Knowledge Acquisition with Natural Language Processing in the Food Domain: Potential and Challenges

Michael Wiegand and **Benjamin Roth** and **Dietrich Klakow**[1]

Abstract. In this paper, we present an outlook on the effectiveness of natural language processing (NLP) in extracting knowledge for the food domain. We identify potential scenarios that we think are particularly suitable for NLP techniques. As a source for extracting knowledge we will highlight the benefits of textual content from social media. Typical methods that we think would be suitable will be discussed. We will also address potential problems and limits that the application of NLP methods may yield.

1 Introduction

Food plays an essential role in each of our lives. We do not only need it to survive but it has also significant social and cultural aspects. Within the last fifty years, research in artificial intelligence (AI) has brought immense achievements for human society with the result that, nowadays, AI technology is available in many parts of our life. Due to the importance of food in our society and the general applicability of AI methods, it is only a natural consequence that research in the area of AI has also addressed tasks in the food domain.

In this paper, we focus on one specific branch in artificial intelligence, namely natural language processing (NLP). NLP can be defined as the task of extracting meaningful content from natural language utterances. Research in artificial intelligence addressing food-related tasks up to now focused on human-computer interaction [2, 4, 11, 13, 14, 16, 17, 22], knowledge engineering [1, 7] and image/video processing [21, 28]. Since there has only been very little research examining the usefulness of NLP in tasks related to the food domain, we outline some directions of research that given the current state of the art we envisage to yield some potential. More precisely, we want to describe some scenarios in which NLP can be leveraged in order to extract *knowledge*. The resulting roadmap is the main contribution of this paper.

The basic task that all these scenarios involving NLP underlie is the conversion of some written natural language text, i.e. some unstructured data, to some structured text. For example, a text, such as Sentence 1, should be transformed to some relation (similar to a logic formula), such as Example 2.

1. I use shortcrust pastry for my apple pie.
2. *Ingredient-of(shortcrust pastry, apple pie)*

It is then the task of other disciplines, such as knowledge engineering, to incorporate these data into an information system that supports a user in their decision making. (This step will not be covered in this paper.) In this paper, we exclusively focus on knowledge extraction from written text. This work expands on our preliminary findings presented in [29] which describes empirical work of knowledge extraction in the food domain for German.

[1] Spoken Language Systems, Saarland University, Germany, email: Michael.Wiegand, Benjamin.Roth, Dietrich.Klakow@lsv.uni-saarland.de

2 The Main Purpose of Artificial Intelligence in the Food Domain

If one categorizes existing research in artificial intelligence dealing with food according to their purposes, one actually finds that most of them serve the same purpose. This research proposes technologies that attempt to fix some undesirable behavior. In [12], such methods are called *corrective technologies*. Much of previous work supports a user in cooking a meal [13, 14, 22, 25]. In these cases, the undesirable behavior can be considered the uncertainty or inexperience of how to prepare a meal. Another line of research deals with supporting people with health-related problems, e.g. following a specific diet [2, 4, 11, 17]. In these works, the undesirable behavior can be considered some improper diet. Some of previous work may not seem to address the fixing of an undesirable behavior, but in most of these cases they *indirectly* address this issue as some intermediate problem is solved. For example, the task of detecting how much food has been consumed from a plate [16], the task of analyzing drinking activity [28] or mastication [21] can be seen as intermediate steps that need to be dealt with in order to fully support humans in performing a diet.

We assume that a viable task in the food domain in which NLP can be applied should also address the fixing of some undesirable behavior. We even think that those two major scenarios presented above (i.e. *preparing a meal* and *following a health-related diet*) are actually ideal scenarios for applying NLP methods. We will motivate this in detail in Section 5.

3 Benefit of State-of-the-art NLP in General

With today's hardware capacity, a prominent advantage of NLP is that it can process text at a speed that significantly exceeds human performance and hence larger amounts of texts can be processed.

