Data-driven Knowledge Extraction for the Food Domain

Michael Wiegand, Benjamin Roth, Dietrich Klakow
Spoken Language Systems
Saarland University
D-66123 Saarbrücken
{michael.wiegand|benjamin.roth|dietrich.klakow}@lsv.uni-saarland.de

Abstract

In this paper, we examine methods to automatically extract domain-specific knowledge from the food domain from unlabeled natural language text. We employ different extraction methods ranging from surface patterns to co-occurrence measures applied on different parts of a document. We show that the effectiveness of a particular method depends very much on the relation type considered and that there is no single method that works equally well for every relation type. We also examine a combination of extraction methods and also consider relationships between different relation types. The extraction methods are applied both on a domain-specific corpus and the domain-independent factual knowledge base *Wikipedia*. Moreover, we examine an open-domain lexical ontology for suitability.

1 Introduction

There has been only little research on natural language processing in the food domain even though there is a high commercial potential in automatically extracting knowledge involving food items. For example, such knowledge could be beneficial for virtual customer advice in a supermarket. The advisor might suggest products available in the shop that would potentially complement the items a customer has already in their shopping cart. Additionally, food items required for preparing a specific dish or typically consumed at a social occasion could be recommended. The advisor could also suggest an appropriate substitute for a product a customer would like to purchase if that product is out of stock.

In this paper, we present methods to automatically extract knowledge from the food domain. We apply different relation extraction methods, such as simple manually designed surface patterns or statistical co-occurrence measures, on both a domain-specific corpus and the open-domain factual knowledge base *Wikipedia*. These large corpora are exclusively used as *unlabeled* data. In addition to the corpora, we also assess an open-domain lexical ontology. Moreover, we combine these methods and harness the relationship between different relation types. Since these methods only require a low level of linguistic processing, they have the advantage that they can provide responses in real time. We show that these individual methods have varying strength depending on which particular food relation type is considered.

Our system has to solve the following task: It is given a *partially instantiated relation*, such as *Ingredient-of(FOOD-ITEM=?, pancake)*. The system has to produce a ranked list of possible values that are valid arguments of the unspecified argument position. In the current example, this would correspond to listing ingredients that are necessary in order to prepare *pancakes*, such as *eggs*, *flour*, *sugar* and *milk*. The entities that are to be retrieved are always food items. Moreover, we only consider binary relations. The relation types we examine (such as *Ingredient-of*) are domain specific.

2 Related Work

Previous work on relation extraction focused on domain-independent semantic relations, such as hyponyms (Hearst, 1992; Snow et al., 2006; Pantel and Pennacchiotti, 2006), meronyms (Girju et al., 2003; Pantel and Pennacchiotti, 2006), synonyms (Chklovski and Pantel, 2004), general purpose analogy relations (Turney et al., 2003) and general relations involving persons or organizations (Ji et al., 2010).

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. 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. The work which is most closely related to this paper is (Wiegand et al., 2012a) in which extraction methods are examined for the relations that we also address in this paper. However, this paper extends the preliminary study presented in (Wiegand et al., 2012a) in many ways: Apart from providing a more detailed explanation for the performance of the different extraction methods, we also compare them on different types of text data (i.e. domain-specific and domain-independent data). Moreover, we assess in how far existing general-purpose resources (we examine *GermaNet* (Hamp and Feldweg, 1997)) might help for this task. In addition, we propose ways to improve extraction performance by combining different extraction methods and considering inter-relationships between the different relation types.

Many approaches to recognize relations employ some form of patterns. These patterns can be either manually specified (Hearst, 1992), fully automatically learned (Girju et al., 2003) or semi-automatically learned (Chklovski and Pantel, 2004; Pantel and Pennacchiotti, 2006). The levels of representation that are considered in these patterns also vary. For some tasks, elaborate patterns using syntactic information are applied (Hearst, 1992; Girju et al., 2003; Chklovski and Pantel, 2004; Pantel and Pennacchiotti, 2006). For others, very simple lexical patterns are employed (Turney and Littman, 2003). In particular for web-based approaches, the latter are much easier to cope with as they allow the patterns to be used as ordinary queries for search engines. In addition to the usage of patterns, some statistical co-occurrence measures have also been successfully used to extract certain relations (Turney and Littman, 2003). While patterns are more generally applicable, the usage of co-occurrence measures is only effective if large amounts of data, for instance the web, are used as a dataset.

