be more effective for conditional healthiness classification than using a prior healthiness lexicon.

From this statistical analysis, we can conclude that there are tendencies of particular health conditions and food items to co-occur with particular classes or, more precisely, groups of classes. For supervised learning, we introduce two simple features that exploit these correlations. *condPrior* always predicts for each instance in which a particular health condition occurs the class label with which it has mostly been observed on our dataset. For example, every instance of *Durchfall (diarrhoea)* is classified as *CAUSE* (Table 10). Similarly, *foodPrior* always predicts for each instance in which a particular food item occurs the class label with which it mostly co-occurs. For example, every instance of *Tee (tea)* is classified as *SUIT* (Table 11). Table 12 summarizes the two prior features we use in this article.

We already stated in Section 2.1 that while the health conditions chosen on our dataset are, by and large, frequently occurring expressions, there were no restrictions made on the food items we consider. As a consequence, the occurrences of food items will follow a power law distribution (Zipf's law), that is, there are many infrequent food items, i.e. 40% of them are singletons. Further evidence regarding the different distributions of health conditions and food items can be obtained by considering the average number of instances associated with a particular item: The average number of instances associated with a health condition is 64 while it is only 7 for food items. For rare food items, in particular singletons, it does not make sense to determine the most frequently observed class label, as these observations are not very reliable. (They could happen by chance.) Another negative side-effect of including them for our feature *foodPrior* would be that they overfit. For instance, *foodPrior* would predict the correct label for every singleton. In order to measure a more realistic impact of this feature, we restrain the prediction to only those food items which occur at least 10 times on our dataset. This also means that for almost 30% of the instances no prediction can be made by *foodPrior* as these instances contain target food items that are too rare.

3.7 Manually Chosen Keywords

In various text classification tasks, the performance of classifiers can usually be improved by using task-specific lexicons that comprise words that are predictive for the classes that are to be detected. One usually applies such lexicons for features that ask if (or how many) of those predictive words occur in a text passage that is to be classified. Such features are usually less sparse than plain bag-of-words features since they generalize over individual words and may also make predictions for words that have not been observed in the training data (provided the words are contained in the respective lexicons).

For the set of task-specific linguistic features (Section 3.2), we already included such type of features, i.e. *causeTS* and *causeEC* (Table 5). Since in this article, we present a fairly novel task with new class labels, with the exception of the class *CAUSE*, there do not exist any appropriate lexicons tailored to these classes that we could use. As a consequence, it is difficult to create such word lists in an unbiased manner. The only way to create them is with the help of our dataset. However, this may result in overfitting those lexicons to the data on which we test our classifiers.

Table 10 Distribution of the different classes among the 10 most frequent target health conditions.

	Conditional Healthy						Conditional Unhealthy					
	SUIT		BENEF		PREVENT		UNSUIT		WORSEN		CAUSE	
Condition	Freq	Perc	Freq	Perc	Freq	Perc	Freq	Perc	Freq	Perc	Freq	Perc
Durchfall (diarrhoea)	32	15.8	40	19.7	2	1.0	4	2.0	4	2.0	**121**	**59.6**
Sodbrennen (heartburn)	26	14.2	39	21.3	54	2.2	10	5.5	1	0.6	**103**	**56.3**
Kopfschmerzen (headache)	23	13.3	42	24.3	17	9.8	2	1.2	0	0.0	**89**	**51.4**
Husten (cough)	34	23.5	**93**	**61.1**	2	1.4	4	2.8	5	3.5	7	4.8
Halsschmerzen (sore throat)	53	42.4	**67**	**53.6**	0	0.0	3	2.4	1	0.8	1	0.1
Schwangerschaft (pregnancy)	32	28.3	17	15.0	0	0.0	**64**	**56.6**	0	0.0	0	0.0
Diabetes (diabetes)	**50**	**56.2**	7	7.9	1	1.1	26	29.2	1	1.1	4	4.5
Übelkeit (nausea)	9	11.4	30	38.0	1	1.3	3	3.8	0	0.0	**36**	**45.6**
Neurodermitis (dermatitis)	18	26.1	3	4.4	0	0.0	**34**	**49.3**	4	5.8	10	14.5
Fieber (fever)	24	38.1	**29**	**46.0**	0	0.0	3	4.8	0	0.0	7	11.1

Table 11 Distribution of the different classes among the 10 most frequent target food items.

	Conditional Healthy						Conditional Unhealthy					
	SUIT		BENEF		PREVENT		UNSUIT		WORSEN		CAUSE	
Food Item	Freq	Perc	Freq	Perc	Freq	Perc	Freq	Perc	Freq	Perc	Freq	Perc
Tee (tea)	**46**	**48.9**	36	38.3	3	3.2	2	2.1	0	0.0	7	7.5
Kaffee (coffee)	12	16.9	**19**	**26.8**	10	14.1	11	15.5	2	2.8	17	23.9
Honig (honey)	16	25.4	**38**	**60.3**	0	0.0	2	3.2	2	3.2	5	7.9
Alkohol (alcohol)	0	0.0	1	2.0	0	0.0	**29**	**59.2**	0	0.0	19	38.8
Hühnersuppe (chicken broth)	18	45.0	**20**	**50.0**	1	2.5	1	2.5	0	0.0	0	0.0
Fleisch (meat)	7	18.0	6	15.4	7	18.0	**11**	**28.2**	1	2.6	7	18.0
Milk (milk)	3	8.3	**11**	**30.6**	0	0.0	8	22.2	4	11.1	10	27.8
Cola (coke)	**15**	**41.7**	9	25.0	1	2.8	2	5.6	3	8.3	6	16.7
Fett (fat)	3	8.6	1	2.9	3	8.6	8	22.9	0	0.0	**20**	**57.1**
Wein (wine)	5	15.6	2	6.3	0	0.0	1	3.1	0	0.0	**24**	**75.0**

Table 12 Description of the prior features.

Feature	Description
condPrior	Always assign an instance with a particular target health condition the relation type with which it mostly co-occurs.
foodPrior	Always assign an instance with a particular target food item the relation type with which it mostly co-occurs.

Despite this risk, we manually create such lists in that fashion. However, being aware of the fact that the result may be biased, we regard the corresponding features as an *upper bound* for lexical features in general.

For each relation type, we consider two types of features (Table 13). *KW* contains plain features that consider the keywords for the pertaining relation type in the target sentence. In addition, *SyntKW* includes for each relation type a feature that asks whether the target sentence contains at least one keyword (for the pertaining relation type) that is either syntactically related to the target food item or the target health condition. *SyntKW* is analogous to *polarSynt* (Table 7) and has also been included with a similar motivation (Section 3.4).

In order to obtain the keywords, in a first step, each noun, verb or adjective in a target sentence considered to be predictive towards the label that the corresponding instance was assigned was marked. We only annotated unigrams. In a second step, all expressions that had thus been marked were collected, and for feature extraction, any mention of them is considered a keyword. For illustration, Table 14 lists the 10 most frequent keywords for each relation type on our dataset.

Unfortunately, not in all instances such keywords could be identified. Table 15 shows the proportion of instances for which at least one keyword was identified. The categories for which the fewest keywords could be identified are *SUIT*, *PREVENT* and *UNSUIT*. Typical sentences for these categories (in which no keywords could be marked) are Sentences (22)-(24). These examples neither contain keywords with parts of speech other than the ones we consider nor keywords being multiword expressions. In all cases, the relations are inferred. In Sentence (22), the fact that the speaker has *Blasentee (diuretic tea)* in her cupboard is the consequence of regularly suffering from cystitis. One, therefore, concludes that *Blasentee (diuretic tea)* is a suitable food item for this health condition. (We refrained from labeling this instance as *BENEF* as there is no explicit mention that the consumption of that tea actually cures the illness.) In Sentence (23), one infers that the reason why the Eskimo did not suffer from heart diseases is that they mostly ate fish. There is no lexical cue explicitly indicating that causal relationship.[10] Sentence (24) is a rhetoric question. This figure of speech clearly implies some disapproval on the part of the speaker regarding the behaviour of the addressee, that is, the speaker thinks that the consumption of *Marmelade (jam)* is unsuitable for people suffering from diabetes.

---

[10] The fact that the target health condition is negated may indicate that this sentence expresses some type of conditional unhealthiness, however, there is no lexical cue that helps us to identify the subtype *PREVENT*.

Table 13 Features derived from keyword lexicons.

Subgroup	Feature
KW	Number of keywords of relation type *SUIT*
	Number of keywords of relation type *BENEF*
	Number of keywords of relation type *PREVENT*
	Number of keywords of relation type *UNSUIT*
	Number of keywords of relation type *WORSEN*
	Number of keywords of relation type *CAUSE*
SyntKW	Is target food item syntactically related to keyword of relation type *SUIT*?
	Is target food item syntactically related to keyword of relation type *BENEF*?
	Is target food item syntactically related to keyword of relation type *PREVENT*?
	Is target food item syntactically related to keyword of relation type *UNSUIT*?
	Is target food item syntactically related to keyword of relation type *WORSEN*?
	Is target food item syntactically related to keyword of relation type *CAUSE*?
	Is target health condition syntactically related to keyword of relation type *SUIT*?
	Is target health condition syntactically related to keyword of relation type *BENEF*?
	Is target health condition syntactically related to keyword of relation type *PREVENT*?
	Is target health condition syntactically related to keyword of relation type *UNSUIT*?
	Is target health condition syntactically related to keyword of relation type *WORSEN*?
	Is target health condition syntactically related to keyword of relation type *CAUSE*?

(22) Ich habe immer *Blasentee* im Schrank (habe so etwa 3-4 mal jährlich eine Blasenentzündung).
(I always have some *diuretic tea* in my cupboard (usually get cystitis two to three times a year).) *LABEL: SUIT*

(23) Und die Eskimos lebten früher überwiegend von *Fisch*, sie kannten keine Herzkrankheiten.
(And in the past, the Eskimo predominantly consumed *fish*; they did not have any heart diseases.) *LABEL: PREVENT*

(24) Du hast Diabetes und kochst tonnenweise *Marmelade* ein?
(You suffer from diabetes and you are canning tons of *jam*?) *LABEL: UNSUIT*

Instances like Sentences (22)-(24) cannot be modelled reliably by state-of-the-art computational models. By employing the keyword features, we may obtain an estimate of how good we would be able to detect the different relation types with an ideal list of keywords but not being able to deal with inferred relations (such as Sentences (22)-(24)).

4 Experiments

In this section, we describe our experiments to detect the different classes that describe some relationship between target food item and health condition. We only focus on the prediction of the four relation types *SUIT* (Section 2.3.1), *BENEF* (Section 2.3.2), *UNSUIT* (Section 2.3.4) and *CAUSE* (Section 2.3.6). (Instances of the other relation types remain in our dataset and should be considered exclusively as negative data instances.[11]) We do not address the classification of the two negative classes *EMBREL* (Section 2.3.7) and *NOREL* (Section 2.3.8) since people are not interested in sentences in which a food item and a health condition occur but there is no

---

[11] By *negative* (data) instances, we mean those instances that have not been tagged with the relation type that is to be extracted.