The type of information that can be extracted is usually restricted to content that can be detected with the help of some surface patterns. Surface patterns usually comprise lexical knowledge, but it may also include syntactic and semantic knowledge. We will illustrate this with an example. For a relation instance, such as *Can-be-Substituted-with(butter, margarine)*, there are many different ways of how this relation instance can be expressed in natural language text as exemplified by Sentences 3-5.

3. I use margarine instead of butter.
4. Butter is often substituted by margarine.
5. For the apple pie we used margarine; I forgot to buy butter at the supermarket.

Sentences 3 and 4 can be recognized with the help of NLP. This will be explained in more detail in Section 6.3.2. To infer this relation from Sentence 5, however, extra-linguistic (pragmatic) knowledge would be required as there are no explicit lexical cues indicating the given relation. The food items *margarine* and *butter* are in different clauses and there is no syntactic relationship between the words that could indicate some relation. Only by knowing that having forgotten butter at the supermarket is a justification of using margarine, one can infer that the speaker would normally have taken butter. From this we can conclude that butter and margarine can be exchanged with each other. This is some domain-specific knowledge that is extremely hard, if not impossible, to acquire. In other words, with state-of-the-art NLP it is not possible to fully comprehend an entire text. A deeper understanding of text can only be obtained if an ontology encoding word knowledge complements the linguistic analysis. Such endeavours only work for extremely closed-domain scenarios. Moreover, they are much less efficient.

4 Social Media as the Source for Natural Language Processing

When using NLP for a new domain, one also needs to answer what text source should be used for extracting content. Of course, not any arbitrary text source is applicable. In order to qualify as a source, the text type needs to meet the two following criteria:

Firstly, the text type needs to contain sufficient domain knowledge. In other words, if we choose a text type that only infrequently contains content regarding food, then we are not very likely to extract any significant amount of knowledge. In the past, most research in NLP has been carried out on news corpora [15]. The topic that is predominant on this text type are political affairs rather than food-related issues. Consequently, this text type would be of little value for knowledge extraction of food relations.

Secondly, the text type should not only contain knowledge about food that is already widely available in structured format (such as databases). Otherwise, there would hardly be any point in extracting knowledge from those texts as it would already be available.

Given these requirements, we argue that one particularly promising text type for the extraction of food-related knowledge are *social media*. By social media, one understands those types of media (not necessarily only textual data) that allow some interaction between the people who produce information and the people who consume it. Moreover, in social media the same person can assume both of these roles; a recipient of some information can be the producer of some other information in a different situation. The person who produces content can be any arbitrary *user*. This has led to the coinage of the term *user generated-content*.

The texts from the social-media domain that we are primarily interested in are *internet forums* and *weblogs*. Apart from the fact that large amounts of such texts are actually publicly available, e.g. they can be downloaded via web crawlers, there is also a significant proportion dealing with food-related issues. This is because food is a central issue in everyday life and, nowadays, almost every part of everyday life is reflected in social media in some way or the other. Furthermore, we assume that the kind of food-related information that can be found in social media is, to a large extent, complementary to what is found in existing resources. (In Section 6.2, we will give a typical example of the type of available resources that contains information regarding food items.) The reason for that is that existing resources contain (uncontroversial) factual content regarding food items. For instance, there are ontologies that arrange food items in a hyponymy (*is-a*) relationship. (Thus one can read off which food items are a subtype or supertype of another item.) Social media, on the other hand, also contain much subjective information. On the web, users may not only exchange recipes but also discuss which combination of food items they *think* is appropriate or which items can be used instead of each other. In addition to that, they may also exchange their *experience* with certain types of food, in particular, if they are on a diet or have certain health conditions (such as *allergies*, *diabetes* or *irritable bowel syndrome*).

In our first preliminary work on knowledge acquisition in the food domain [29], a crawl from an internet forum has successfully been used. Since that work is done on German, the largest German website dealing with food-related issues, namely *chefkoch.de*[2], has been crawled. The resulting data collection comprises $418,558$ webpages.

In the following section, we show that the information potentially contained in these data (described above) would also be extremely valuable for real-life scenarios.