3 Data and Resources

For our experiments we use a crawl of *chefkoch.de*[1] as a domain-specific dataset. *chefkoch.de* is the largest web portal for food-related issues in the German language. Note that we only consider the *forum* of this website for our experiments. The website also contains some more structured information, such as a *recipe*-section, but this knowledge could only be extracted by writing a rule-based parser processing the idiosyncratic format of those webpages which would – unlike the approaches examined in this paper – not be generally applicable. We obtained the crawl by using *Heritrix* (Mohr et al., 2004). The plain text from the crawled set of web pages is extracted by using *Boilerpipe* (Kohlschutter et al., 2010). The final domain-specific corpus consists of 418,558 webpages (3GB plain text).

In order to use Wikipedia, we downloaded the current dump of the German version of Wikipedia.[2] This pre-processed corpus contains 385,366 articles (4.5GB plain text).[3] All corpora are lemmatized by using *TreeTagger* (Schmid, 1994). In order to have an efficient data access we index the corpora with *Lucene* (McCandless et al., 2010). As a domain-independent lexical database, we use *GermaNet* (Hamp and Feldweg, 1997) which is the German counterpart of *WordNet* (Miller et al., 1990). We use Version 5.3.

[1] `www.chefkoch.de`

[2] The dump was downloaded in the fourth quarter of 2011.

[3] Note that we only processed articles from Wikipedia that contain mentions of food items.

4 The Different Relations Types

In this section, we will briefly describe the four relation types we address in this paper. We just provide English translations of our German examples in order to ensure general accessibility.

- ***Suits-to(FOOD-ITEM, EVENT)*** 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*>.
- ***Served-with(FOOD-ITEM, FOOD-ITEM)*** describes food items that are typically consumed together. Examples are <*fish fingers, mashed potatoes*>, <*baguette, ratatouille*> or <*wine, cheese*>.
- ***Substituted-by(FOOD-ITEM, FOOD-ITEM)*** 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*>.
- ***Ingredient-of(FOOD-ITEM, DISH)*** denotes some ingredient of a particular dish. Examples are <*chickpea, falafel*> or <*rice, paella*>.

5 Challenges of this Task

The extraction of relations from the food domain yields some particular challenges. The most striking problem of this task is that the language to be employed to express a specific relation type can be very diverse. For example, the relation type *Ingredient-of* is often expressed in the context of a cooking instruction. Thus, the language that is used may be very specific to the procedure of preparing a particular dish. For instance, Example 1 expresses the relation instance *Ingredient-of(cooking oil, pancake)*. As such, it would be extremely difficult to employ some textual patterns in order to detect this relation as the relevant entities *cooking oil* and *pancake* are not contained within the same sentence. Even if there were a means to acquire such a long-distance pattern, one may doubt that this pattern would capture other relation instances, such as *Ingredient-of(mince meat, lasagna)*, as many of those involve other procedural patterns.

1. Pour some *cooking oil* into the pan. Add 1/6 of the dough and fry the *pancake* from both sides. [Relation: *Ingredient-of(oil, pancake)*]

The inspection of our data using some seed examples displaying prototypical relation instances of our relation types (e.g. <*hot dog, fries*> for *Served-with*) revealed that – despite the language variability – there are some very simple and general textual patterns, for example the pattern *FOOD-ITEM and FOOD-ITEM* for *Served-with* as illustrated in Example 2. However, many of those simple patterns are ambiguous and can also be observed with other relation types. For instance, the pattern mentioned above could also imply the relation type *Substituted-by* as in Example 3.

2. We both had a *hot dog* and *fries*. [Relation: *Served-with(fries, hot dog)*]
3. I'm looking for a nice fish-recipe for someone who does not like plain fish but who eats *fish fingers* and *fish cake*. [Relation: *Substituted-by(fish fingers, fish cake)*]

Since three of four of our relation types are relation types between two entities of type FOOD-ITEM[4], there is a high likelihood that two relation types are confused with each other. This would also suggest that the remaining relation type, which is a relation type between entities of types FOOD-ITEM and EVENT, namely *Suits-to*, is easier to cope with. An obvious solution to detect this relation type is just to consider the co-occurrence of two entities with these particular entity types. For a mention of that relation type, such as Example 4, this would work. However, some mechanism must be provided in order to distinguish those meaningful co-occurrences from coincidental ones, such as Example 5.[5]

4. There will be six of us at *Christmas*. I'd like to prepare a *goose*. [Relation: *Suits-to(goose, Christmas)*]
5. Last *Christmas*, I got a moka-pot. Unfortunately, I don't know how to make a proper *espresso*.

