A Gold Standard for Relation Extraction in the Food Domain

Michael Wiegand, Benjamin Roth, Eva Lasarcyk, Stephanie Köser, Dietrich Klakow

Spoken Language Systems
Saarland University
D-66123 Saarbrücken, Germany
{Michael.Wiegand|Benjamin.Roth|Dietrich.Klakow}@lsv.uni-saarland.de
{evaly|skoeser}@coli.uni-saarland.de

Abstract

We present a gold standard for semantic relation extraction in the food domain for German. The relation types that we address are motivated by scenarios for which IT applications present a commercial potential, such as *virtual customer advice* in which a virtual agent assists a customer in a supermarket in finding those products that satisfy their needs best. Moreover, we focus on those relation types that can be extracted from natural language text corpora, ideally content from the internet, such as web forums, that are easy to retrieve. A typical relation type that meets these requirements are pairs of food items that are usually consumed together. Such a relation type could be used by a virtual agent to suggest additional products available in a shop that would potentially complement the items a customer has already in their shopping cart. Our gold standard comprises structural data, i.e. relation tables, which encode relation instances. These tables are vital in order to evaluate natural language processing systems that extract those relations.

Keywords: Domain-specific Relation Extraction, Information Extraction, Food Domain

1. Introduction

In this paper, we present a gold standard for semantic relation extraction in the food domain for German. There has only been very little research on natural language processing in the food domain. This is not only true for German but also for all other languages including English. We believe that one reason for this is that there hardly exist any labeled data for this particular domain on which to evaluate automatic systems.

There are several scenarios for commercial applications which would benefit from systems that automatically acquire knowledge about food items, for example, a virtual customer advisor in a supermarket suggesting additional products available in the shop that would potentially complement the items a customer has already in their shopping cart. In another situation the customer would like to purchase an item which is not available. Ideally, the advisor could suggest an appropriate substitute that is available. Customer advice may also take into consideration some specific disposition of a customer, e.g. a customer suffering from diabetes or irritable bowel syndrome may want to be recommended some products that are particularly suitable for their disposition.

Natural language processing and text mining are tools that could be instrumental in acquiring the necessary knowledge for these scenarios. The large amounts of textual data on the web may present an alternative to manually compiled domain knowledge which is typically expensive to produce. But in order to find robust methods for these tasks, some manually labeled data are required for evaluating the output of automated systems. The data that we provide in our new resource are manually compiled relation instances of certain relation types that are relevant to potential applications in the food domain. For example, for the relation type *Can-be-Served-with(FOOD-ITEM, FOOD-ITEM)* a typical entry (i.e. relation instance) could be the tuple <*fish fingers*, *mashed potatoes*>.[1] This means that our resource contains structured data. The reason for not annotating a text corpus is that in comparison to the time-consuming sentence annotation, our method allows to capture much more different relation instances than would be possible to capture by the annotation of sentences (using the same amount of time for annotation). Ultimately, we would like to evaluate systems that populate databases. For such an evaluation, the more different relation instances are available to compare an extraction method against, the more accurately the method can be evaluated. However, we even believe that our gold standard can be used to obtain a labeled text corpus where mentions of instances of a particular relation type are annotated. This conversion is typically achieved by applying some form of *distant supervision* (Mintz et al., 2009), i.e. extracting sentences with mentions of the entities of a particular tuple and assuming that their joint occurrence will express this particular relation. For instance, a joint occurrence of the tuple mentioned above could be the following sentence:

> I usually have *mashed potatoes* with my *fish fingers*.

Such sentences can in turn be used as labeled training data for supervised learning. This is an established learning paradigm that has also been successfully applied in NIST's benchmark task for *knowledge base population* (Ji et al., 2010), a task that is dedicated to relation extraction.

2. Related Work

Most work examining the extraction of semantic relations focus on domain-independent relations, such as hyponyms (Hearst, 1992; Snow et al., 2006),

[1] Despite the fact that our resource exclusively contains German entries, we will just provide English translations in order to ensure general accessibility.

meronyms (Girju et al., 2003), synonyms (Chklovski and Pantel, 2004), general purpose analogy relations (Turney et al., 2003) and general relations involving persons or organizations (Ji et al., 2010).