5 Scenarios

One possible scenario that could make use of NLP and that also motivates our previous work in the food domain [29] is virtual customer advice. We will now describe this setting and highlight what benefits NLP would bring about in this task. Moreover, we will also outline possible extensions to this scenario. We will focus on this single scenario because it has many interesting facets that yield many possibilities of applying different NLP methods. Moreover, this scenario has an obvious commercial potential. Commercial viability is important for many new technologies to be developed, as it may foster cooperation between academia and industry.

The specific use case that is presented in [29] is assisting a customer in a supermarket in doing their shopping. Typical situations that could arrive are that

a) a customer wants to organize an event and does not know what food items or dishes are appropriate for that occasion;
b) a customer wants to prepare a meal but does not know what ingredients are necessary;
c) a customer wants to purchase a product that is currently out of stock and does not know what suitable substitutes there are;
d) a customer has a certain health disposition (e.g. may be suffering from diabetes) and does not know which products are most suitable for them.

All these cases are typical everyday life situations, all of which exhibit a user need that cannot be immediately satisfied by information that is available in a supermarket.[3] In principle, these problems could be solved by a large knowledge base containing relevant relations. For a), a relation table listing food items for diverse events would be required; for b), it would be a list of ingredients of different dishes; for c), it would be a table containing pairs of food items that can be substituted with each other; and for d), it would be a table listing food items recommended for people with a particular health disposition. Social media cover those everyday-life problems but, to a large extent, this information is only available in unstructured natural language text (e.g. as entries in an internet forum) rather than structured relation tables. Since we already stated in Section 3 that the speed of processing natural language text with NLP software can

[2] `www.chefkoch.de`

[3] Of course, a shopping assistant could be consulted but most supermarkets will not have sufficient human resources to assist every customer with their individual problems.

significantly exceed human performance, the choice of using NLP on extracting this knowledge from those weblogs or internet forums seems self-evident.

While [29] focuses on extracting structured knowledge, we also think that it would also be worthwhile providing entire (natural language) text passages in which particular relations have been found. The resulting applications may not be necessarily linked to the shopping scenario mentioned above, though. As already outlined in Section 2, health-related issues play a major role when it comes to the topic of food. Instead of building applications that just contain the knowledge of what types of food are recommendable for people with a certain disposition or the information about which food items are healthy or unhealthy, we assume that providing contextual information might be beneficial in several respects. Contextual information helps a user to understand how a system has arrived at some specific information. Thus, a user gains some trust in the knowledge-extraction system. In the ideal case, the context actually provides some explanation or justification for a specific claim. This additional information is in particular important if a claim that has been found is controversial or, at least, not immediately comprehensible. For instance, if a system extracts the knowledge *Is-Healthy(chocolate)*, a user would remain unsatisfied with this unexpected claim unless they are given some further background information as Sentence 5 does.

5. Chocolate is healthy because it's high in magnesium and provides vitamin E and vitamin B.

In particular, recent advances in shallow discourse processing might help to retrieve those passages which do not only contain a specific relation but also some justification [27] for it.

6 Methods

We will now describe a generic architecture which needs to be implemented in order to extract the type of knowledge from the food domain that we previously described. This architecture (illustrated in Figure 1) is a generalization of the system presented in [29].

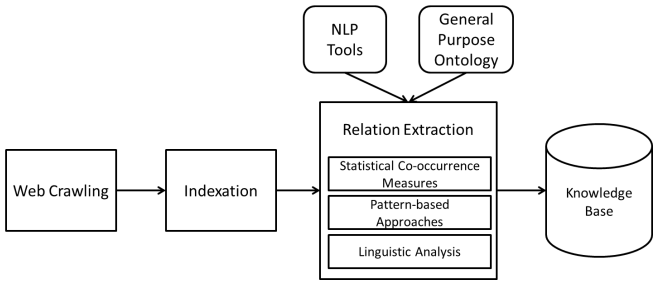

Figure 1. Generic architecture for knowledge acquisition using NLP.