6 Method

In the following, we describe the individual extraction methods that are examined in this paper. The first three methods (Sections 6.1-6.3) are taken from (Wiegand et al., 2012a).

6.1 Surface Patterns (PATT)

As surface patterns, manually compiled patterns are exclusively considered. The patterns comprise a set of few generally-applicable and fairly precise patterns. As a help for building such patterns, Wiegand et al. (2012a) recommend to look at mentions of typical relation instances

[4]Note that the entity type DISH in *Ingredient-of* is a subset of FOOD-ITEM.

[5]That is, Example 5 does *not* express the relation *Suits-to(espresso, Christmas)*.

in text corpora, e.g. <*butter, margarine*> for *Substituted-by* or <*mince meat, meat balls*> for *Ingredient-of*.

As already stated in Section 5, the formulation of such patterns is difficult due to the variety of contexts in which a relation can be expressed. Wiegand et al. (2012a) confirm this by computing lexical cues automatically with the help of statistical co-occurrence measures, such as the *pointwise mutual information*, which have been run on automatically extracted sentences containing mentions of typical relation instances. The output of that process did not reveal any significant additional patterns.

The final patterns exclusively use lexical items immediately before, between or after the argument slots of the relations. Table 1 illustrates some of these patterns. The level of representation used for those patterns (i.e. word level) is very shallow. However, these patterns are precise and can be easily used as a query for a search engine. Other levels of representation, e.g. syntactic information, would be much more difficult to incorporate. Moreover, Wiegand et al. (2012a) report that they could not find many frequently occurring patterns using these representations to find relation instances that could not be extracted by those simple lexical patterns. Additionally, since the domain-specific data to be used comprise informal user generated natural language, the linguistic processing tools, such as syntactic parsers, i.e. tools that are primarily built with the help of formal newswire text corpora, are severely affected by a domain mismatch.

The extraction method PATT comprises the following steps: Recall from the task description in Section 1 that we always look for a list of values for an unspecified argument in a partially instantiated relation (PIR) and that the unspecified argument is always a food item. Given a PIR, such as *Substituted-by(butter, FOOD-ITEM=?)*, we partially instantiate each of the pertaining patterns (Table 1) with the given argument (e.g. *FOOD-ITEM instead of FOOD-ITEM* becomes *FOOD-ITEM instead of butter*) and then check for any possible food item (e.g. *margarine*) whether there exists a match in our corpus (e.g. *margarine instead of butter*). The output of this extraction process is a ranked list of those food items for which a match could be found with any of those patterns. We rank by the frequency of matches. Food items are obtained using GermaNet. All those lexical items are collected that are contained within the synsets that are hyponyms of *Nahrung* (English: *food*).

6.2 Statistical Co-occurrence (CO-OC)

The downside of the manual surface patterns is that they are rather sparse as they only fire if the exact lexical sequence is found in our corpus. As a less constrained method, one may therefore also consider statistical co-occurrence. The rationale behind this approach is that if a pair of two specific arguments co-occurs significantly often (at a certain distance), such as *roast goose* and *Christmas*, then there is a likely relationship between these two linguistic entities.

By applying a co-occurrence measure one may be able to separate meaningful from coincidental co-occurrences as exemplified in Examples 4 and 5 in Section 5. As a co-occurrence measure, we consider the *normalized Google distance (NGD)* (Cilibrasi and Vitanyi, 2007) which is a popular measure for such tasks. The extraction procedure of CO-OC is similar to PATT with the difference that one does not rank food items by the frequency of matches in a set of patterns (all containing the given entity) but the correlation score with the given entity. For instance, given the PIR *Suits-to(FOOD-ITEM=?, Christmas)*, one computes the scores for each food item from our (food) vocabulary and *Christmas* and sorts all these food items according to the correlation scores.

It is believed that this approach is beneficial for relations where the formulation of surface patterns is difficult – this is typically the case when entities involved in such a relation are realized within a larger distance to each other. Thus, CO-OC would tackle one challenge that was presented in Section 5.