As far as domain-specific relation extraction is concerned, the domain that has been receiving the greatest attention is most likely the biomedical domain. The types of relations that are examined here involve entities, such as genes, proteins or drugs (Cohen and Hersh, 2005).

There has also been some work on relation extraction in the food domain. The most prominent research addresses ontology or thesaurus alignment (van Hage et al., 2010), a task in which concepts from different sources are related to each other. In this context hyponomy relations (van Hage et al., 2005) and part-whole relations (van Hage et al., 2006) have been explored. Both hyponomy and synonymy relations have also be examined as a means of query expansion for search (Liu and Li, 2011). While in both (van Hage et al., 2005) and (van Hage et al., 2006) the semantic relations are extracted or learned from various types of data, Liu and Li (2011) do not explicitly mention how they obtain the semantic relations involving food items.

Existing knowledge bases, such as GermaNet (Hamp and Feldweg, 1997), may contain some information relevant to the food domain but since they are designed as open-domain ontologies they do not cover such specific relation types as we do in our resource. Even domain-specific ontologies, such as *AGROVOC*[2] or *USDA National Nutrient Database*[3], only cover hyponomy or part-whole relations and nutrient content. Our gold standard contains different relation types. This is mainly due to the fact that the aforementioned resources are primarily designed for experts e.g. in the agricultural sector, while our gold standard has been created for building applications dealing with every-day-life shopping scenarios.

3. Terminology

In the following, we briefly define some terminology which we subsequently use in the remainder of this paper.

3.1. Relation

By a relation we understand an n-tuple consisting of a predicate and its arguments, i.e. *Predicate*(Arg_1, ..., Arg_n). The relations that we consider in this work are typed. This means that the arguments of a relation are restricted to certain entity types, such as FOOD-ITEM, DISH, EVENT or DISPOSITION.

3.2. Instantiated Relation

By instantiated relation, we understand a relation where all its arguments are instantiated, i.e. they denote some specific entity. Examples are *Can-be-Served-with(fish fingers, mashed potatoes)*, *Can-be-Substituted-by(apple, pear)* or *Recommended-for-People-with(chicken broth, flu)*. We may also refer to this type of relation as *relation instance*.

[2] aims.fao.org/website/AGROVOC-Thesaurus/sub

[3] ndb.nal.usda.gov/ndb/foods/list

3.3. Partially Instantiated Relation

By partially instantiated relation, we understand relations where not all arguments are instantiated. Examples are *Can-be-Served-with(baguette, ??)* or *Recommended-for-People-with(??, diarrhoea)*.

For the annotation, partially instantiated relations were given to the annotators. The annotators collect suitable values for unspecified argument slots. The choice of the relations mostly depends on the scenario for which we envisage this resource to be useful (see also Section 1.). Several of our (binary) relations contain argument slots with the same entity types and these relations also happen to be reflexive, e.g. *Can-be-Substituted-by(margarine, butter)* is the same as *Can-be-Substituted-by(butter, margarine)*. For those relations we need not specify which argument slot is specified as this is some redundant information.

4. The Different Relations

In the following, we describe the semantic relations for which we provide annotation. The choice of the relation types was mainly influenced by the scenarios for which we think IT applications would be viable in this domain. One obvious scenario is *virtual customer advice* that has already been outlined in Section 1. Considering this very scenario, the relation types that matter should correspond to the information needs a customer has that cannot be (immediately) satisfied by means that are already provided in a typical supermarket, such as price or expiry date (these information are usually provided by the packaging of the products or contact labels).

4.1. *Suits-to(FOOD-ITEM, EVENT)*

Suits-to describes a relation about food items that are typically consumed at some particular cultural or social event. Examples are <*roast goose*, *Christmas*> or <*popcorn*, *cinema visit*>.

The list of partially instantiated relations given to the annotators always specifies the event. The annotator, therefore, has to look for appropriate food items.