Table 14 The 10 most frequent keywords for each class.

SUIT	BENEF	PREVENT	UNSUIT	WORSEN	CAUSE
gut	helfen	vorbeugen	weglassen	fördern	bekommen
(good)	(help)	(prevent)	(leave out)	(increase)	(get)
können	wirken	vermeidbar	verzichten	verschleimen	verursachen
(be allowed to)	(be effective)	(avoidable)	(do without)	(congest)	(cause *v*)
empfehlen	Mittel	vermeiden	meiden	verschlimmern	führen (zu)
(recommend)	(remedy)	(avoid)	(avoid)	(worsen)	(lead (to))
dürfen	Hausmittel	schützen	tabu	reizen	kriegen
(may)	(home remedy)	(protect)	(taboo)	(irritate)	(catch)
vertragen	lindern	protektiv	Problem	umhauen	reagieren
(agree with)	(abate)	(protective)	(problem)	(knock out)	(react)
umstellen	bekämpfen	herzschützend	einschränken	stark	kommen (von)
(switch)	(combat)	(heart-protecting)	(cut back)	(strong)	(come (from))
schwören	lösen	ersparen	Verzicht	schüren	auslösen
(swear by)	(expectorate)	(spare)	(waiver)	(stir up)	(trigger)
rutschen	Heilmittel	abwehren	vermeiden	schleimbildend	machen
(slide down)	(cure)	(combat)	(avoid)	(mucous)	(create)
(es) geht	Medizin	*N/A*[*]	schlecht	ruinieren	hervorrufen
(okay)	(medicine)		(bad)	(ruin)	(give rise to)
geeignet	vertreiben	*N/A*[*]	schädlich	gefährlich	Ursache
(suitable)	(combat)		(harmful)	(dangerous)	(cause *n*)

[*] for *PREVENT* only 8 keywords were manually extracted

Table 15 Proportion of instances with at least one keyword (in target sentence).

SUIT	BENEF	PREVENT	UNSUIT	WORSEN	CAUSE
21.73	56.97	13.51	32.39	73.33	72.20

relationship between them (*NOREL*), or there is a relationship but this is embedded in a context so that one cannot conclude that the relation type holds (*EMBREL*). The remaining two relation types *PREVENT* (Section 2.3.3) and *WORSEN* (Section 2.3.5) are not considered since they are too infrequent (Table 2); neither of these relation types covers more than 3% of the instances. Our initial experiments with these types showed that this size does not suffice for supervised classification. We assume that the sparseness of those two relation types is not an artifact of our dataset but reflects their general distribution. Therefore, ignoring them does not mean overlooking crucial relation types. Presumably, the relation type *PREVENT* is rare since food items are a less reliable means to prevent illnesses from breaking out than, for instance, vaccinations. *WORSEN* is rare since it is actually a variation of *UNSUIT* in which the deterioration is explicitly mentioned. Our comparison of instances of *UNSUIT* and *WORSEN* actually revealed that the food items that were mentioned for a particular health condition do not vary between these two relation types. Therefore, we would recommend subsuming instances of *WORSEN* by *UNSUIT* in future work.[12]

Each instance to be classified is a sentence in which there is a co-occurrence of a target food item and a health condition along its respective context sentences (Section 2.2). The dataset was parsed using the Stanford Parser (Rafferty and Manning,

---

[12] Since the instances labeled as *WORSEN* only cover approximately 1% of our entire dataset (Table 2), we are convinced that the choice of treating them as negative instances or as instances of type *UNSUIT* will not affect the overall results of our experiments.

2008). We carry out a 5-fold cross-validation on our manually labeled dataset. For the supervised classifier, we chose Support Vector Machines (Joachims, 1999). As a toolkit, we use *SVMLight*[13] with a linear kernel.

4.1 Comparison of Task-Specific Features

In our first experiment we look at various task-specific features. By that we mean the task-specific linguistic features (Section 3.2), the sentiment features (Section 3.4) and the features derived from a healthiness lexicon (Section 3.5).

We contrast them with two baseline features being *coocc* and the word-based features *word* (Section 3.1). **coocc** is an unsupervised classifier that considers all instances of our dataset as positive instances (of the class which is examined, i.e. *SUIT*, *BENEF*, *UNSUIT* or *CAUSE*). In other words, this baseline indicates how well the mere co-occurrence of the target food item and the target health condition predicts any of our four classes.[14]

4.1.1 Comparison of Feature Groups

We first compare the performance of different feature groups. Table 16 shows the results of this comparison for each class. If we use either the sentiment features (i.e. *polarWord* and *polarSynt*) or the features derived from the healthiness lexicon (i.e. *healthTarget* and *healthCont*) in isolation, for many classes the resulting performance remains as the simple baseline *coocc*. For the healthiness features, this may come as a surprise, particularly if one recalls that there is some correspondence between conditional healthiness and the prior healthiness as shown in Table 8. One should recall, however, that our dataset does not exclusively consist of sentences in which the target food item is claimed to be suitable or unsuitable given a particular health condition (in that case, the scores produced by the healthiness features would presumably have been higher). There is also a large number of sentences in which the co-occurrence of the target food item and health condition does neither express suitability nor unsuitability, that is, these are instances of the types *NOREL* or *EMBREL* (Table 2). For the exclusion of those cases, the health features are totally ineffective.

Table 16 also shows that considering four fine-grained classes (i.e. *SUIT*, *BENEF*, *UNSUIT* and *CAUSE*) instead of two coarse-grained (i.e. *suitability* and *unsuitability*) is appropriate, since different features are effective for the different fine-grained classes. This mostly concerns the two relation types *SUIT* and *BENEF* (which would be merged to the class *suitability*). As far as *BENEF* is concerned, both the sentiment and the healthiness features beat the baseline feature *coocc*. We have following explanations: If a food item is classified as *BENEF*, the food item has some curing properties. In order to express this property, one has to use polar expressions. In addition, having some curing function also usually coincides with healthiness. The situation is different for *SUIT* which to our definition in Section 2.3.1 means that one does not

---

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

[14] We also experimented with variations restricting the co-occurrence to a fixed window size. However, we did not obtain better classifiers than with the plain (sentence-wise) co-occurrence.

expect a negative effect on the health of an afflicted person. In terms of verbalizing such weaker property, we assume less usage of polar expressions. In addition, a food item that is not harmful is less likely to be healthy (in general) than a food item with a curing function. In Table 15, we found that the proportion of keywords within the set of instances labeled as *BENEF* is much larger than the proportion within instances labeled as *SUIT*. This is consistent with our observation that for *BENEF* polar expressions have a notable impact on the classification performance as polar expressions are some kind of keywords.

As far as *UNSUIT* and *CAUSE* are concerned, the healthiness features do not work at all. In Table 8, we already found that conditional unhealthiness correlates with prior unhealthiness much less than conditional healthiness does with prior healthiness. The sentiment features only have a marginal positive effect on *UNSUIT* and *CAUSE* which means that negative polar expressions are not that indicative of conditional unhealthiness, either.

The simple word-based features *word* are much more effective than the sentiment and healthiness features. On all relation types *coocc* is beaten by a large margin. The same can be said about our task-specific features *ling*. In addition, the combination of the two types of features also results in some improvement. Adding the healthiness and sentiment features, however, has no notable effect. From these results, we conclude that a contextual analysis (as enabled by the features *word* and *ling*) is more effective than focusing on external knowledge (as the healthiness and sentiment features do) in this classification task.

4.1.2 Comparison of Individual Features

We now turn to the performance of individual features. Table 17 shows the most highly ranked task-specific features for the four different classes according to Chi-Square ranking. The feature ranking was carried out with *WEKA* (Witten and Frank, 2005). We consider the individual features from the three different high-level feature types, that is, sentiment features, healthiness features and linguistic features.

Even though we find features from all feature groups on those rankings, the top 3 ranks are always linguistic features. The healthiness features are pretty rare on *SUIT* and *BENEF*. Among those few that occur are the features that describe the healthiness of the target rather than the healthiness of contextual food items. This implies that the healthiness status of neighbouring food items is less relevant. The sentiment features are exceptionally frequent on the highest ranks of *BENEF*. This is consistent with Table 16. The strongest sentiment features are those that check whether there is a syntactic relationship between some polar expression and the target food item or health condition.

In terms of *absolute* score of the individual features, we find that even the most highly ranked features of *SUIT* are much weaker than their counterparts of the other relation types. For *BENEF*, there are five features with an exceptionally high score: Two features look for synonyms of the word *healthy*, i.e. *synoHlth(EC|TS)*, one additional feature checks whether the health condition is preceded by *against*, i.e. *againstCond*, and two further sentiment features check whether the target health condition is syntactically related to some positive polar expression, i.e. *sPosCondSynt* and

Table 16 Performance of different feature sets (*word*: word-based features (Section 3.1); *ling*: task-specific linguistic features (Table 5)).

Features	SUIT			BENEF			UNSUIT			CAUSE		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
coocc	16.7	**100.0**	28.6	19.1	**100.0**	32.1	9.5	**100.0**	17.3	18.4	**100.0**	31.1
polarWord	17.7	89.8	29.6	25.4	76.8	38.2	9.5	100.0	17.3	22.1	65.5	33.0
polarSynt	17.9	93.0	30.1	38.5	41.8	40.1	9.8	88.3	17.7	20.1	96.1	33.2
polarWord+polarSynt	17.5	87.4	29.2	25.6	73.5	37.9	10.2	88.3	18.3	20.0	92.5	33.3
healthTarget	16.7	**100.0**	28.6	21.2	89.0	34.2	9.5	**100.0**	17.3	18.4	**100.0**	31.1
healthCont	16.7	**100.0**	28.6	20.4	96.7	33.7	9.5	**100.0**	17.3	18.4	**100.0**	31.1
healthTarget+healthCont	16.7	**100.0**	28.6	22.6	86.8	35.8	9.5	**100.0**	17.3	18.4	**100.0**	31.1
word	31.6	52.6	39.5	59.1	59.8	59.5	37.4	31.7	34.3	52.8	53.9	53.4°
ling	24.5	82.7	37.8	**60.0**	49.9	54.5	34.6	50.0	40.9‡	34.0	73.6	46.5
word+ling	**37.3**	56.6	**45.0**•‡	53.7	70.0	60.8	**49.8**	41.3	**45.1**°‡	**55.6**	63.1	**59.1**•†
word+ling+healthTarget+healthCont+polarWord+polarSynt	36.1	56.2	43.9°‡	57.9	64.6	**61.1**	41.6	44.6	43.1°‡	50.7	71.5	**59.3**•†

significantly better than *ling* at °: $p < 0.1$, •: $p < 0.05$; significantly better than *word* at †: $p < 0.1$, ‡: $p < 0.05$ (based on paired t-test)

posCondSynt. The only feature with an exceptionally high correlation score for *UN-SUIT* is the feature checking whether the food item is negated, i.e. *negFood*. If we revisit the most frequent keywords for that class (Table 14) we find that many of these words (e.g. *weglassen (leave out)*, *verzichten (do without)*, *vermeiden (avoid)* and *tabu (taboo)*) convey a request to stop consuming the target food item. These are *shifters* (Wilson et al, 2005), i.e. (predominantly) verbs, nouns and adjectives that express some negation but without being traditional negation expressions. Due to the semantic similarity between shifters and negation words, we can conclude that the strong predictiveness of *negFood* is consistent with these shifter keywords. (We will also discuss the importance of negation in our error analysis in Section 5.6.) The fact that the *cause*-keywords from Girju et al (2003), i.e. *cause(EC|TS)*, are exceptionally predictive for *CAUSE* comes as no surprise and proves that this domain-independent list of cues is also predictive for our task.