6.3 Relation between Title and Body of a Webpage (TITLE)

Rather than computing statistical co-occurrence at a certain distance, one may also consider the co-occurrence of entities between title and body of a webpage. Wiegand et al. (2012a) argue that

Relation Type	#Patterns	Examples
Suits-to	6	FOOD-ITEM at EVENT; FOOD-ITEM on the occasion of EVENT; FOOD-ITEM for EVENT
Served-with	8	FOOD-ITEM and FOOD-ITEM; FOOD-ITEM served with FOOD-ITEM; FOOD-ITEM for FOOD-ITEM
Substituted-by	8	FOOD-ITEM or FOOD-ITEM; FOOD-ITEM (FOOD-ITEM); FOOD-ITEM instead of FOOD-ITEM
Ingredient-of	8	DISH made of FOOD-ITEM; DISH containing FOOD-ITEM

Table 1: Illustration of the manually designed surface patterns.

entities mentioned in the title represent a predominant topic and that a co-occurrence with an entity appearing in the body of a webpage may imply that the entity has a special relevance to that topic and denote some relation. The co-occurrence of two entities in the body is more likely to be coincidental. None of those entities needs to be a predominant topic.

The extraction procedure of this method selects those documents that contain the given argument of a PIR (e.g. *lasagna* in *Ingredient-of(FOOD-ITEM=?, lasagna)*) in the title and ranks food items that co-occur in the document body of those documents according to their frequency.

6.4 Wikipedia Links (LINK)

Since we also evaluate Wikipedia as a corpus for our relation extraction task, we also want to take into account a feature that is specific to this type of resource, namely *Wikipedia links*. According to the guidelines of Wikipedia[6], links are typically used to connect some article X to another article Y that a reader of article X might be also interested in. Similar to TITLE, we want to examine whether these links have any specific semantics for our domain task.

Using Wikipedia links we extract relations in the following way: The given argument of a PIR is the source article and we rank food items whose articles are linked to from this source article according to their frequency.[7]

6.5 GermaNet - Sibling Synsets in the Hyperonym Graph (GERM)

Finally, we also examine a general-purpose ontology for German, namely GermaNet. This resource organizes different general relations (e.g. *hyperonymy*, *hyponomy* or *meronymy*) between different synsets being groups of words with a similar meaning. The assignment of a given food item (or event) to a particular synset is simple as these expressions are usually unambiguous in our domain. Since GermaNet is an open-domain ontology it does not specialize for the relations that we consider in this work. Of our relation types, only *Substituted-by* can be modeled with the help of GermaNet.[8] We found that the *sibling* relationship between different synsets in the hyperonym graph encodes a very similar concept. For instance, *apple*, *pear*, *quince* and *guava* are siblings (their immediate hyperonym is *pome*) and, therefore, they are likely to be substituted by each other. Of course, the degree of similarity also depends on the location of those siblings within that graph. The more specific a synset is (i.e. the deeper it is within the graph), the more similar are its *siblings* to it. We found that the type of similarity that we want to model can only be reliably preserved if the target synset is actually a leaf node. Otherwise, we would also obtain *meat* and *pastries* as an entry for *Substituted-by*. They, too, are siblings (*solid food* is their immediate hyperonym) but these entries are not leaves in the hyperonym graph.

Unlike the other extraction methods there is no straightforward way for this method to provide a ranking of the food items that are extracted. That is why we evaluate them in random order.

7 Experiments

We already stated in Section 1 that the unspecified argument value of a partially instantiated relation (PIR) is always of type FOOD-ITEM. This

[6] `en.wikipedia.org/wiki/Wikipedia:Link`

[7] We do not apply any correlation measure for LINK because of the same reasons as we do not apply them for TITLE.

[8] Conceptually speaking, a second relationship, namely *Ingredient-of*, could be recognized with the help of the *meronymy* (part-of) relation of GermaNet. Unfortunately, there exist virtually no entries for food items with regard to that relation.