4.2. *Can-be-Served-with(FOOD-ITEM, FOOD-ITEM)*

This relationship describes food items that are typically consumed together. Examples are <*fish fingers*, *mashed potatoes*>, <*baguette*, *ratatouille*> or <*wine*, *cheese*>.

4.3. *Can-be-Substituted-by(FOOD-ITEM, FOOD-ITEM)*

Can-be-Substituted-by lists pairs of food items that are almost identical to each other in that they are commonly consumed or served in the same situations. Examples are <*butter*, *margarine*>, <*anchovies*, *sardines*> or <*Sauvignon Blanc*, *Chardonnay*>.

4.4. *Ingredient-of(FOOD-ITEM, DISH)*

In this relation, the ingredients of diverse dishes are listed. Examples are <*chickpea*, *falafel*> or <*rice*, *paella*>.

The list of partially instantiated relations given to the annotators always specifies a dish. The annotator, therefore, has to look for appropriate ingredients. (We also experimented with the inverse relation but we dropped this as the annotation was less straightforward.)

4.5. *Recommended-for-People-with(FOOD-ITEM, DISPOSITION)*

This relation describes food items that are commonly considered alleviating or (at least) harmless for people having a specific disposition. Examples are <*lemon*, *cold*> or <*red meat*, *iron deficiency*>.
The list of partially instantiated relations given to the annotators always specifies a disposition. The annotator, therefore, has to look for appropriate food items.

4.6. *Not-Recommended-for-People-with(FOOD-ITEM, DISPOSITION)*

This relation describes food items that are commonly considered harmful to people having a particular disposition. Examples are <*alcohol*, *pregnancy*> or <*chocolate*, *diabetes*>.
The list of partially instantiated relations given to the annotators always specifies a disposition. The annotator, therefore, has to look for appropriate food items.
This relation and its counterpart (i.e. *Recommended-for-People-with*) may not only be interesting for information extraction, in general, but also for research in sentiment analysis. This is because detecting whether an item is suitable for a person with a specific disposition or not in some way mirrors the detection of (opinion) polarity towards certain target entities (Kessler and Nicolov, 2009; Jakob and Gurevych, 2010).

4.7. *Healthy and Unhealthy Food Items*

In addition to the relations mentioned above, we will also provide a classification (i.e. a unary relation) for each food item in our vocabulary (approximately 3300 words) regarding its healthiness. We distinguish between the following categories:

1. definitely healthy (e.g. *broccoli*)
2. mostly healthy (e.g. *cheese*)
3. mostly unhealthy (e.g. *coffee*)
4. definitely unhealthy (e.g. *potato crisps*)

Even though there exist other databases that list nutrient content of food items, such as the *USDA National Nutrient Database*, these databases are hardly accessible to laymen due to their complex labels. Our categorization that only employs four different labels presents a much simpler alternative.
Apart from its intrinsic relevance the aspect of healthiness may be particularly significant for the relations *Recommended-for-People-with* and *Not-Recommended-for-People-with*. If only (or at least mostly) healthy items appear in the former relation and unhealthy items appear in the latter relation, this would mean that a simple solution (or approximation) of covering the two relations would first retrieve food items relevant to a specific disposition. Then, the appropriateness or inappropriateness could be determined by answering the question whether a retrieved item is healthy or not. Thus, the detection of healthiness would be an auxiliary task for the detection of the two relations. We are, of course, aware that using a global healthiness detection as an auxiliary task is a simplification for the detection of relation types dealing with dispositions. Indeed, there are cases in which a food item generally considered healthy might not be suitable for a person with a certain disposition (e.g. *apples*, though considered very nutritious, should be avoided by people suffering from an *apple allergy*). If there are, however, only very few of these cases, the proposed approximation would serve as legitimate solution to the problem, especially, since it would be more efficient than the alternative that would require the modeling of *context-dependent* healthiness.