We find it interesting that another exceptionally strong feature for the relation type *CAUSE* checks whether the target food item precedes the health condition in the target sentence, i.e. *foodBefCond*. Apparently, the temporal order of events that underlie a causation relation, i.e. the consumption of some food item followed by the outbreak of a disease, is reflected in the order in which these events are mentioned in a written text. To further substantiate this hypothesis, we also measured the proportion of cases in which the target food item preceded the health condition on all instances and on instances labeled as *CAUSE*. The result is displayed in Table 18. It indeed shows that while, in general, there is no notable tendency for a particular order (55.4%), in 78.8% of the cases labeled as *CAUSE*, however, the food item precedes the health condition.

In Section 3.4, we pointed out the similarity between the sentiment features and the task-specific linguistic features *synoHlth(EC|TS)* and *synoUnh(EC|TS)*. Table 17 gives evidence that the small set of these domain-specific expressions is more effective than the sentiment features (*synoHlthTS* is the strongest feature for *SUIT* and *BENEF* while *synoHlthEC* is the second strongest feature for *UNSUIT*).

On all feature rankings, we observe both features that consider the entire context (suffix *-EC*) and features that exclusively consider the target sentence (suffix *-TS*). This means that it is inconclusive which contextual scope is most informative for this task.

Table 19 shows the best performing feature subset for each relation type using a best-first forward selection. For this feature selection, we, again, use *WEKA*. This form of feature selection is complementary to that depicted in Table 17 as it excludes redundant features. It supports our previous observation that the sentiment and healthiness features are much less important than the task-specific linguistic features. No relation type has more than one sentiment and healthiness feature (each) on the list. The best performing feature set for *UNSUIT* even exclusively comprises linguistic features. Apart from that, some features seem to be much more important than their Chi-Square ranking position suggests. This is especially true for *irrealis*, which appears on three of the four lists. Although its individual predictiveness towards the different classes is marginal (Table 17), the information it encodes is unique and cannot be obtained by other existing features (Table 19).

Table 17 Top 15 task-specific features according to Chi-Square feature ranking for each individual class.

SUIT		BENEF		UNSUIT		CAUSE	
Feature	**Score**	**Feature**	**Score**	**Feature**	**Score**	**Feature**	**Score**
synoHlthTS$^\diamond$	46.1	synoHlthTS$^\diamond$	315.3	negFood$^\diamond$	262.3	causeTS$^\diamond$	172.6
foodBefCond$^\diamond$	43.5	againstCond$^\diamond$	310.7	synoHlthEC$^\diamond$	18.8	foodBefCond$^\diamond$	141.1
againstCond$^\diamond$	31.4	synoHlthEC$^\diamond$	253.8	againstCond$^\diamond$	18.0	causeEC$^\diamond$	114.1
negFood$^\diamond$	25.5	sPosCondSynt*	238.5	sPosCondSynt*	15.7	targetDefUnh$^\bullet$	62.9
causeTS$^\diamond$	21.5	posCondSynt*	212.3	causeEC$^\diamond$	14.8	negTS*	62.4
posCondSynt*	20.5	sPosTS*	93.4	foodBefCond$^\diamond$	14.5	synoHlthEC$^\diamond$	60.2
sPosCondSynt*	19.4	sNegTS*	90.9	synoUnhEC$^\diamond$	13.6	negEC*	53.9
sPosTS*	18.1	sPosFoodSynt*	90.4	targetUnh$^\bullet$	11.8	synoHlthTS$^\diamond$	51.6
posTS*	14.8	negFood$^\diamond$	60.9	synoHlthTS$^\diamond$	11.1	targetUnh$^\bullet$	50.9
synoHlthEC$^\diamond$	14.0	sPosEC*	47.4	targetDefUnh$^\bullet$	10.9	targetHlth$^\bullet$	41.8
causeEC$^\diamond$	12.8	posFoodSynt*	44.7	defUnhTS$^\bullet$	10.5	posEC*	41.7
targetUnh$^\bullet$	10.6	sNegEC*	36.2	causeTS$^\diamond$	10.5	againstCond$^\diamond$	38.6
sNegTS*	10.5	diseaseEC$^\diamond$	31.7	sNegEC*	9.9	posCondSynt*	35.2
negCond$^\diamond$	9.7	targetDefUnh$^\bullet$	29.6	sPosEC*	8.9	negFoodSynt*	34.5
irrealis$^\diamond$	9.5	posTS*	28.8	defUnhEC$^\bullet$	7.7	sPosFoodSynt*	33.9

$^\diamond$: linguistic features (Table 5); *: sentiment features (Table 7); $^\bullet$: features derived from healthiness lexicon (Table 9)

Table 18 Comparison of order between food item and health condition.

Instances	Percentage of food items preceding health condition
all instances	55.42
instances (manually) labeled as *CAUSE*	78.84

Table 19 List of the best subset of task-specific features for each individual class.

Class	Features
SUIT	synoHlthTS$^\diamond$, againstCond$^\diamond$, negFood$^\diamond$, causeTS$^\diamond$, irrealis$^\diamond$, hlthTS$^\bullet$, negCond$^\diamond$
BENEF	synoHlthTS$^\diamond$, againstCond$^\diamond$, synoHlthEC$^\diamond$, sPosCondSynt*, negFood$^\diamond$, diseaseEC$^\diamond$, boundary$^\diamond$, irrealis$^\diamond$, causeTS$^\diamond$, defHlthTS$^\bullet$
UNSUIT	negFood$^\diamond$, againstCond$^\diamond$, causeEC$^\diamond$, synoUnhEC$^\diamond$, prom$^\diamond$
CAUSE	causeTS$^\diamond$, foodBefCond$^\diamond$, causeEC$^\diamond$, negTS*, synoHlthEC$^\diamond$, boundary$^\diamond$, otherFood$^\diamond$, irrealis$^\diamond$, againstCond$^\diamond$, targetDefUnh$^\bullet$

$^\diamond$: linguistic features (Table 5); *: sentiment features (Table 7); $^\bullet$: features derived from healthiness lexicon (Table 9)

4.2 Impact of Generic Linguistic Features

In this section, we examine the generic linguistic features. In the previous section, we found that the task-specific linguistic features systematically have a positive impact on classification performance. We want to know whether this contribution can be equally achieved by more generic (and thus simpler) features and/or whether the generic features can be usefully combined with the task-specific features. As a baseline feature set we use the word-based features to which the other features are added. Table 20 shows the result of this comparison.

The table shows that the syntactic relation path from the target food item to the target health condition, i.e. *path*, is less helpful than the part-of-speech features, i.e. *pos*. For three out of the four relation types (i.e. *SUIT*, *UNSUIT* and *CAUSE*), *pos* increases performance when added to *word*. *path* only manages improvements for the

Table 20 Performance of generic linguistic features (Table 6).

Features	SUIT			BENEF			UNSUIT			CAUSE		
	Prec	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**
word	31.6	52.6	39.5	59.1	59.8	59.5	37.4	31.7	34.3	52.8	53.9	53.4
word+path	32.3	53.6	40.4$^\bullet$	59.8	59.4	59.6	36.2	33.3	34.7	56.8	54.5	55.6°
word+pos	37.1	49.3	42.3$^\bullet$	57.0	60.5	58.7	38.2	39.2	38.7$^\bullet$	56.3	59.2	57.7°
word+path+pos	37.0	44.3	40.3	**66.1**	52.8	58.7	37.0	37.5	37.3$^\bullet$	58.1	59.2	58.7$^\bullet$
word+ling	37.3	**56.6**	**45.0**$^\bullet$	53.7	**70.0**	60.8	**49.8**	41.3	**45.1**$^\bullet$	55.6	63.1	59.1°
word+pos+ling	**38.6**	53.6	44.9$^\bullet$	58.0	66.7	**62.0**$^\bullet$	45.6	**42.9**	44.2$^\bullet$	**59.8**	**68.2**	**63.7**$^{\bullet\ddagger}$

significantly better than *word* at $^\circ$: $p < 0.1$, $^\bullet$: $p < 0.05$; significantly better than *word+ling* at †: $p < 0.1$, ‡: $p < 0.05$ (based on paired t-test)

relation types *SUIT* and *CAUSE*. A combination of the two features does not result in a systematic improvement. The word-based features combined with the part-of-speech information are not as good as the word-based features combined with the task-specific linguistic features (*ling*). This means that the task-specific features cannot be replaced by the generic features. However, the combination of these features, i.e. *pos* and *ling*, results in some further improvement for the relation types *BENEF* and *CAUSE*.

4.3 Impact of Food and Health Condition Priors

In this section, we investigate the impact of the food and health condition priors (Table 12). Each of the priors is evaluated as a stand-alone feature and in combination with the best subset of features we obtained in our previous experiments, i.e. *word+pos+ling*. Table 21 displays the results.

The table shows that the priors themselves often largely outperform the standard baseline *coocc*. The most effective prior uses the knowledge about the target health condition, i.e. *condPrior*. However, the performance of the feature set *word+pos+ling* is significantly better in three out of four cases (*UNSUIT* is the only exception). This means that beyond this prior knowledge there is much more information that can be learnt from labeled textual data. As far as *UNSUIT* is concerned, there is a high fluctuation among the different folds (in cross-validation) so that the overall improvement of the feature set *word+pos+ling* is not statistically significant. Nevertheless, the classifiers based on that feature set have (overall) a much higher precision than the classifiers based on priors (which can be preferable for certain practical applications). The fact that we also gain a notable performance increase by combining the prior features with the feature set *word+pos+ling* for the relation types *SUIT* and *CAUSE* means that these two types of information are complementary to a certain extent.

4.4 Impact of Manually Chosen Keywords

Table 22 compares the best (overall) feature set from our previous experiments, i.e. *word+pos+ling+condPrior+foodPrior* from Table 21 (we denote it as *no KW* in Ta-

Table 21 Performance of prior features (Table 12).