6.1 Creating an Offline Index

In order to extract knowledge for the type of scenarios that we presented in Section 5, text processing needs to be carried out on large amounts of data, i.e. texts comprising several million words. Texts must first be assembled from the web. For such a purpose, publicly available web crawlers, such as *Heritrix* [23], can be used. Processing these text documents in a naive way (e.g. iterating through all files line by line) is not feasible as it would take too much time to complete the process. Imagine, a system is asked to find evidence for *Is-Healthy(broccoli)*. The first step would be finding passages (or sentences) in which the two linguistic entities *broccoli* and *healthy* co-occur. In order to obtain such text passages in real time, the text documents need to be converted into an *index data structure* that allows for efficient retrieval. For example, a widely used toolkit that carries out such a conversion and also enables the retrieval of text passages using that representation format is *Lucene* [18]. The algorithms that these tools implement are conceptually very similar to web search engines, such as *Google*, but these tools can be very flexibly tailored to a specific application. For example, one can determine how the index representing the data collection is going to be arranged. Moreover, much more sophisticated queries can be formulated in order to retrieve specific text passages.

6.2 Resources for Detecting Relevant Entities

Even though we want to extract knowledge from textual data, we also need some initial knowledge about our task domain. For instance, if we want to extract the knowledge of what types of food are usually consumed at a particular event, one needs to know the set of possible events and a list of all kinds of food. The most appropriate way to obtain such information is by incorporating general-purpose ontologies. For English, for example, one could make use of *WordNet* [19] which is a lexical database that lists semantic relations, such as *hyponymy* (*is-a* relation) or *meronymy* (*part-of* relation). These relations are not formulated between words but concepts which are groups of words with a similar meaning, i.e. *synonyms*. To obtain all words denoting food items one merely has to collect the words associated with the concepts that are hyponyms of *food*. The advantage of using such ontology instead of a mere list of food items is that it allows some inferences which might be useful for knowledge extraction. Imagine, for example, one is able to extract from text the knowledge *Suits-to(cheese, picnic)* (i.e. cheese is an appropriate type of food that can be consumed on a picnic). From this knowledge, one could also derive that this information also holds for a particular subtype of cheese, e.g. *cheddar*. Moreover, there can also be situations in which the knowledge of synonyms is beneficial. For instance, if the knowledge *Can-be-Substituted-by(zucchini, eggplant)* is extracted (i.e. zucchini is a good substitute for eggplant), and a user wants to obtain substitutes for *aubergine*, then the knowledge from an ontology that *aubergine* and *eggplant* refer to the same type of vegetable helps to infer that *Can-be-Substituted-by(zucchini, aubergine)*.

6.3 Relation Extraction

Once a text passage has been found in which two different target entities, for instance, the two food items *zucchini* and *aubergine* occur, one further needs to determine whether a particular relation holds between those items (e.g. *Can-be-Substituted-by(zucchini, aubergine)*). This is the task of *relation extraction*.

6.3.1 Statistical Co-occurrence

The simplest way to establish a relation is by measuring the statistical co-occurrence of entities between which there is potential relation. Imagine, for instance, we want to extract the food items that

are typically consumed at a given event, i.e. *Suits-to(FOOD-ITEM, EVENT)*. One possible way of extracting that knowledge is by measuring for each possible event which of the entire set of food items co-occurs with it. The more often two expressions co-occur with each other, the more likely there holds a specific relation between them. For example, *roast goose* will more often co-occur with *Christmas* than *banana* will co-occur with it, as *roast goose* is a dish typically consumed at *Christmas*. The strength of co-occurrence can be determined by applying standard measures, such as *point-wise mutual information* [6]. Statistical co-occurrence is particularly suitable for extracting relations which are difficult to grasp by means of textual patterns. For instance, in [29] it was found that lexical cues or phrases (see Section 6.3.2) to indicate the relation type *Suits-to* (e.g. cues of the form *X is usually consumed/eaten at/on (event) Y*) were less effective than co-occurrence measures. In particular, if the entities involved in a relation do not appear close to each other, a statistical co-occurrence method may be suitable.