5. Annotation Guidelines and Statistics

5.1. General Set-up and Instructions

In order to have a reliable gold standard, each relation was separately annotated by two native speakers of German. As already stated in Section 3.3., annotators were given partially instantiated relations and they were to provide possible values for the unspecified argument position. We compiled a list of those relations comprising both very common and some more exotic items.
As we could not expect the annotators to compile the lists in an impromptu manner, they were allowed to consult various information sources for research, such as the internet. However, in order to obtain unbiased results they were specifically asked not to focus on a particular source, e.g. a particular website.
As we observed that the semantic strength varies quite notably across the different relation instances (e.g. though *apples* and *eggs* are both valid ingredients for *pancake*, the latter is a necessary ingredient while the former is not; therefore the relation instance *Ingredient-of(eggs, pancake)* is much stronger than *Ingredient-of(apples, pancake)*) we marked each instance as either a *strong* or *weak* relation instance.
We also set a timeout for finding argument values for one partially instantiated relation. This should ensure that we do not obtain too far-fetched relation instances. The specific timeout varies throughout the different relation types.
Table 1 illustrates some annotated partially annotated relations from our gold standard.

5.2. Defining the Food Vocabulary

A consistent annotation requires the usage of a unique vocabulary. For that reason, the annotators had to map each entity they had in mind to a synset in GermaNet, if there were an appropriate synset available. Using GermaNet synsets may be helpful for automatic processing. It is the only resource that allows to identify mentions of food items in natural language text, as this domain-specific entity type is not covered by off-the-shelf named-entity recognizers for German (Chrupala and Klakow, 2010; Faruqui and Padó, 2010). In case neither the term an annotator originally came up with nor any of its synonyms was contained in GermaNet, the concept was added in an additional vocabulary list. Of course, the two annotators notified each other about updates in this secondary vocabulary so that the usage of that term list was guaranteed to be consistent among the annotations as well. We also maintained the terms the

Suits-to(??, picnic)
sandwiches^s, wraps^s, noodle salad^s, potato salad^s, fruit salad^s, meat balls^s, filet of pork, vegetables^s, apples^s, melons^s, strawberries^s, muffins^s, biscuits^s, antipasti^s, ice tea, chicken salad^s, baguette^s, gazpacho, wine, vegetable pie, lemonade^s, mineral water^s, sugared slices, sausages^s, beer, cheese^s, chicken legs, apple turnover,
Can-be-Served-with(??, falafel)
lettuce^s, coleslaw, sauce^s, yoghurt^s, tomato salad^s, olives onions^s, sesame paste^s, pita^s, cucumbers^s, radish, fries, carrots
Can-be-Substituted-by(??, porridge)
millet gruel^s, muesli, semolina pudding^s, cornflakes, grits^s, oat meal, rice pudding, groats^s
Ingredient-of(??, apple pie)
apples^s, flour^s, eggs^s, sugar^s, cinnamon^s, yeast^s, baking powder^s, butter^s, milk^s, margarine^s, honey^s, almonds, almond paste, baking soda, sour cream, coconut oil, walnuts, raisins, lemons^s, custard^s, vanilla^s, orange juice^s, puff paste, poppy seeds, vanilla sugar^s, rum, gelatin, cider, cranberries, apricot jam^s, sugar powder, chocolate
Recommended-for-People-with(??, diabetes)
dietary fibre^s, fish^s, vegetables^s, lettuce^s, fruits, potatoes^s, magnesium,^{s} low-fat yoghurt^s, low-fat cheese^s, mineral water^s, unsweetened tea^s, muesli, pasta, colza oil^s, olive oil^s, rice^s, lean meat, whole-grain products^s, oat meal^s, linseed oil^s, corn oil^s, soya, vitamin C^s, vitamin D^s
Not-Recommended-for-People-with(??, diabetes)
alcohol^s, pastries^s, butter, soft drinks^s, sugar^s, convenience products^s, fat, sweets^s, honey^s, rice pudding, fructose^s, lactose^s, fries^s, sweetened bread spread^s, sorbite^s, nicotine^s, white bread^s, sausages

Table 1: Illustration of annotated relations (^{s} denotes items with strong association).

annotators originally came up with for the sake of completeness.