Features	SUIT			BENEF			UNSUIT			CAUSE		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
cooc	16.7	**100.0**	28.6	19.1	**100.0**	32.1	9.5	**100.0**	17.3	18.4	**100.0**	31.1
condPrior	36.4	23.2	28.4	33.6	54.6	41.6	32.5	61.7	42.6	39.3	82.8	53.3
foodPrior	30.4	27.7	29.0	41.3	38.4	39.8	28.3	26.7	27.7	36.9	27.3	31.4
word+pos+ling	38.6	53.6	44.9•	58.0	66.7	62.0•	45.6	42.9	44.2	**59.8**	68.2	63.7°
word+pos+ling+condPrior	37.9	58.1	45.8•	58.5	65.8	61.9•	38.8	55.0	45.5	57.0	81.1	67.0•
word+pos+ling+foodPrior	36.6	59.2	45.3•	**66.2**	59.2	62.5•	**51.1**	40.4	45.1	55.9	70.6	62.4
word+pos+ling+condPrior+foodPrior	**39.7**	61.9	**48.4**•‡	57.4	71.8	**63.8**•	44.6	54.6	**49.1**	57.4	82.2	**67.6**•†

significantly better than both *condPrior* and *foodPrior* at °: $p < 0.1$, •: $p < 0.05$; significantly better than *word+ling+pos* at †: $p < 0.1$, ‡: $p < 0.05$ (based on paired t-test)

Table 22 Performance of features derived from manually extracted keywords (Table 13).

	SUIT			BENEF			UNSUIT			CAUSE		
Features	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**
no KW	39.7	61.9	48.4	57.4	**71.8**	63.8	44.6	54.6	49.1	57.4	**82.2**	67.6
KW	41.0	**64.7**	50.2†	66.1	68.3	67.2†	**53.5**	54.6	**54.0**	65.6	78.5	**71.5**‡
KW+SyntKW	**43.0**	61.4	**50.6**†	**66.2**	68.5	**67.3**†	45.8	**58.3**	51.3	**66.4**	77.0	71.3‡

significantly better than *no KW* at †: $p < 0.1$, ‡: $p < 0.05$ (based on paired t-test)

ble 22) with manually extracted keywords. Even though we do not address the automatic extraction of instances of the relation types *PREVENT* and *WORSEN*, we included keyword features for these relation types (Table 13). Thus, the classifier may use them as negative features for the detection of the four remaining relation types (i.e. *SUIT*, *BENEF*, *UNSUIT* and *CAUSE*).

Overall, the inclusion of such keywords results in some increase in F-score. Adding syntactic information (*SyntKW*) has no notable impact on performance. Given the contribution of other features we explored in our previous experiments and considering that all features based on manually extracted keywords may also be biased towards our dataset, their impact is, in general, comparatively moderate. Thus, we conclude that extracting keywords for the different relation types is hardly one of the most promising directions for further improvement.

5 Error Analysis

In this section, we present a detailed error analysis. It mostly focuses on explaining why certain features did not show the expected positive effect in our experiments. We address healthiness features (Section 5.1), linguistic features (Section 5.2) and sentiment features (Section 5.3). We also try to uncover why bag of words performs so well (Section 5.4) and also compare it with syntactic features (Section 5.5). Regarding the different relation types, we have a closer look at *SUIT* and *UNSUIT* (Section 5.6), which systematically scored lower than the remaining relation types. Since we use a dataset originating from user-generated web documents, we also need to address the general text quality of our corpus (Section 5.7). Finally, in Section 5.8, we merge the four fine-grained classes to two coarse-grained classes in order to check whether low classification scores are the result of an inventory of classes that are not sufficiently well-defined.

5.1 Healthiness Features

In our evaluation, none of the features derived from the healthiness lexicon (Section 3.5) helped to produce better results than the bag of words and the task-specific features (Section 3.2), i.e. *word* and *ling* in Table 16. Basically, this result should not be considered as an issue for error analysis. It just means that *prior* healthiness as encoded in the healthiness features (Table 9) does not sufficiently correlate with *conditional* healthiness. This should not come as a surprise if we recall the results of

Table 8 displaying the food items strongly correlating with conditional healthiness. It was found that many food items generally considered healthy (e.g. *Pilze (mushrooms)* or *Spargel (asparagus)*) are often considered problematic for certain types of health conditions – so they are *conditional unhealthy*. As a matter of fact this insight should be regarded as some further motivation to consider conditional healthiness as a separate task which requires a different kind of feature engineering than prior healthiness. (Otherwise, if there were a very strong correlation between conditional and prior healthiness, we could solve the current task by *re-using* the methods from solving prior healthiness classification.)

5.2 Linguistic Features

In our previous experiments, we could establish that, as a whole, our set of task-specific features both improves a set of standard features (i.e. *word+path+pos* in Table 20) and that some individual features are especially predictive (e.g. *againstCond* for *BENEF*, *negFood* for *UNSUIT* or *foodBefCon* for *CAUSE* (Table 17)). However, among the set of those task-specific features, there are also features that neither appear to be predictive (Table 17) nor provide complementary information to the remaining features (Table 19). The features that fall within this category are *question*, *side*, *scope* and *weird*.

Table 23 lists the frequency of those less effective features in our dataset. For comparison, we also list the frequency of some effective features. If features are too rare, they are very unlikely to improve classification performance. Table 23 shows, however, that the less effective features are not less frequent than the effective features. From that we conclude that the sparsity of those features is not the reason for them to perform poorly.

Table 24 shows which other features (i.e. effective features) strongly correlate with those poor performing features. If there are strongly correlating features, this explains why these four less effective features were not listed on Table 19 (that only lists features that provide complementary information). Furthermore, if there are features overlapping with those less effective features, the correlating features should then be considered a better alternative. The fewer overlapping features there are, the more unique the information is of a particular feature. We find that the features *negFoodSynt*, *boundary* and *sNegFoodSynt* are the three features most strongly correlating with *scope*, having a Chi-Square score of 98.9, 85.4 and 60.9, respectively. What all these features have in common is that they *syntactically* restrict the mention of the food item in a sentence: It comes as no surprise that *boundary* and *scope* correlate, as we introduced the former as a back-off feature of the latter (Table 5).[15] The other two features, i.e. *negFoodSynt* and *sNegFoodSynt*, do not restrict target food item and health condition to be in the same clause (as *scope* does) but demand target food item and negative polar expression to be syntactically related. The feature *scope* can be considered a pre-requisite of the features *negFoodSynt* and *sNegFoodSynt*, so there is

---

[15] In Section 3.2 we already speculated that the performance of *scope* may be affected by a bad parse quality due to the noise contained in our language data. Obviously, the noise really affects syntactic processing and some of the features that depend on that information, such as *scope*.

Table 23 Frequency of effective/less effective linguistic features.

Feature	Frequency	Percentage of Corpus
Less Effective Features		
scope	819	32.3
weird	280	11.1
side	222	8.8
question	180	7.1
Effective Features		
negFood	384	15.2
synoHlthTS	292	11.5
causeTS	183	7.2
againstCond	180	7.1

Table 24 Feature overlap of less effective features with other features (we only list features with a correlation of Chi-Square 50 or higher).

Feature	Correlated Features	Chi-Square Score
question	boundary	579.7
scope	negFoodSynt	98.9
	boundary	85.4
	sNegFoodSynt	60.9
side	*no strongly correlated features*	*N/A*
weird	foodBefCond	113.0

some linguistic connection between these features. The strong correlations between *question* and *boundary* as well as between *weird* and *foodBefCond* may not appear intuitive. However, one has to recall that two statistically correlated features do not have to be conceptually related. For the feature *side*, we could not find any overlapping features. From that we conclude that even though this feature carries some unique information (Table 24), the information that this feature encodes (i.e. that the target food item occurs as a side dish) is just not relevant for the task examined in this article.

5.3 Sentiment Features

Even though the usage of features from sentiment analysis, or more precisely, polarity classification (Section 3.4) seemed very intuitive for this classification task, for none of the different classes to be detected did the usage of those features produce significantly better results than bag of words and the task-specific features (Section 3.2), i.e *word* and *ling* in Table 16.

In the following, we try to find reasons for this behavior. We take a closer look at the relation type *BENEF* (Section 2.3.2).[16] Intuitively, this type is most likely to correlate with sentiment information, since if some food items are supposed to have some beneficial properties, this should be expressed with some (explicit) positive po-

---

[16] We assume that for the other three relation types, i.e. *SUIT*, *UNSUIT* and *CAUSE*, sentiment is not a predictive feature. Causation (as conveyed by *CAUSE*) is quite different from positive or negative sentiment, so it does not come as a surprise that sentiment features are not effective for this relation type. *SUIT* and *UNSUIT* will be discussed in a dedicated section of this error analysis (i.e. Section 5.6).

lar expressions. From the positive relation types we consider (e.g. *SUIT*, *BENEF* and *PREVENT*), *BENEF* is the relation type with the largest proportion of keywords in the target sentence (Table 15). Having a large proportion of keywords is a pre-requisite for a relation type for which sentiment features are effective (sentiment features can be regarded as a subset of keyword features).

We manually annotated all instances labeled as *BENEF* for which no positive polar expression matched according to our sentiment lexicon. This amounts to 42% of the utterances. Still, all those utterances express something positive. We distinguish between the following reasons why the sentiment lexicon could not establish this information:

(I) There is actually a positive polar expression but it is missing from the sentiment lexicon.

(II) There is actually a positive polar expression but there has been a spelling error or an error in processing (e.g. part-of-speech tagging) so that the positive polar expression could not be matched.[17]

(III) The relation type *BENEF* is inferred and there is no explicit sentiment in the utterance.

(IV) Some form of negation is involved.

Sentiment analysis is far from being solved. This entails that there does not exist any *exhaustive* sentiment lexicon. Therefore, it is only natural that errors caused by (I) occur.

There are similar reasons for (II). Tools in natural language processing are known to be error-prone. On user-generated content (as our dataset), errors may not only be caused by the brittleness of NLP software but also by spelling/grammar mistakes in the actual language to be analyzed.

Sentiment does not have to be explicitly expressed. That is, there are utterances that convey some sentiment even though no explicit polar expression is employed (III). Sentence (25), for example, expresses some positive sentiment towards *Salbeitee (salvia tea)*. One can only infer that sentiment with the help of world knowledge. We must know that reducing the level of blood sugar is the goal of a successful treatment of diabetes. Literally speaking, the reduction of the level of blood sugar just denotes a process that can take place within the body of a living being. Therefore, phrases like this are unlikely to be contained in a sentiment lexicon.

(25) *Salbeitee* senkt bei <u>Diabetes</u> den Zuckerspiegel.
    (*Salvia tea* reduces the level of blood sugar in case of <u>diabetes</u>.)

Finally, we need to address negation (IV). Typically, negation is known to reverse polarity, for instance, a negated negative polar expression can be interpreted as a positive polar expression (as in [**not** *bad*$^-$]$^+$). This form of negation, however, is not present in our data as far as the relation type *BENEF* is concerned. The first difference between the common form of negation and the one that can be found in sentences labeled with our target relation types is that we do not encounter ordinary

---

[17] In order to match a polar expression, not only the word token listed in the sentiment lexicon has to match but also its part-of-speech tag.

Table 25 Statistics of utterances labeled as *BENEF* where no positive polar expression matched according to the sentiment lexicon.