The major shortcoming of this method is that it is completely oblivious of the context in which the entity pairs appear. This is, in particular, insufficient if there can be more than one relation holding between an entity pair. For example, if we applied this very method in order to extract relation instances of the type *Can-be-Substituted-by(FOOD-ITEM, FOOD-ITEM)* as in *Can-be-Substituted-by(fish fingers, fish cake)*, then this would mean that we would have to consider all potential pairs of food items. Unfortunately, a frequent co-occurrence between two food items does not necessarily mean that this particular relation type, i.e. *Can-be-Substituted-by*, holds between those items. This is because there could also be another relation holding between these two items, e.g. *Can-be-Served-with(FOOD-ITEM, FOOD-ITEM)* as in *Can-be-Served-with(fish fingers, mashed potatoes)*. With regard to an entity pair of type <*FOOD-ITEM*, *EVENT*>, there is actually only one likely relation type, namely *Suits-to*. Therefore, in order to decide whether there holds such a relation between a specific event and a specific food item, one just needs to measure the degree of co-occurrence. For extracting relations, such as *Can-be-Substituted-by(FOOD-ITEM, FOOD-ITEM)* or *Can-be-Served-with(FOOD-ITEM, FOOD-ITEM)*, on the other hand, more complex processing involving a context-based analysis is required.

6.3.2 Pattern-based Approaches

As motivated in the previous section, for some relation types, more context-aware methods, so-called *pattern-based* approaches, are necessary in order to extract instances from text.

One obvious solution to obtain such patterns is by manually writing them as it has been done in [29]. The advantage of this acquisition method is that it usually yields very precise patterns. The disadvantages are that the patterns are expensive to produce as they require expert knowledge and, moreover, tend to have a limited coverage. Of course, by considering levels of representation going beyond the mere lexical surface form (as in our preliminary work [29]) and using some linguistic resources, one could achieve some generalization. For example, consider the simple sequential surface pattern an expert may come up with, such as *replace X by Y*, in order to extract relation instances of type *Can-be-Substituted-by(FOOD-ITEM, FOOD-ITEM)*. This pattern would match Sentence 6. Sentences 7 and 8, on the other hand, would be missed.

6. You can replace butter by margarine.
7. Butter is often replaced by margarine.
8. Butter is often substituted by margarine.

With more sophisticated levels of representation that are available by state-of-the-art technology, these two sentences could also be matched. By using a pattern that does not only employ lexical information but also syntactic information, such as $X \uparrow$`logical-object` *replace* $\downarrow$`by-object` Y, Sentence 7 could also be matched. This pattern normalizes constructions, such as passive voice. The pattern says that the relation holds between *X* and *Y* if there is the verb *replace* and its logical object (this corresponds to the direct object in Sentence 6 but to the syntactic subject in Sentence 7 – both constituents are *butter*) is *X*, while its by-object is *Y* (this corresponds to *margarine*). In order to be able to match also Sentence 8 with a pattern, one would additionally need to know that *replace* and *substitute* are synonyms. General-purpose ontologies, such as *WordNet* [19], can provide such knowledge.

Another method to obtain patterns is to learn them from text. In order to do so, one needs labeled contexts, e.g. if patterns for relation type *Can-be-Substituted-by(FOOD-ITEM, FOOD-ITEM)* are to be learnt, sentences where instances of that relation are expressed have to be collected. A sufficiently large amount of such labeled data enables state-of-the-art supervised machine learning methods, such as *Support Vector Machines* [26], to be applied. A model produced by such a classifier is a weighted set of features which allows relation instances to be extracted. In principle, the features can be similar to the manually designed patterns. However, one typically uses a much larger set of features (patterns). One does not need to include exactly those features that are predictive. This is usually learnt by the classifier, i.e. highly weighted features roughly correspond to the predictive patterns. The features that are chosen as input for the learning algorithm can consequently be much more generic than manually designed surface patterns. Typical examples are words or word sequences between the arguments of a relation in a training sentence or the syntactic relationship between the arguments (as shown above). Since the feature space is usually fairly large, the resulting models that are learnt can be much more expressive than a set of manually defined surface patterns. In particular, the coverage of those models may be much higher.