When a new concept needed to be added, the situation often occurred that several surface realizations (i.e. spelling variations, synonyms etc.) were eligible. Therefore, the annotators were instructed to choose that word form that is most commonly used. Such frequency information can be approximated by the number of hits a common search engine, such as *Google*, returns for a specific word form when used as a query.

There are also some word forms which are ambiguous, even within the food domain. For example, in German the term *Salat* may refer to both *lettuce* (i.e. the salad plant) and *salad* (i.e. the complete dish that in addition to the plant includes, for instance, dressing and herbs). This is why the annotators kept another (common) record which specified the reading they agreed upon for a specific word form. Ideally, the reading should correspond to the prototypical usage of the word, provided there is one.

5.3. Information Sources for Research

The types of information sources that the annotators made use of for their research were very diverse and also vary throughout the different relation types. The usage of the internet is, of course, essential for this annotation. User forums played a large role in finding entries, in particular, for the relation type *Suits-to*. For the relations involving dispositions, however, forums that provide expert advice (e.g. from medical doctors or complementary health practitioners) were found more useful. Such advice was also frequently discovered on websites from pharmacies or health insurances. Cookbooks (both online and hard copy) were an invaluable resource especially for finding entries for the relation type *Ingredient-of*. Online glossaries listing ingredients of food items were often used to determine typical characteristics that are vital for establishing similar food items for relation type *Can-be-Substituted-by*. As far the usage of search engines is concerned, we found that image search is much more effective for this task than text search. In particular for relation type *Can-be-Served-with*, the images retrieved for food items very often contained other food items, such as side dishes. In the retrieved text snippet, however, this information was less often contained.

5.4. Relation-specific Guidelines

For some relation types, the annotation is pretty straightforward. For others, some additional coding instructions were necessary:

5.4.1. *Can-be-Substituted-by*

For the relation type *Can-be-Substituted-by*, we explicitly formulated properties of food items that qualify as a substitute. For instance, for many dishes containing meat there exist vegetarian alternatives. As far as fruits and vegetables are concerned, the substitutes are often found within the set of food items that are grown during the same season. Moreover, if a food item is known for a particular nutrient content, then a substitute often shares this substance. These properties were just given as a *help* for finding food items. They should not be considered as hard criteria that substitutes always have to meet. In other words, there may be food items that are valid substitutes even though they do not possess any of these properties. However, our observation was that these properties are an effective means to find a large set of substitutes.

For this relation type, we let the annotators include synonyms. However, in order to distinguish them from the rest, they were specially marked. The purpose of including synonyms is that we anticipate automatic extraction methods to confuse near-synonyms (that should actually be found) and actual synonyms. A gold standard containing both categories should thus allow a more informative error analysis.

5.4.2. *Can-be-Served-with* versus *Ingredient-of*

Some given dishes were to be annotated for both the relation types *Can-be-Served-with* and *Ingredient-of*. In some of those cases, it is difficult to decide for some food items in which relation they should be used. A special rule of thumb was devised to help to distinguish these two relation types: If a food item is involved in the main processing

step of preparing a dish (i.e. cooking, frying, baking etc.) then it is a typical *ingredient*. If, however, the food item is added afterwards (e.g. *croutons* that are added to a *soup* or *parmesan* that is added to a *pasta dish*), then this is typically considered as a food item that is *served with* the dish. Of course, there are also food items which can be involved in the main processing step of a dish but may also be added afterwards. These were the only cases in which we allowed both relation types to be used. A typical example is <*sausage*, *pea soup*>. One may consider sausages as ingredients (in that case they are included during the main processing step, i.e. cooking) or they may be a side-dish (in that case they are not included in the main processing step, i.e. they are added to the dish once the soup has been cooked).

5.4.3. Relation Types Involving Dispositions

For the relation types involving dispositions (i.e. Sections 4.5. and 4.6.), a scientific proof of a causal relationship between a particular food item and a particular disposition was *not* required. There are several reasons that justify this decision: Firstly, it would have been beyond the scope of this annotation effort to verify that for every relation. Secondly, since we added a restriction that a relation must at least have been found at two different locations on the web, we assume that our gold standard will not include any doubtful individual opinion. Moreover, we believe that *popular wisdom* might be readily accepted by many customers.