Category	Percentage
againstCond	32.2
negation/shifter (except *against*) negating health condition *(proportion of shifters: 90%)*	17.6
unigram positive polar expression missing from sentiment lexicon	17.6
no positive polar expression at all	16.6
no positive polar expression in target sentence but there is some positive polar expression in the remaining context	11.2
multiword positive polar expression (missing from sentiment lexicon)	2.9
spelling error/error in processing	2.0

polar expressions involved in these constructions, but health conditions. Health conditions, if they represent some illness (e.g. *headache* or *cancer*), can also be considered as a negative polar expression. These expressions are not systematically listed in our sentiment lexicon (approximately 25% of them are missing in our German lexicon). Given that our list of health conditions is known in advance, however, this problem could be easily solved by adding all of those health conditions to our sentiment lexicon. What makes this form of negation really difficult is that in almost 90% of the cases, negation is not conveyed by typical negation words (e.g. *kein (no)*, *nicht (not)*). Instead, so-called *shifters* (Wilson et al, 2005), i.e. (predominantly) verbs, nouns and adjectives whose meaning is similar to negation expressions (e.g. *bekämpfen (combat)*, *lindern (abate)*), are used.[18] Unfortunately, the means to form negations with shifters are pretty diverse, since there is a large set of shifters (26)-(29).[19] Due to the fact that there do not exist any robust lexicons of shifters[20], we could not properly model this phenomenon with one exception. While most individual shifters occurred only very rarely, there was one shifter *against* which occurred very often (29), which lead us to the task-specific feature *againstCond* (Table 5).

(26) Nach zwei Tassen *Tee* sind meine <u>Halsschmerzen</u> schon fast **weg**$_{shifter}$.
     (After two cups of *tea*, my <u>sore throat</u> is almost **gone**$_{shifter}$.)

(27) Der <u>Durchfall</u> hat sich **eingestellt**$_{shifter}$, als ich ihm *Karotten* gab.
     (His <u>diarrhoea</u> **subsided**$_{shifter}$ after I gave him *carrots*.)

(28) Die furchtbaren <u>Kopfschmerzen</u> werde ich jetzt mal mit *Kaffee* **bekämpfen**$_{shifter}$.
     (I will now **combat**$_{shifter}$ my dreadful <u>headache</u> with *coffee*.)

(29) **Gegen**$_{shifter}$ den <u>Durchfall</u> kannst Du *Bananen* essen.
     (You may eat *bananas* **against**$_{shifter}$ your <u>diarrhoea</u>.)

Table 25 displays the distribution of the different cases in which no positive polar expression according to the sentiment lexicon matched. By far the most frequent case is the shifter based on *against*, i.e. *againstCond*. This is also consistent with our evaluation of task-specific features in Table 17 where *againstCond* is the second strongest feature. Other frequent cases in Table 25 are negations other than *againstCond* and

---

[18] Please note that the situation is different in the case of *UNSUIT*, where common negation words predominate (Section 5.6, Table 28).

[19] More than 60% of the shifters occur as singletons on the sentences labeled as *BENEF*.

[20] The shifter lexicon from Wilson et al (2005) just contains 67 shifters which, when translated to German, would have only a marginal impact on our dataset.

unigram polar expressions missing from our lexicon. These polar expressions form a subset of manually chosen keywords (Section 3.7) and, indeed in Table 22, we saw a significant improvement on *BENEF* caused by adding all manually chosen keywords. The remaining cases are fairly difficult to cope with, even if we had better lexical resources. This also concerns the case when only a polar expression in the context is present (but not the target sentence). If a polar expression appears in one of the two sentences preceding or following the target sentence, it is pretty difficult to decide whether this mention refers to the relation between food item and health condition expressed in the target sentence. So far, there do not exist reliable means in NLP to establish intersentential relationships. Multiword expressions could also be covered by a better sentiment lexicon, however, with a proportion of only 2.9%, it would be unlikely to significantly increase classification performance.

While Table 25 displays all cases in which no (positive) polar expression from our sentiment lexicon matched, for 58% of the sentences labeled as *BENEF* in our gold standard, there is a match. Yet, using polar expressions does not significantly improve classifiers trained on bag of words (Table 16). We will examine the role of bag-of-words features in the next section and try to find a reason why it is so difficult to beat that baseline on our dataset.

5.4 Bag of Words

Our previous evaluation (Tables 16 and 20) established that bag of words poses a strong baseline. On the other hand, several features that are based on a list of cue words (e.g. *polarCont* from Table 7) did not have the expected impact on classification (at least not when added to bag of words). The reason for this may be that bag of words captures a considerable amount of the information that is contained in such word-list features. This is particularly true if there are some frequent (discriminatory) cue words. By visual inspection we found that for some relation types there are some recurring language patterns, e.g. for *BENEF*, *X **hilft** gegen Y (X **helps** against Y)* or for *CAUSE*, *von X **bekomme** ich Y (I **get** Y from X)*. Sentences (30)-(36) illustrate this for the relation type *BENEF*. Inevitably, with bag of words, a classifier will pick up correlations, for instance, a correlation between (the positive polar expression) *hilft* and *BENEF* or between *bekomme* and *CAUSE*. Moreover, we also assume that the type of domain we analyze may additionally support the effectiveness of such features. Users that post entries in food-related forums rarely pay much attention to *stylistic diversity*. In other words, it does not matter greatly to them if certain formulations are constantly repeated. This situation is certainly different in other text types, such as news editorials or movie reviews[21] where the authors are also more inclined to appeal to the readers with their language and, therefore, make use of a more diversified vocabulary.

(30) *Natron **hilft** gegen* <u>Sodbrennen</u>.
    (*Baking soda **helps** against* <u>heartburn</u>.)

---

[21] We mention these text types since some of our word-list features, such as polar expressions, have successfully been applied to such domains and are known to improve bag-of-words baselines (Wilson et al, 2005; Ng et al, 2006).

(31) Bei Übelkeit und Erbrechen **hilft** *Ingwer.*
        (*Ginger* **helps** in case of nausea and vomiting.)
(32) *Honig* **hilft** sehr gut bei Husten.
        (*Honey* **helps** very much in case of cough.)
(33) *Holundersaft* ist gesund und **hilft** gegen Fieber.
        (*Elderberry juice* is healthy and **helps** against fever.)
(34) *Kaffee* **hilft** mir auch oft bei Kopfschmerzen.
        (*Coffee* also often **helps** me when I got a headache.)
(35) Mich hat die Grippe erwischt, da **hilft** nur *Hühnersuppe.*
        (The flu got me, only a *chicken broth* will **help**.)
(36) ... und auch bei Bauchweh und Durchfall **hilft** *Cola.*
        (... and even in case of stomach ache and diarrhoea, *coke* will **help** you.)

In order to substantiate our intuition regarding the relationship of our task-specific features and bag of words, we extracted for the strongest task-specific features (throughout the different classes) the words (unigrams) most strongly correlated with them according to Chi-Square statistics. The results are displayed in Table 26. Indeed, there are some very frequently occurring unigrams, such as *helfen (help)*, *gut (good)* or *verursachen (cause)*. Our cue-word lists, such as our sentiment lexicon or *cause*-cues, become less important in the light of such expressions. Apparently, there is really little lexical diversity on our dataset, which means that bag-of-words classifiers can easily pick up the importance of particular words towards the classes to be predicted.

The table also shows that different features (e.g. *synoHlthTS* or *againstCond*) overlap with the same words, i.e. *helfen (help)* or *gegen (against)*. This is an indication that even though the features are meant to model different things, to some extent they capture the same information.[22] However, one should not conclude from Table 26 that features, such as *synoHlthTS*, *negFood* or *posCondSynt*, are identical. In fact, we have proven the opposite in Table 19 by computing the most predictive subset of complementary features using a best-forward subset selection. (For several classes, these different features are contained in the resulting best-forward subset, which means that they are, at least to some extent, complementary.)

Several of our features also express a conjunction of different properties. For instance, *negFood* indicates a mention of a food item that is also negated. Similarly, *posCondSynt* conveys the occurrence of a positive polar expression that is also syntactically related to some health condition. However, these kinds of features still preserve a very high correlation towards particular words. From that we conclude that to a large extent, the information contained in those features mainly stems from the expressiveness of individual (high-frequency) words rather than the co-occurrence with some additional linguistic property. For example, as far as *negFood* is concerned, the mere presence of simple negation expressions within a sentence may already reveal a substantial amount of information as to which type of relation between the health condition and the target food item holds.

---

[22] The counterintuitive result that *helfen (help)* appears on list of *againstCond* and *gegen (against)* appears on the list of *synoHlthTS* can be explained by the fact that the two words basically form a collocation *helfen gegen (help against)* (see also Sentences (30) and (33)).

Table 26 Words that highly correlate with the strongest task-specific features (the numbers in square brackets in the column *Words* denote the Chi-Square score of each word towards the respective feature and the frequency of co-occurrence with that feature; we only list features with a Chi-Square score of 100 or higher from Table 17; we only list words with a correlation towards the task-specific feature of Chi-Square 50 or higher; in case of task-specific features that only marginally differ in scope, e.g. *synoHlthTS and synoHlthEC*, we only list the stronger feature in this table).

Feature	Words
synoHlthTS$^\diamond$	helfen (help) [1380.5, 169]; gegen (against) [157.7, 85]; gesund (healthy) [84.0, 12]
againstCond$^\diamond$	gegen (against) [1928.4.7, 180]; helfen (help) [149.8, 52]
negFood$^\diamond$	nicht (not) [273.6, 111]; kein (no) [273.3, 102]; ohne (without) [158.9, 42]; verzichten (do without) [73.1, 20]
sPosCondSynt*	helfen (help) [664.7, 128]; gut (good) [190.1, 74]; gegen (against) [184.6, 95]
posCondSynt*	helfen (help) [453.3, 131]; gegen (against) [143.8, 106]; gut (good) [136.3, 79]
causeTS$^\diamond$	verursachen (cause) [667.7, 52]; führen (lead to) [336.6, 27]; auslösen (trigger) [141.2, 12]; zu (to) [77.6, 49]; können (can) [67.1, 35]; auftreten (occur) [64.0, 6]
foodBefCond$^\diamond$	*no strongly correlated words present*

$^\diamond$: linguistic features (Table 5); *: sentiment features (Table 7)

In Section 4.1.2, we already pointed out the strength of the feature *foodBefCond* compared to other task-specific features. The analysis displayed in Table 26 additionally confirms that the information encoded cannot be expressed by a subset of words, as no word strongly correlates with this feature. Therefore, *foodBefCond* is a feature that is pretty much unique in the information that it encodes.

5.5 Bag of Words vs. Syntactic Features

In our evaluation, we employed two variations for some features: one plain feature and the other combined with some syntactic restriction. For instance, for the features derived from the sentiment lexicon (Table 7), we included the plain features (*polarCont*) and the features where food item and polar expression also have to be syntactically related (*polarSynt*). Similarly, we divided the set of features using the keyword lexicon (Table 13) into the plain keyword features (*KW*) and the features for which the keyword has to be either syntactically related to the target food item or the health condition (*SyntKW*).

The effectiveness of these syntactically restrained features is somewhat mixed (Tables 17 and 22). In the case of the sentiment features (Table 17), there is a general limitation that sentiment expressions only strongly correlate with one relation type, i.e. *BENEF*. However, for that particular relation type, the strongest syntactic features are much more effective than the strongest plain features (i.e. *sPosCondSynt/posCondSynt* with Chi-Square scores of more than 200 vs. *sPosTS/sNegTS* with Chi-Square scores of just above 90). For the keyword features, we could not find any evidence that the syntactic features significantly improve the plain features (Table 22).