Partially Instantiated Relations (PIRs)	#PIRs
Suits-to(FOOD-ITEM=?, EVENT)	40
Served-with(FOOD-ITEM, FOOD-ITEM=?)	58
Substituted-by(FOOD-ITEM, FOOD-ITEM=?)	67
Ingredient-of(FOOD-ITEM=?, DISH)	49

Table 2: Statistics of partially instantiated relations in gold standard.

is because these PIRs simulate a typical situation for a virtual customer advisor, e.g. such an advisor is more likely to be asked what food items are suitable for a given event, i.e. *Suits-to(FOOD-ITEM=?, EVENT)*, rather than the opposite PIR, i.e. *Suits-to(FOOD-ITEM, EVENT=?)*. The PIRs we use are presented in Table 2.[9]

We use the gold standard from (Wiegand et al., 2012b) for evaluation.[10] For each relation, a certain number of PIRs has been manually annotated (see also Table 2).

Since our automatically generated output are ranked lists of food items, we use *precision at 10 (P@10)* and *mean reciprocal rank (MRR)* as evaluation measures. The two metrics are to some extent complementary. While P@10 evaluates the matches with the gold standard on the 10 most highly ranked items not taking into account on what positions the correct items appear, MRR just focuses on the highest ranked correct item but it also considers the corresponding ranking position.

7.1 Individual Evaluation of the Different Extraction Methods

Table 3 compares the different individual methods on all of our four relation types. (Note that for CO-OC, we consider the best window size for each respective relation type.) For each relation type, the best extraction is achieved with the help of our domain-specific corpus (chefkoch.de). This proves that the choice of the corpus is at least as important as the choice of the method.

Table 3 also shows that the performance of a particular method varies greatly with respect to the relation type on which it has been applied. For *Suits-to*, the methods producing some reasonable output are CO-OC and TITLE. For *Served-with*, PATT and CO-OC are effective. For *Substituted-by*, the clear winner is PATT. Not even the lexical resource GermaNet (GERM) can be harnessed in order to have some comparable output. For *Ingredient-of*, TITLE performs best. For that relation type, LINK also provides some reasonable performance. In terms of coverage (P@10 is the more indicative measure for that), we obtain much better results for *Ingredient-of* than for the other relation types. This can be ascribed to the fact that this is obviously the most discussed relation type. Since LINK and TITLE-Wikipedia are very similar in their nature, we manually inspected the output of those methods for some queries in order to find out why LINK performs so much better. We found that LINK is usually a proper subset of what is retrieved by TITLE-Wikipedia. This subset is much more relevant for *Ingredient-of* than the larger list from TITLE-Wikipedia. The fact that *Ingredient-of* is the only relation type which can be properly extracted with the help of Wikipedia does not come as a surprise as the knowledge encoded in that relation type is mostly factual while the other relation types are influenced by social conventions (mostly *Suits-to*) and common taste (*Served-with* and *Substituted-by*). The latter two issues are less present in Wikipedia.

7.2 Interpreting the Results

The results of Table 3 prove that we can partly solve the challenges presented in Section 5. *Suits-to* and *Ingredient-of* can be successfully extracted using methods that bypass the modeling of the difficult surface realizations. TITLE is very effective for *Ingredient-of*, i.e. it produces a fairly unambiguous output. Thus, for this relation type, we have found a method that does not confuse this relation type with the other relation types exclusively involving entities of type FOOD-ITEM (i.e. *Served-with* and *Substituted-by*).

Table 4 takes a closer look at CO-OC in that it compares different window sizes. (Note that we only consider MRR as this measure is more sensitive to changes in ranking quality than P@10.) This comparison may shed some light into why

[9] Since the two relation types *Served-with* and *Substituted-by* are reflexive, the argument positions of the PIRs do not matter.

[10] Following Wiegand et al. (2012a), we carried out our experiments on an earlier version of that gold standard. Therefore, the statistics regarding PIRs differ between this work and (Wiegand et al., 2012b).