The downside of this method is of course the time effort involved in labeling the data. A standard way of acquiring such data would be annotating large amounts of textual data, sentence by sentence, and mark the entities (in the text) between which there holds the target relation type, e.g. *Can-be-Substituted-by*. Fortunately, there are alternative methods that try to reduce that annotation effort. The class of methods commonly referred to as *distant supervision* [20] is a fairly recent methodology that could be applicable. It makes certain (simplifying) assumptions about the realization of relations that can drastically speed up the annotation process. Instead of annotating texts from scratch, one could, for instance, define a set of prototypical arguments of the target relation type, e.g. <*fish fingers*, *fish cake*> or <*margarine*, *butter*> for *Can-be-Substituted-by*, and then consider sentences in which those entities co-occur, for example Sentence 9, as positive training data.

9. I often use *margarine* instead of *butter*.

So, instead of *directly* labeling sentences, one just needs to formulate argument pairs. The remaining steps of this method can be done fully automatically. This annotation is much less time consuming since common argument pairs can have quite many matches within a large corpus. Moreover, the gold standard used in our preliminary work [29] introduced in [30] could be used for that very purpose.

Of course, there are limitations to this approach. One assumes that the co-occurrence of two entity pairs will always represent the target

relation type. However, in some sentence their occurrence could be co-incidental, such as the co-occurrence of *margarine* and *butter* in Sentence 10 (although if the chosen argument pairs are good prototypes, this situation will rarely occur and thus not critically mar the quality of the training data).

10. I just went to the supermarket to buy some *margarine*, *butter*, cheese, vegetables and potatoes.

6.3.3 Beyond Simple Patterns – Why further linguistic analysis might be helpful

Most pattern-based approaches focus on a propositional level of how a relation is expressed. However, we observed that for some relation types it is vital to consider the *embedding* of those propositions as it may discard the (general) validity of the proposition. For instance, consider the relation instance *Is-Healthy(beans)*. With a pattern-based approach, it would be easy to detect a typical occurrence, such as Sentence 11. Imagine, for example, that a pattern sequence *X BE*[4] *healthy* has been acquired. However, this sequence would also fire in Sentences 12-16 even though none of these sentences entails that statement.[5]

11. *Beans are healthy*.
12. I don't think that *beans are healthy*.
13. I really wonder whether *beans are healthy*.
14. My aunt claims that *beans are healthy*. (But this is wrong!)
15. *Beans are healthier* than chocolate.
16. It could be that *beans are healthy*.

Sentence 12 is negated, Sentence 13 is an indirect question, Sentence 14 reports somebody else's belief, Sentence 15 is a comparison, while in Sentence 16 there is a modal embedding. Some appropriate linguistic analysis (even with current state-of-the-art NLP technology) should be able to detect these types of embeddings. It involves common tasks, such as *negation detection* [3, 24] (Sentence 12) or *opinion holder extraction* [5] (Sentence 14) that are mostly dependent on lexical and syntactic information. This linguistic analysis could be used as an additional rule that undoes an erroneous detection of a relation by a pattern-based approach.

7 Difficulties and How They can be Solved

We already pointed out in Section 3 that for state-of-the-art NLP it is not possible to achieve a full textual understanding. In particular, relations that require some pragmatic knowledge cannot be extracted. In this section, we will not discuss the difficulties of NLP with regard to that particular problem but focus on difficulties that typically arise with those types of methods that we proposed in the previous sections. We believe that these difficulties are more imminent problems to the task and that they are also more likely to solved (at least partially) in the near future.