As far as the syntactic keyword features (*SyntKW*) are concerned, we consider it unlikely that the lacking effectiveness is due to flaws in the feature extraction. This is since the underlying resource (the Stanford parser for German) for establishing whether two words are syntactically related is also employed for the syntactic senti-

ment features, which have been found effective (e.g. *sPosCondSynt* or *posCondSynt* in Table 17).

Complementary to the analysis of the keyword and sentiment features, we also examined *generic linguistic features* (Table 6) which exclusively encode syntactic knowledge. However we distinguish between features incorporating part-of-speech information and features encoding information from a syntactic parse tree. The evaluation of those features (Table 20) revealed that part-of-speech information is more effective than information derived from the syntactic parse tree. This result mirrors the only mild effectiveness of the syntactic keyword/sentiment features. Those features, too, incorporate information from a syntactic parse tree rather than part-of-speech information.

In general, syntactic features do not always have to outperform shallow/bag-of-words features. A detailed study for text classification (on the document level) examining various datasets was presented by Moschitti and Basili (2006). Since we obtain similar results on German data as Moschitti and Basili (2006) do on English data (i.e. their syntactic features do not help on that task), we conclude that our results indicating minor effectiveness of many syntactic features cannot be (exclusively) ascribed to the fact that we examine German language data (German NLP-tools are known to be less robust than their English counterparts).

A further noteworthy recent example for lacking effectiveness of syntactic features is the NIST benchmark on slot-filling for knowledge-based population from 2013, in which the top-scoring system (Roth et al, 2014) did not incorporate syntactic features while many of the other participating systems did. (Roth et al (2014) also carry out ablation studies with syntactic features on their system further supporting that, on the task of slot-filling for knowledge-based population, such features are less helpful.)

5.6 The Difficult Relation Types: *SUIT* and *UNSUIT*

Throughout our experiments, we found that the two classes *SUIT* and *UNSUIT* score consistently lower than the other two classes *BENEF* and *CAUSE*. We found the following reason for that:

For both *SUIT* and *UNSUIT*, there are a lot fewer predictive lexical cues than for *BENEF* and *CAUSE*. Three individual results of our experiments support that claim. Firstly, for *SUIT* and *UNSUIT*, the proportion of keywords manually annotated (Table 15) is much smaller, i.e. 21.73 and 32.39, than for the other two classes, i.e. 56.97 and 72.20. Secondly, among the set of effective linguistic features (Table 17), only for *BENEF* and *CAUSE* were there lists of cue words, i.e. *synoHlthTS*, *synoHlthEC*, *sPosCondSynt* and *posCondSynt* for *BENEF*, and *causeTS* and *causeEC* for *CAUSE*. Thirdly, the bag-of-words feature set[23] (Table 16) produces much higher F-scores for *BENEF* (59.4) and *CAUSE* (53.4) than for *SUIT* (39.5) and *UNSUIT* (34.3).

For most of those cases labeled as either *SUIT* or *UNSUIT* where no keyword is present, we did not find any obvious and recurring linguistic pattern that one could

---

[23] We consider our unigram bag-of-words features as a typical example of lexical information.

use. For most of these sentences, one has to *infer* the pertaining relation. In Sentence (37), for example, it is obvious to the reader that the speaker's grandmother drinks that much *mineral water* since her doctor told her to do so because of her diabetes. However, this is extremely difficult to recognize automatically. Likewise, in Sentence (38), one must have the knowledge that acidic food should be avoided if one suffers from heartburn.[24] The speaker of the utterance recommends adding sugar to tomato sauce as a means to neutralize the acidity of *tomatoes*. From that one must infer that if this special trick was not applied, tomatoes would contain too much acid. From that we conclude that, plain tomatoes are unsuitable for people suffering from heartburn.

(37) Meine Oma hat nun <u>Diabetes</u> und ihr hängt das *Mineralwasser* schon zum Hals raus.
(My grandma suffers from <u>diabetes</u> and is completely fed up with drinking *mineral water*.) *LABEL: SUIT*

(38) Noch ein Tipp für Leute mit <u>Sodbrennen</u>: etwas Zucker in die [Tomaten-]Sauce geben, das nimmt die Säure der *Tomaten*.
(Here is a tip for people with <u>heartburn</u>: add some sugar to the [tomato] sauce, it will reduce the acid of the *tomatoes*.) *LABEL: UNSUIT*

To further substantiate that the classes *SUIT* and *UNSUIT* are difficult and that the features we previously examined are, in principle, correctly implemented, but just do not sufficiently correlate with those classes, we manually annotated the sentences labeled as either *SUIT* or *UNSUIT*. We assign a label indicating the most predominant cue that indicates to a human annotator that the particular class label is present. In this annotation, we annotate linguistic categories rather than just the presence of lexical cues (Table 15).[25] We consider polar expressions (both contained and not contained in our sentiment lexicons), negation (we distinguish between the negation words and further shifters (cp. Sentences (26)-(29) in Section 5.3)), discourse relations indicating causal/conditional relations (anchored by an explicit discourse connective, e.g. *wenn (if)* or *weil (because)*)[26], and lexico-syntactic cues (involving neither negation nor other types of lexical items mentioned before in this list). Finally, there are also sentences in which inferences are necessary (e.g. Sentences (37) or (38)).

The results for *SUIT* are displayed in Table 27 and for *UNSUIT* in Table 28, respectively.[27]

The distribution of the categories is consistent with our previous evaluation. As far as *SUIT* is concerned, half of the sentences require human inferencing. The only other frequent cues are discourse relations with explicit discourse connectives. We

---

[24] Note that we do not consider *acid* as a negative polar expression. Acid is not harmful per se. (For example, our digestive system depends on gastric acid.) This is also reflected by the fact that it is not contained in our sentiment lexicon.

[25] Unlike the lexical cues, some of our categories also consider additional syntactic information.

[26] Example: I eat *almonds* **because** I suffer from <u>dermatitis</u>.

[27] Notice that there is no direct correspondence between the categories from Table 15 stating the proportion of instances with manually annotated keywords and Tables 27 and 28. Table 15 only considers unigram keywords being either nouns, verbs or adjectives while the annotation in Tables 27 and 28 is unrestricted. In other words, keywords annotated in Table 15 are *not* the sum of categories excluding *inferences* in Tables 27 and 28. There will also be other constructions marked in those tables that were not captured by the restricted annotation from Table 15.

Table 27 Manual annotation of predictive cues among the sentences labeled as *SUIT*.

Cue	Percentage
inferences	49.5
discourse relations (with explicit discourse cue, e.g. *weil (because)*)	21.0
lexico-syntactic cues	15.7
polar expression (contained in sentiment lexicon)	6.1
polar expression (not contained in sentiment lexicon)	4.4
uncovering relation requires consideration of further context (i.e. preceding/following sentence(s))	2.3
negation word (with complex construction involving health condition/food item)	1.0
negated polar expression	0.0
food item negated by negation word	0.0
food item negated by shifter	0.0
shifter (with complex construction involving health condition/food item)	0.0

Table 28 Manual annotation of predictive cues among the sentences labeled as *UNSUIT*.

Cue	Percentage
food item negated by negation word	42.1
inferences	18.2
food item negated by shifter	11.3
negation word (with complex construction involving health condition/food item)	8.9
negated polar expression	6.5
polar expression (contained in sentiment lexicon)	4.9
shifter (with complex construction involving health condition/food item)	2.4
discourse relations (with explicit discourse cue, e.g. *weil (because)*)	2.0
polar expression (not contained in sentiment lexicon)	1.6
uncovering relation requires consideration of further context (i.e. preceding/following sentence(s))	1.2
lexico-syntactic cues	0.0

subsequently tried to automatically detect these relations, however, we failed to improve performance. One reason for this is that German discourse connectives are ambiguous (e.g. *da (because/there)*).

As far as *UNSUIT* is concerned, we see further evidence that negation is, by far, the most important linguistic phenomenon to be addressed. Furthermore, it is the case that the target food item being negated is the most relevant form of negation. The table also shows that we miss some forms of negation, in particular, those involving shifters, i.e. many words not contained in our negation lexicon. There is still a significant amount of cases which require inference, but it is considerably lower than for *SUIT*. Unlike *SUIT*, other forms of lexical cues not involving polar expressions or negation do not seem to play a role for *UNSUIT*.

A comparison of Table 27 and Table 28 may suggest that it should be easier to detect the relation type *UNSUIT* than the relation type *SUIT*. Still, in our evaluation we obtained similar F-scores for both classes (Tables 16, 20, 21 and 22). The reason for this may lie in the class distribution (Table 2). From a feature perspective, *UNSUIT* may be easier to handle, however, this effect is less visible in our evaluation since *UNSUIT* is much rarer (9.49%, Table 2) than *SUIT* (16.44%, Table 2). The latter occurs almost twice as often as the former. Typically, it is more difficult to detect rare classes (i.e. *UNSUIT*) than frequently observed classes (i.e. *SUIT*).

Despite the fact that *SUIT* and *UNSUIT* score much lower in comparison to *BENEF* and *CAUSE*, we assume that further significant improvements could be achieved in the presence of more training data. Figure 2 displays the learning curve using the full feature set (including even keywords), i.e. KW from Table 22. The figure suggests that at least the curves for *SUIT* and *UNSUIT* are far from being saturated (in contrast to *BENEF*).[28] (We assume that the more training data is added, the more information can be learned about less frequent lexical features.) Still, one may doubt that given sufficient training data, one will ever obtain scores similar to those obtained for *BENEF* or *CAUSE*.

In this error analysis we established that a very large amount of instances labeled as *SUIT* require some form of inference (Table 27). These sentences are unlikely to be resolved with current NLP technologies. At this point, we should wonder whether it should be equally important to extract all types of sentences automatically that express our target relations. In Section 1, we briefly sketched as a potential application of our task an intelligent search engine for forum entries. Such a component would typically be used by laymen possessing no knowledge about the algorithms that underlie the search engine. As a consequence, the sentences extracted should be easy to verify. Examples where the target relation is present but must be inferred, such as Sentences (37), may not be perceived as particularly convincing to such users. Explicit alternatives, such as Sentence (39) should be preferable. Such alternatives contain some form of explicit (lexical) cues and are much more likely to be extracted automatically. Therefore, we think it does not matter too greatly that some relation instances (i.e. the implicit ones requiring inference) cannot be extracted automatically.

(39) Bei <u>Diabetes</u> **empfehle** ich *Mineralwasser*.
(For *diabetes*, I **recommend** <u>mineral water</u>.) *LABEL: SUIT*

5.7 Textual Quality of the Corpus

Since our text corpus has been extracted from user-generated content of the web, we need to address its textual quality. Obviously, we cannot expect the level of correctness encountered on newspaper corpora that are typically employed for NLP tasks. In order to measure the amount of noise in our dataset, we randomly sampled 100 sentences from our corpus in which a food item and some health condition co-occurred and manually annotated the types of errors that were encountered.