		Suits-to		Served-with		Substituted-by		Ingredient-of	
Method	**Resource**	**P@10**	**MRR**	**P@10**	**MRR**	**P@10**	**MRR**	**P@10**	**MRR**
GERM	GermaNet	*NA*	*NA*	*NA*	*NA*	0.191	0.322	*NA*	*NA*
PATT	Wikipedia	0.000	0.000	0.048	0.131	0.024	0.177	0.000	0.000
CO-OC	Wikipedia	0.138	0.417	0.086	0.152	0.076	0.315	0.114	0.215
TITLE	Wikipedia	0.095	0.186	0.076	0.173	0.051	0.160	0.267	0.186
LINK	Wikipedia	0.000	0.000	0.083	0.155	0.058	0.214	0.400	0.646
PATT	chefkoch.de	0.023	0.133	**0.343**	**0.617**	**0.303**	**0.764**	0.076	0.331
CO-OC	chefkoch.de	**0.340**	**0.656**	0.310	0.584	0.172	0.553	0.335	0.581
TITLE	chefkoch.de	0.300	0.645	0.171	0.233	0.049	0.184	**0.776**	**0.733**

Table 3: Comparison of the different individual methods (for CO-OC the best window size is considered).

Window	**Suits-to**	**Served-with**	**Substituted-by**	**Ingredient-of**
2	0.371	**0.584**	**0.553**	0.372
5	0.511	0.545	0.537	0.496
10	0.579	0.544	0.527	0.536
20	0.644	0.532	0.534	**0.581**
50	**0.656**	0.469	0.512	0.558
sentence	0.525	0.550	0.515	0.544
document	0.618	0.431	0.500	0.377

Table 4: MRR of *CO-OC* using different window sizes.

a particular method only works for certain relations. Small window sizes are very effective for *Served-with* and *Substituted-by*. This means that the entities involved in those relations tend to appear close to each other. This is a pre-requisite that our short-distance patterns (PATT) fire. For the other relation types, in particular *Suits-to*, the entities involved can be fairly far apart from each other (i.e. 50 words in between). For such relation types short-distance patterns are not effective.

Table 4 also shows that more natural boundaries for the entities involved in a relation, i.e. the sentence and the entire document, are less effective than choosing a fixed window size.

7.3 Combination of Extraction Methods

Table 5 compares the performance of the best individual method for each relation type with some combination. The combination always uses the best performing individual method (for each respective relation type) and the method which in combination with the best gives the largest improvement (this is usually the second best method). We experimented with standard merging methods of rankings, such as linear interpolation or multiplication of the inverted ranks. However, they did not result in a notable improvement. Presumably, this is due to the fact that the output of several methods – this is usually the method that is combined with the best individual method – cannot be suitably represented as a ranking as the entries are more or less equipollent. This is most evident for GERM (as discussed in Section 6.5) but it also applies for LINK. For the latter, we may create a ranking based on frequency but since most links are only observed once or twice, this criterion is hardly discriminant. We therefore came up with another combination procedure that reflects this property. We mainly preserve the ranking produced by the best individual method but boost entries that also occur in the output of the other method considered since these entries should be regarded most reliable. We empirically increase the rank of those entries by n ranking positions. In order to avoid overfitting we just consider three configurations for n: 5, 10 and 20. Table 5 shows that, indeed, some improvement can be achieved by this combination scheme.

7.4 Relationship between Relation Types

Finally, we also examine whether one can improve performance of one relation type by considering some relationship towards another relation type. Recall that *Served-with*, *Substituted-by* and *Ingredient-of* are all relation types between two food items. Therefore, there is a chance that those three relation types get confused. We found

	Suits-to		Served-with		Substituted-by		Ingredient-of	
	P@10	**MRR**	**P@10**	**MRR**	**P@10**	**MRR**	**P@10**	**MRR**
best individual	0.340	0.656	0.343	0.617	0.303	0.764	**0.776**	0.733
combination	**0.365**†	**0.722***	**0.378**‡	**0.648***	**0.310**	**0.794**	0.773	**0.835**‡
methods	CO-OC$_{ch}$+TITLE$_{ch}$		PATT$_{ch}$+CO-OC$_{ch}$		PATT$_{ch}$+GERM		TITLE$_{ch}$+LINK	
ranking increase	10 ranks		20 ranks		5 ranks		5 ranks	

Table 5: Comparison of the best individual method and the best combination of (two) methods. $_{ch}$ indicates that this method has been applied on *chefkoch.de*; significantly better than *best individual* *: at $p < 0.1$; †: at $p < 0.05$; ‡: at $p < 0.01$ (paired t-test).