For the relation type *Recommended-for-People-with*, we allowed both entries that may prevent the outbreak of a particular disposition and entries that are considered to contribute to curing it. We did not specially mark these two types, however. This is mostly due to the fact that food items often qualify for both subtypes.

A major problem that the annotators faced with these two relation types was finding an appropriate level of granularity for food items. The other relation types, more or less, required listing fairly specific food items. (Presumably, this is due to the fact that the given items of the partially instantiated relations were also fairly specific.) For the relation types that deal with dispositions, however, there can be cases in which entire groups of food items are beneficial or harmful for people suffering from a particular disposition. For instance, for a person having a *cold*, *fruits*, in general, may support their recovery. The question is what terms should be used to describe this issue. Should all fruits be listed or just the term *fruits*? We devised another rule of thumb that basically asked the annotators to stick to the level of granularity that is mentioned in text. As our gold standard is primarily designed for evaluating NLP systems, it should employ that terminology which will be mostly observed as input. If the given disposition were *cold*, this would mean including the general expression *fruits*. (To be more precise, we additionally listed some fruit types that are predominant in this context, such as *oranges*, as they will also be found in sentences in which this relation is expressed.)

For some dispositions, recommended/beneficial food items are not thought to be ingested but used in a different manner. For instance, some types of teas might be recommended for gurgling, or some herbs might be used as bath additive. We also included some of those items but marked them in a special way, so that they will not be confused with the regular food items that are ingested.

Finally, in order to prevent a contradictory annotation, the annotators were also asked to avoid food items being used with a specific disposition in both relation types.

5.4.4. Annotating Healthiness

In order to avoid a subjective healthiness annotation, we set up several additional guidelines. The decision whether a particular food item is healthy or not should be answered by primarily considering the following questions:

- What is the fat content of the food item? (Annotators should in particular focus on saturated fatty acids.)
- What is the sugar content of the food item?
- What is the salt content of the food item?
- Does the food item contain a substantial amount of artificial additives? Does the food item undergo extensive industrial processing?
- What are experts' opinions regarding the nutrient content of the food item?

The annotators used a set of websites that contained reliable information regarding these properties.[4] They were also asked to consider quantities of these food items that are realistically consumed by an average human. (Note that these realistic quantities differ greatly from those given by manufacturers of many snacks or sweets!) To simplify the annotation, the annotators were also given a list of prototypes for each category (e.g. *broccoli* for *definitely healthy*, *fruit juice* for *mostly healthy* etc.). Exotic items that were unknown to annotators and that were not mentioned on those reference websites were excluded from annotation.

5.5. Agreement in Annotation

We found that as far as the binary relations were concerned (Sections 4.1.-4.6.) there was only a match of 35% among the food items that the annotators suggested. After discussing with them a sample of partially instantiated relations whose suggested food items contained particularly few matches, we found that both annotators were consistently following our strict annotation guidelines. We found neither any evidence for underspecified instructions in the guidelines nor any systematic differences in the annotations of the two annotators[5] implying that they might have annotated two different concepts. Therefore, we conclude that the low number of matches simply draws from the fact that, apparently, two human annotators detect more relevant items than just one. This is quite plausible since the annotators come from two different regions of Germany (i.e.

[4] Example websites are:
`www.naehrwertrechner.de`
`ndb.nal.usda.gov/ndb/foods/list`

[5] For instance, one annotator might have focused on listing specific food items while the other might have employed more general terminology.

Northern and Southern Germany) whose cuisines differ in several respects.

For the unary relation labeling food items with regard to healthiness (Section 4.7.), a stronger agreement is required as an evidence for the suitability of our annotation guidelines. Indeed, the agreement was much stronger. We computed Cohen's Kappa (κ) on a sample of 100 food items and measured $\kappa = 0.61$ considering all classes. For binary classification[6], we measured $\kappa = 0.78$. Both scores can be considered as some *good agreement.* For this annotation, we will release a merged version in addition to the individual annotations. (For the remaining annotations, we will exclusively provide the individual annotations.)