Table 29 lists the results of this evaluation. With regard to grammar, we distinguished between obvious errors and non-standard grammatical constructions. For instance, we encountered many sentences in which the subject is missing (40). Even though such constructions are considered ungrammatical in formal written language, they are acceptable and commonly found in informal language (including spoken language).

---

[28] It may come as a surprise that the classes *BENEF* and *CAUSE* already produce high scores with only 25% of the training data. This is due to the fact that we consider the full feature set that includes many features based on word lists. Typically, such features are particularly effective if only few training data are present. Such features generalize over individual word occurrences and are less sparse than bag of words.
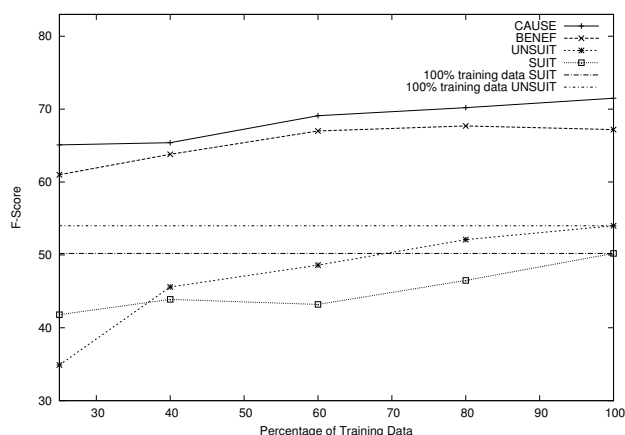
Fig. 2 Learning curve for the different classes using the full feature set including keywords (i.e. KW from Table 22).

(40) Heute bleibe ich bei *Zwieback*; **hatte letzte Nacht furchtbares <u>Sodbrennen</u>**.
     (Today, I stick to *zwieback*; **had a terrible <u>heartburn</u> last night**.)

We also added very informal expressions to our list (including informal abbreviations, such as *vll* for *vielleicht (perhaps)*), despite the fact that they are not errors, strictly speaking. Still, they are a potential error source for part-of-speech tagging or syntactic parsing since they are treated as unknown words.[29]

Table 29 confirms that there is a significant amount of errors present in our dataset. These errors are also likely to affect our automatic extraction procedure.

One notable error source that has been caused by pre-processing our text corpus are errors in sentence boundary detection which occur in 25% of the sentences. We used a standard tool for German, i.e. the German model from *OpenNLP*[30], and are not aware of a more recent/robust tool, so there are no realistic alternatives for this processing step that could reduce those errors significantly.

In general, most of the errors are inherent to this text type. In order to better handle those domain idiosyncrasies, it would be worthwhile to employ a part-of-speech tagger and/or a syntactic parser which has been specially trained on such social media texts containing informal language. However, such an undertaking is beyond the scope of the research presented in this article. Neither is there a publicly available tool of that sort for German.

Despite the present formal errors contained in our text corpus, we still think that our resource is appropriate for this research. The amount of spelling mistakes is actually low. We could hardly find any health conditions/diseases or drugs (which can be quite complex) written incorrectly. Most of the errors are simple typos. The amount

---

[29] The training data for those tools typically originate from the news domain, where such expressions cannot be found.

[30] opennlp.sourceforge.net/projects.html

Table 29 Distribution of error types manually annotated on a random sample of 100 sentences (a sentence may contain several errors at the same time).

Error Type	Percentage
non-standard grammatical construction	26
incorrect sentence boundary detection	24
spelling error	19
punctuation error	16
very informal/non-standard wording	16
tokenization error	12
grammatical error	6

of grammatical errors (excluding non-standard usage) is even lower than the amount of spelling mistakes. Given these observations, our text corpus is unlikely to represent a particularly poor choice of language data. Further, it is very unlikely that there exists a similar alternative textual source from the web with a similar coverage on the issues we want to extract.

5.8 Fine-Grained vs. Coarse-Grained Analysis

In this section, we contrast the fine-grained classification that we examined in the previous sections with a more coarse-grained classification. Rather than addressing four different classes, we could alternatively merge the four classes into two large classes, i.e. one positive class (combining *SUIT* and *BENEF*) and one negative class (combining *UNSUIT* and *CAUSE*). Through this experiment, we want to provide further evidence that our fine-grained class inventory is well-defined. This would not be the case if it were significantly easier to just distinguish between two strongly opposing classes and thus achieve a much higher classification performance (than on the fine-grained set-up). This would indicate that the fine-grained classifiers confuse the different classes too much.

Table 30 compares the performance of the fine-grained and coarse-grained setting. We compare two different classification approaches: *individual* considers fine-grained classes, that is, in the training data there are four classes (to be predicted); in *merged*, there are only two coarse-grained classes (i.e. *SUIT+BENEF* and *UN-SUIT+CAUSE*). In order to have a meaningful comparison, we must produce a common output format for both classification approaches. Therefore, we converted the output of the fine-grained classifier to the output of the coarse-grained classifier. So, even though *individual* originally predicts four different classes, after conversion, there are only two class labels that are predicted (either prediction of *SUIT* and *BENEF* is converted to the coarse-grained class *SUIT+BENEF*, while either prediction of *UNSUIT* or *CAUSE* is converted to the coarse-grained class *UNSUIT+CAUSE*).

If our fine-grained class inventory were inappropriate, that is, we were artificially separating the two coarse-grained classes into two halves each, then we would end up having very similar (perhaps even almost identical) instances in two separate (fine-grained) classes. Such partition of instances would become fairly problematic for a fine-grained classifier because it would be forced to learn inconsistent information (i.e. two similar instances belonging to different classes). The result of this would

Table 30 Performance on the coarse-grained classes (i.e. *SUIT+BENEF* and *UNSUIT+CAUSE*).

Classifier	SUIT+BENEF			UNSUIT+CAUSE		
	Prec	Rec	F1	Prec	Rec	F1
individual (fine-grained)	61.72	75.36	67.86	64.26	71.87	67.82
merged (coarse-grained)	60.60	82.10	69.73•	57.16	78.61	66.19

•: significantly better than *individual* at $p < 0.05$

be that the fine-grained classifier (*individual*) performs much worse than the coarse-grained classifier (*merged*). (The latter should be better since it would be trained on meaningful classes.) If, however, the fine-grained class inventory is appropriate (this means that the different classes would contain sufficiently different instances), the result would be that there is no notable difference between *merged* and *individual*.

Table 30 shows that for the two positive classes, i.e. *SUIT* and *BENEF*, the classifier trained on a merged positive class performs slightly better than the classifier trained on the individual classes. The difference between *individual* and *merged* is even statistically significant. However, for the negative classes there is no such effect. From that we conclude that *SUIT* and *BENEF* are more similar to each other than *UNSUIT* and *CAUSE*. A fine-grained classifier has a tendency to confuse the two positive classes. However, given that there is only a slight decrease in performance by *individual* on *SUIT+BENEF*, we do not think that our fine-grained class inventory is insufficiently well-defined.

6 Related Work

In the health/medical domain, the majority of research focuses on domain-specific relations involving entities, such as genes, proteins and drugs (Cohen and Hersh, 2005). More recently, the prediction of epidemics from social media (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 also been work on processing health-care claims (Popowich, 2005) and detecting sentiment in health-related texts (Sokolova and Bobicev, 2011).

Research in the food domain using natural language processing has so far focused on the following tasks: 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, 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 Wiegand et al (2012b), extraction methods for domain-specific relations in the food domain have been examined. The relations that are dealt with are motivated by customer needs in a supermarket. These relations answer the questions which food items can be consumed together, which food items are typically offered on a particular social event and which food items can be substituted by each other. Unlike this work, where individual utterances are evaluated, Wiegand et al (2012b) exclusively evaluate relation extraction based on an aggregate of multiple utterances. The extraction methods

employed, i.e. lexical surface patterns and statistical co-occurrence, are fairly simple and require little linguistic processing. Wiegand et al (2012a) is an extension of Wiegand et al (2012b) in which those extraction methods are tested on different corpora, thus showing that domain-specific texts are necessary in order to obtain reasonable performance. Moreover, the different extraction methods are combined. Wiegand et al (2014) examine the extraction of individual utterances in a supervised learning scenario and incorporate common food categories (e.g. *fruits*, *meat*, *vegetables* etc.) as features.

Our work also bears some relation to the recent work on sentiment analysis addressing *goodFor/badFor* events, i.e. events that have a positive or negative effect on the entities involved in them (Deng and Wiebe, 2014; Deng et al, 2014). Food items could also be interpreted as such *events*, while the persons with specific health conditions are the entities onto whom the food items have either a positive or negative impact. Rather than looking at explicit cues, as we do, Deng and Wiebe (2014) and Deng et al (2014) largely focus on *opinion implicatures*, i.e. (defeasible) implicit opinions, that are uncovered by automatic inference. There are several reasons, however, why we think that this approach cannot be immediately applied on our task setting. Firstly, the text source examined for *goodFor/badFor* events are editorials or blogs. These texts are longer monothematic discourses where there is much context with explicit sentiment information to infer the relation involving *goodFor/badFor* events. Our contexts are much shorter and largely multi-topic (as a matter of fact, several relations expressing the suitability or unsuitability of food items are expressed as a side note) where such inference mechanism is likely to fail. Secondly, their work relies on a gold annotation of explicit sentiment (which is used to infer the implicatures). In realistic situations that kind of information is not available, and automatic methods to detect it are still fairly error-prone. Moreover, on our task, we found that there is only a strong correlation between explicit sentiment (i.e. polar expressions) and one out of four classes we want to extract (i.e. *BENEF*).

The work that is most closely related to this article is Wiegand and Klakow (2013a) in which the contextual healthiness of food items is extracted from natural language texts. In both this article and Wiegand and Klakow (2013a), some contextual analysis is carried out on sentences that contain (hopefully health-related) mentions of food items. The fundamental difference to this article is that in Wiegand and Klakow (2013a) the *general* or *prior* healthiness of food items is to be determined instead of *conditional* healthiness. In Section 3.5, we explained the difference between these two concepts in detail. It has also a notable impact on the usability of resources. As far as general/prior healthiness is concerned, a healthiness lexicon can be used as a source for prior features. Indeed, Wiegand and Klakow (2013a) confirm a strong correlation between the content of the healthiness lexicon and the information contained in their text corpus. In this article, however, such a resource was considered less useful. Wiegand and Klakow (2013a) and this article also work with different instance spaces: While Wiegand and Klakow (2013a) consider sentences with co-occurrences of the word *healthy* and some food item, in this article, we consider sentences in which food items co-occur with mentions of particular health conditions. (Due to these different instance spaces, different feature sets are also applied in these two different scenarios.) Prior healthiness is a binary classification problem

(i.e. either a food item is healthy or not) while in this article we identified several sub-categories of suitability and unsuitability which we also tried to separate. Wiegand and Klakow (2013a) spend some considerable effort in finding reliable utterances regarding healthiness. As far as conditional healthiness is concerned, we do not filter these cases (e.g. *hedging*), as most relations between food items and health conditions are weak and thus associated with a relatively high degree of uncertainty. Wiegand and Klakow (2013a) also experiment with different types of classifiers, including aggregate-based and rule-based classification, while in this work we focus on feature engineering for supervised learning.