Suits-to(?, picnic)	Served-with(broccoli, ?)	Substituted-by(beef roulades, ?)	Ingredient-of(?, falafel)
sandwiches*	broad noodles	goulash*	chickpea*
fingerfood	potatoes*	roast*	cooking oil*
noodle salad*	salt potatoes*	roast beef*	garlic*
meat balls*	croquettes	braised meat*	water
potato salad*	sweet corn	marinated beef*	coriander*
melons*	spaetzle	rolled pork*	onions*
fruit salad*	noodle casserole	roast pork*	parsley*
small sausages	fillet of pork*	cutlet	flour*
sparkling wine	mushrooms*	ragout	cumin*
baguette*	rice*	rabbit	salt*

Table 7: The 10 most highly ranked food items for some automatically extracted relations; *: denotes match with the gold standard.

Method	P@10	MRR
Served-with$_{ind}$	0.343	0.617
Served-with$_{comb}$	0.378	0.648
Served-with$_{ind}$ + ¬Substituted-by$_{ind}$	0.393	0.698
Served-with$_{comb}$ + ¬Substituted-by$_{comb}$	**0.431**†	**0.754***

Table 6: Filtering *Served-with* with the output of *Substituted-by*; *ind*: best individual method from Table 5; *comb*: combination method from Table 5; significantly better than *Served-with*$_{ind}$ + ¬*Substituted-by*$_{ind}$ *: at $p < 0.05$; †: at $p < 0.01$ (paired t-test).

that *Served-with* is most affected by this as it gets mostly confused with *Substituted-by*. This comes as no surprise as Table 4 showed that these relation types are very similar with respect to the distance in which their participating entities appear to each other. We try to improve the extraction of *Served-with* by deleting those entries that have also been retrieved for *Substituted-by* (we denote this by ¬*Substituted-by*). Table 6 shows that this filtering method largely increases the performance of *Served-with*.

Table 7 illustrates some automatically generated output using the best configuration for each relation type. Even though not all retrieved entries match with our gold standard, most of them are (at least) plausible candidates. Note that for our gold standard we aimed for high precision rather than completeness.

8 Conclusion

In this paper, we examined methods for automatically extract knowledge for the food domain from unlabeled text. We have shown that different relation types require different extraction methods. We compared different resources and found that a domain-specific corpus consisting of web forum entries provides better coverage of the relations we are interested in than the open-domain data we examined. Further improvement can be achieved by combining different methods that may also rely on different resources and using interrelationships between different relation types. Since our methods only require a low level of linguistic processing, they may serve for applications that have to provide responses in real time.

More information about this work including a demo can be found at: `www.lsv.uni-saarland.de/personalPages/michael/relFood.html`

Acknowledgements

This work was funded by the German Federal Ministry of Education and Research (Software-Cluster) under grant no. “01IC10S01”.

References

Timothy Chklovski and Patrick Pantel. 2004. VerbOcean: Mining the Web for Fine-Grained Semantic Verb Relations. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 33–40, Barcelona, Spain.

Rudi Cilibrasi and Paul Vitanyi. 2007. The Google Similarity Distance. *IEEE Transactions on Knowledge and Data Engineering*, 19(3):370 – 383.

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)*, pages 80–87, Edmonton, Canada.

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

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, Nantes, France.

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)*, Gaithersburg, MD, USA.

Christian Kohlschutter, Peter Fankhauser, and Wolfgang Nejdl. 2010. Boilerplate Detection using Shallow Text Features. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*, pages 441–450, New York City, NY, USA.

Michael McCandless, Erik Hatcher, and Otis Gospodnetić. 2010. *Lucene in Action*. Manning Publications, 2nd edition.

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

Gordon Mohr, Michael Stack, Igor Ranitovic, Dan Avery, and Michele Kimpton. 2004. An Introduction to Heritrix, an open source archival quality web crawler. In *Proceedings of the International Web Archiving Workshop (IWAW)*.

Patrick Pantel and Marco Pennacchiotti. 2006. Expresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations. In *Proceedings of the International Conference on Computational Linguistics and Annual Meeting of the Association for Computational Linguistics (COLING/ACL)*, pages 113–120, Sydney, Australia.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49, Manchester, United Kingdom.

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)*, pages 801–808, Sydney, Australia.

Peter Turney and Michael Littman. 2003. Measuring Praise and Criticism: Inference of Semantic Orientation from Association. In *Proceedings of ACM Transactions on Information Systems (TOIS)*, pages 315–346.

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)*, pages 482–489, Borovets, Bulgaria.

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

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

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.

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

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