5.6. Statistics

Table 2 lists the amount of partially annotated relations that have been annotated for each relation type. Moreover, the average number of values for the unspecified argument slot in a partially annotated relation is displayed. We list these figures for both annotators individually and also differentiate between entries that have either a strong or a weak association towards a particular relation (Section 5.1.).[7] As we do not provide a merged version of the dataset, we also consider the union of the partially annotated relations from the two annotators. (We do not consider the intersection, as our analysis comparing the annotations of the two annotators suggested that the annotations are complementary rather than concordant (Section 5.5.).) With regard to this union, we ignore the strength labels as there are frequent cases of food items listed by both annotators being assigned different strength information. This suggests that this label is rather subjective.

The most striking result from Table 2 is that, on average, Annotator 1 has found more entities that Annotator 2. Moreover, this difference is greater on items marked with *weak* association. We ascribe this general quantitative difference to the different search strategies of the annotators, such as the choice of websites they used for their research. We deliberately let the annotators develop their own search strategy as otherwise we would have posed some bias onto their annotation.

Finally, we also have a look at how the different relation types overlap. For that experiment we exclusively consider those pairs of relation types which share the same types of entity pairs, i.e. *Can-be-Served-with*, *Can-be-Substituted-by* and *Ingredient-of*, which consider the entity types <FOOD-ITEM, FOOD-ITEM>[8], and *Recommended-for-People-with* and *Not-Recommended-for-People-with*, which consider the entity types <FOOD-ITEM, DISPOSITION>. We measure the overlap of two relation types by counting the number of tuples (e.g. <*sausage*, *pea soup*>) that occur with both relation types. In order to provide numbers that can be compared across the different pairs of relation types, we normalize the number of common tuples by the overall number of tuples that can be observed with either of the relation types of a specific pair. Table 3 lists the overlap not only for each annotator but also for the union of both annotations (similar to Table 2). Ideally, the overlap between the different relation types is very small. This would indicate that we have defined fairly disjoint classes which is a prerequisite for a successful automatic classification.

There is only one pair that has a notably higher overlap of relation types being *Can-be-Served-with* and *Can-be-Substituted-by*. We already addressed this pair in Section 5.4.2. where we also presented additional guidelines to separate these classes. Even though this pair has a higher overlap, this does not necessarily mean that our relation types are ill-defined or the annotators did not consistently annotate. Section 5.4.2. already pointed out that there can be cases in which tuples appear with both relation types. However, we hope that due to our additional guidelines we have managed to reduce these cases.

It may come as a surprise that we included the pair involving dispositions in Table 3. In Section 5.4.3., we presented a guideline that prohibited the usage of a food item to appear with a specific disposition in both relation types. Therefore, the overlap of 0 is just an indication that the annotators obeyed the annotation guidelines. We actually included this list in order to show that if we automatically merge the annotations of the two annotators we may not fully preserve this separation but the *noise* that is thereby introduced (i.e. an overlap of 0.006) can be considered negligible.

Our resource, that is, all annotations that have been described in this paper along their respective coding instructions, will be made available for research purposes.[9]

6. Conclusion

In this paper, we presented a gold standard for semantic relation extraction in the food domain for German. Apart from describing the annotation scheme that we devised in detail and suggesting how the annotation can be used to train and evaluate natural language systems, we also outlined scenarios in which this data set may become very useful. Finally, we hope that our resource will foster new research in natural language processing in the food domain.

Acknowledgements

The work presented in this paper was performed in the context of the Software-Cluster project EMERGENT (www.software-cluster.org). It was funded by the German Federal Ministry of Education and Research (BMBF) under grant no. "01IC10S01". The authors assume responsibility for the content.

7. References

Timothy Chklovski and Patrick Pantel. 2004. VerbOcean: Mining the Web for Fine-Grained Semantic Verb Rela-

[6] We obtained two classes by conflating the categories *definitely (un)healthy* and *mostly (un)healthy*.

[7] We omit the label for *synonyms* in this statistics as it is only included for partially annotated relations of relation type *Can-be-Substituted-by* (Section 5.4.1.). Even for this relation type, that label plays a minor role, as in the annotations of both annotators there is less than one occurrence on average per partially annotated relation.