Finally, Wiegand and Klakow (2013b) examine the problem of detecting *reliable* food-health relationships. Their work also carries out experiments on the dataset examined in this article. The focus of that work is that given a food-health relationship, how can one decide automatically whether that information is perceived as reliable or not. Wiegand and Klakow (2013b) already assume the relationships between food items and health conditions as given (they just read off the labels from the gold-standard), while this article focuses on how these labels can actually be automatically detected.

7 Conclusion

We presented a new annotation scheme for conditional healthiness extraction, that is, the extraction of suitable and unsuitable food items regarding specific health conditions. Rather than considering this task as a binary classification problem we also considered subtypes of suitability and unsuitability.

We examined a plethora of features for supervised classification and found that task-specific resources, such as a healthiness lexicon classifying food items according to prior healthiness, do not systematically help for this task. Instead a feature set comprising bag-of-words features, some task-specific linguistic features and part-of-speech information produces a much better classifier. The task-specific linguistic features we include focus on three groups of features, those being configurational cues implying that the target food item and health condition are related to each other, features checking whether the assumed relation is embedded in a context that invalidates the relation (e.g. negation, irrealis etc.) and some look-up lists containing words indicative of concepts that are relevant for this task.

We also found that some individual food items and health conditions correlate with particular relation types. Using this knowledge as stand-alone prior features mostly outperforms a trivial co-occurrence-based baseline and can also be usefully combined with the remaining features for supervised learning.

A further result of our research is that, to some extent, the four different classes we examine display different properties and therefore show different effectiveness with regard to specific features:

SUIT is the most difficult class producing the lowest overall scores. Apart from the features that basically help for all classes, e.g. generic linguistic features, prior features or only manually chosen keywords, we could not determine any particularly outstanding feature. Most surprisingly, polar expressions do not correlate with this

class. The learning curve from that class, however, indicates that, given more labeled training data than are provided by the current dataset, classifiers with better performance could be achieved. This class contains a high number of instances where some inference is required. Prospective applications incorporating the extraction of conditional healthiness may not suffer too greatly from the fact that those *implicit* relation instances cannot be reached, since users may find *explicit* instances more convincing.

BENEF behaves very differently from *SUIT*, even though they both convey suitability. Overall, we achieve much higher scores. This class is easier to handle because there are more explicit keywords to indicate the relation. One notable subset of keywords are polar expressions (in particular if they are syntactically related to the health condition), however, bag of words can already pick up correlations between most polar expressions and this class. There are frequently occurring words, such as *helfen (help)*, that are fairly unambiguous cues for this class.

UNSUIT displays similar properties as *SUIT* in that it also produces fairly low overall scores. Still, there is one feature that correlates fairly well with this class, namely the target food item being negated. Polar expressions, similar to *SUIT*, do not help. However, the learning curve also indicates further improvement given more training data.

Finally, *CAUSE* displays similar properties to *BENEF*. This class already produces high scores with a simple bag-of-words feature set. In addition, re-using a lexicon with cue words proposed for other datasets in previous publications also results in some notable improvement. Furthermore, a simple feature that indicates that the target food item has been mentioned prior to the health condition (in the sentence to be classified) also strongly correlates with that class.

Acknowledgements

References

Beamer B, Girju R (2009) Using a Bigram Event Model to Predict Causal Potential. In: Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics (CICLing), Mexico City, Mexico, pp 430–441

Bunescu RC, Mooney RJ (2005) Subsequence Kernels for Relation Extraction. In: Proceedings of the Conference on Neural Information Processing Systems (NIPS), Vancouver, Canada

Chahuneau V, Gimpel K, Routledge BR, Scherlis L, Smith NA (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), Jeju Island, Korea, pp 1357–1367

Church KW, Hanks P (1990) Word Association Norms, Mutual Information, and Lexicography. Computational Linguistics 16(1):22–29

Cohen AM, Hersh WR (2005) A survey of current work in biomedical text mining. Briefings in Bioinformatics 6:57 – 71

Deng L, Wiebe J (2014) Sentiment Propagation via Implicature Constraints. In: Proceedings of the Conference on European Chapter of the Association for Computational Linguistics (EACL), Gothenburg, Sweden, pp 377–385

Deng L, Wiebe J, Choi Y (2014) Joint Inference and Disambiguation of Implicit Sentiments via Implicature Constraints. In: Proceedings of the International Conference on Computational Linguistics (COLING), Dublin, Ireland, pp 79–88

Diaz-Aviles E, Stewart A, Velasco E, Denecke K, Nejdl W (2012) Epidemic Intelligence for the Crowd, by the Crowd. In: Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM), Dublin, Ireland

Fisichella M, Stewart A, Cuzzocrea A, Denecke K (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), Pisa, Italy, pp 87–103

Gildea D, Jurafsky D (2002) Automatic Labeling of Semantic Roles. Computational Linguistics 28(3):245–288

Girju R (2003) Automatic Detection of Causal Relations for Question Answering. In: Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering, Sapporo, Japan, pp 76–83

Girju R, Moldovan D (2002) Text Mining and Causal Relations. In: Proceedings of the International FLAIRS Conference (FLAIRS), Pensacola Beach, FL, USA, pp 360–364

Girju R, Badulescu A, Moldovan D (2003) Learning Semantic Constraints for the Automatic Discovery of Part-Whole Relations. In: Proceedings of the Human Language Technology Conference (HLT), Edmonton, Canada, pp 80–87

van Hage WR, Katrenko S, Schreiber G (2005) A Method to Combine Linguistic Ontology-Mapping Techniques. In: Proceedings of International Semantic Web Conference (ISWC), Springer, Galway, Ireland, pp 732 – 744

van Hage WR, Kolb H, Schreiber G (2006) A Method for Learning Part-Whole Relations. In: Proceedings of International Semantic Web Conference (ISWC), Springer, Athens, GA, USA, pp 723 – 735

van Hage WR, Sini M, Finch L, Kolb H, Schreiber G (2010) The OAEI food task: an analysis of a thesaurus alignment task. Applied Ontology 5(1):1 – 28

Hamp B, Feldweg H (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, Madrid, Spain, pp 9–15

Joachims T (1999) Making Large-Scale SVM Learning Practical. In: Schölkopf B, Burges C, Smola A (eds) Advances in Kernel Methods - Support Vector Learning, MIT Press, pp 169–184

Kessler JS, Nicolov N (2009) Targeting Sentiment Expressions through Supervised Ranking of Linguistic Configurations. In: Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM), San Jose, CA, USA

Klenner M, Petrakis S, Fahrni A (2009) Robust Compositional Polarity Classification. In: Proceedings of Recent Advances in Natural Language Processing (RANLP), Borovets, Bulgaria, pp 180–184

Kozareva Z (2012) Cause-Effect Relation Learning. In: Proceedings of the TextGraphs Workshop at ACL, Jeju, Republic of Korea, pp 39–43

Landis JR, Koch GG (1977) The Measurement of Observer Agreement for Categorical Data. Biometrics 33(1):159–174

Miller G, Beckwith R, Fellbaum C, Gross D, Miller K (1990) Introduction to WordNet: An On-line Lexical Database. International Journal of Lexicography 3:235–244

Mintz M, Bills S, Snow R, Jurafsky D (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), Singapore, pp 1003–1011

Moschitti A, Basili R (2006) Complex Linguistic Features for Text Classification: A Comprehensive Study. In: Proceedings of the European Conference in Information Retrieval (ECIR), Sunderland, United Kingdom, pp 181–196

Munro R, Gunasekara L, Nevins S, Polepeddi L, Rosen E (2012) Tracking Epidemics with Natural Language Processing and Crowdsourcing. In: Proceedings of the Spring Symposium for Association for the Advancement of Artificial Intelligence (AAAI), Toronto, Canada, pp 52–58

Ng V, Dasgupta S, Arifin SMN (2006) Examining the Role of Linguistic Knowledge Sources in the Automatic Identification and Classification of Reviews. In: Proceedings of the International Conference on Computational Linguistics and Annual Meeting of the Association for Computational Linguistics (COLING/ACL), Sydney, Australia, pp 611–618

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

Rafferty A, Manning CD (2008) Parsing Three German Treebanks: Lexicalized and Unlexicalized Baselines. In: Proceedings of the ACL Workshop on Parsing German (PaGe), Columbus, OH, USA, pp 40–46

Roth B, Barth T, Wiegand M, Singh M, Klakow D (2014) Effective Slot Filling Based on Shallow Distant Supervision Methods. In: Proceedings of the Text Analysis Conference (TAC), NIST, Gaithersburg, MD, USA

Sokolova M, Bobicev V (2011) Sentiments and Opinions in Health-related Web Messages. In: Proceedings of Recent Advances in Natural Language Processing (RANLP), Hissar, Bulgaria, pp 132–139

Torii M, Yin L, Nguyen T, Mazumdar CT, Liu H, Hartley DM, Nelson NP (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

Wiegand M, Klakow D (2010) Convolution Kernels for Opinion Holder Extraction. In: Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL), Los Angeles, CA, USA, pp 795–803

Wiegand M, Klakow D (2013a) Towards Contextual Healthiness Classification of
    Food Items – A Linguistic Approach. In: Proceedings of the International Joint
    Conference on Natural Language Processing (IJCNLP), Nagoya, Japan, pp 19–27

Wiegand M, Klakow D (2013b) Towards the Detection of Reliable Food-Health Re-
    lationships. In: Proceedings of the NAACL-Workshop on Language Analysis in
    Social Media, Atlanta, GA, USA, pp 69–79

Wiegand M, Roth B, Klakow D (2012a) Data-driven Knowledge Extraction for the
    Food Domain. In: Proceedings of KONVENS, Vienna, Austria, pp 21–29

Wiegand M, Roth B, Klakow D (2012b) Web-based Relation Extraction for the Food
    Domain. In: Proceedings of the International Conference on Applications of Nat-
    ural Language Processing to Information Systems (NLDB), Springer, Groningen,
    the Netherlands, pp 222–227

Wiegand M, Roth B, Lasarcyk E, Köser S, Klakow D (2012c) A Gold Standard for
    Relation Extraction in the Food Domain. In: Proceedings of the Conference on
    Language Resources and Evaluation (LREC), Istanbul, Turkey, pp 507–514

Wiegand M, Roth B, Klakow D (2014) Automatic Food Categorization from Large
    Unlabeled Corpora and Its Impact on Relation Extraction. In: Proceedings of the
    Conference on European Chapter of the Association for Computational Linguistics
    (EACL), Gothenburg, Sweden, pp 673–682

Wilson T, Wiebe J, Hoffmann P (2005) Recognizing Contextual Polarity in Phrase-
    level Sentiment Analysis. In: Proceedings of the Conference on Human Lan-
    guage Technology and Empirical Methods in Natural Language Processing
    (HLT/EMNLP), Vancouver, BC, Canada, pp 347–354

Witten I, Frank E (2005) Data Mining: Practical Machine Learning Tools and Tech-
    niques. Morgan Kaufmann Publishers, San Francisco, US