[8] According to Section 4.4., *Ingredient-of* comprises tuples <FOOD-ITEM, DISH> but DISH is a subtype of FOOD-ITEM.

[9] available at: `www.lsv.uni-saarland.de/personalPages/michael/relFood.html`

		Annotator 1			Annotator 2			Union
Relation Type	**#PIRs**	**strong**	**weak**	**all**	**strong**	**weak**	**all**	**all**
Suits-to	40	35.70	36.43	72.13	26.73	12.98	39.70	87.38
Can-be-Served-with	58	11.78	15.74	27.74	17.86	13.62	31.48	49.93
Can-be-Substituted-by	70	4.73	7.00	11.77	8.51	6.64	15.34	22.89
Ingredient-of	49	18.33	49.84	68.20	25.76	16.65	42.41	82.78
Recommended-for-People-with	47	46.47	19.77	66.23	29.19	17.09	46.28	85.19
Not-Recommended-for-People-with	47	29.62	16.47	46.09	13.11	5.91	19.02	48.32
Average	51.83	24.44	24.21	48.69	20.19	12.15	32.37	62.75

Table 2: Statistics of partially instantiated relations (PIRs) to be annotated per relation type and values found for the unspecified argument slot (the values are listed for each annotator and the union of both annotations).

		Overlap		
Relation Types	**Typed Tuple**	**Annotator 1**	**Annotator 2**	**Union**
Can-be-Served-with $\oplus$ Can-be-Substituted-by	<FOOD-ITEM, FOOD-ITEM>	0.012	0.006	0.012
Can-be-Served-with $\oplus$ Ingredient-of	<FOOD-ITEM, FOOD-ITEM>	***0.126***	***0.027***	***0.117***
Can-be-Substituted-by $\oplus$ Ingredient-of	<FOOD-ITEM, FOOD-ITEM>	0.010	0.012	0.014
Recommended-for-People-with $\oplus$ Not-Recommended-for-People-with	<FOOD-ITEM, DISPOSITION>	0.000	0.000	0.006

Table 3: Overlap of tuples between different relation types having the same entity types (overlap is normalized by the overall number of tuples occurring with either relation type).

tions. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Grzegorz Chrupala and Dietrich Klakow. 2010. A Named Entity Labeler for German: Exploiting Wikipedia and Distributional Clusters. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, La Valletta, Malta.

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

Manaal Faruqui and Sebastian Padó. 2010. Training and Evaluating a German Named Entity Recognizer with Semantic Generalization. In *Proceedings of KONVENS*, Saarbrücken, Germany.

Roxana Girju, Adriana Badulescu, and Dan Moldovan. 2003. Learning Semantic Constraints for the Automatic Discovery of Part-Whole Relations. In *Proceedings of the Human Language Technology Conference (HLT)*.

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

Marti A. Hearst. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 539–545.

Niklas Jakob and Iryna Gurevych. 2010. Extracting Opinion Targets in a Single- and Cross-Domain Setting with Conditional Random Fields. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Boston, MA, USA.

Heng Ji, Ralph Grishman, Hoa Trang Dang, Kira Griffitt, and Joe Ellis. 2010. Overview of the TAC KBP 2010 Knowledge Base Population Track. In *Proceedings of the Text Analytics Conference (TAC)*.

Jason S. Kessler and Nicolas Nicolov. 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.

Li Liu and Fangfang Li. 2011. 3-layer Ontology Based Query Expansion for Searching. In *Proceedings of the International Symposium on Neural Networks (ISNN)*, pages 621 – 628. Springer-Verlag.

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

Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2006. Semantic Taxonomy Induction from Heterogenous Evidence. In *Proceedings of the International Conference on Computational Linguistics and Annual Meeting of the Association for Computational Linguistics (COLING/ACL)*.

Peter D. Turney, Michael L. Littman, Jeffrey Bigham, and Victor Shnayder. 2003. Combining Independent Modules to Solve Multiple-choice Synonym and Analogy Problems. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*.

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

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

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