As stated previously in this paper, the text type we think is most suitable from which to extract knowledge regarding food is user-generated content from the web. Irrespective of the concrete task that is to be carried out on these data, the text type itself already entails a significant problem. User-generated content is typically not subject to any checking that the texts that are produced are suitable. As a consequence of that, texts may contain errors on various levels. Words may be misspelt, sentences may be ungrammatical, wording may be inaccurate or misleading, and even complete statements may be incomprehensible. Moreover, statements may be off-topic or offensive (e.g. flames). Of course, this has a negative impact on NLP methods as the largest part of research in NLP assumes that the texts contain no errors. If words are misspelt, they cannot be properly recognized. WordNet (Section 6.2), for example, cannot anticipate incorrect spelling since it only contains correctly spellt entries. In the previous section we stated that some methods to extract relations require some linguistic analysis. These analyses are typically produced by a parser. Not only ungrammatical sentences may negatively affect the analyses made by such parsers. Most automatic syntactic analyses require that all words of a sentence have been recognized and that both the wording and the syntactic constructions resemble those data on which the parser has been developed. In spite of deviations, a parser may produce an analysis but the analysis may be very wrong. As the majority of NLP tools are developed on regular (*tidy*) newswire texts, one has to expect a significant *domain mismatch* when using those tools on other text types.

Only until recently, the necessity of adapting common NLP tools to other domains, in particular noisy text types as can be found in social media, has been addressed in research. Already initial experiments on that task yield promising results [8, 9, 10]. What this line of research mostly attempts is capturing systematics behind misspelling words and training parsers on those sentences that are more representative of the target domain than traditional newswire texts. (Thus, to some extent, systematic ungrammaticality can be learnt from those data.) As this line of research is still in its infancy, up to now, there are no NLP tools publicly available that have been tuned to these special data.

As a consequence, the question, of course, may arise whether any research on social media is premature and should be carried out despite the lack of NLP tools that work sufficiently well on the user-generated content. Moreover, these adaptation efforts will only be able to solve some of those problems that are inherent to that domain. Inaccurate wording or incomprehensible statements will still remain a problem. Fortunately, not every sentence in user-generated texts contains these errors. After all, even with our preliminary work [29], we could show that some knowledge can be extracted. However, more research needs to be carried out in order to quantify the impact of those errors.

Irrespective of the technical problems that may occur during the automatic extraction of knowledge, one may also wonder how much knowledge is actually encoded in the data available. After all, the text corpora on which data are extracted can only be finite. Moreover, we just mentioned that in some way we rely on relations to be mentioned several times within our text collection. In the worst case, we would only be able to extract frequently occurring relations that are already common knowledge (e.g. relations of the type *Can-be-Substituted-by(butter, margarine)*). In other words, we would extract only that information that is not worth to be included in a specially built knowledge base since every ordinary person already has that knowledge. At this point in time, without some thorough empirical analysis, no definite judgement can be rendered. There is, however, one insight that may support the usefulness of the approach sketched in this paper, which is that social media are rapidly and steadily growing. A natural consequence of this is that the knowledge that can be extracted by state-of-the-art NLP methods may increase. So, a relation that cannot be extracted from textual data today because it is either not contained in those data or occurs too infrequently does not mean that it cannot be extracted from similar domain data in a few years' time. By then, there is much more text available that may al-

[4] By BE all inflectional forms of the verb *to be* are meant, i.e. *am*, *is*, *are*, *was*, *were* etc.

[5] We assume that Sentence 15 may also match as one usually normalizes word forms, so *healthier* would be reduced to the positive *healthy*.

low automatic NLP techniques to successfully extract this relation.

8 Conclusion

In this paper, we presented an outlook on the effectiveness of NLP in applications in the food domain. We identified two potential scenarios, namely advice on preparing meals and health-related issues, that predominate research in the food domain with regard to artificial intelligence and found that these scenarios are also quite suitable for NLP techniques. As a source for extracting knowledge we outlined the benefits of social media. Different extraction methods, ranging from co-occurrence measures to more complex linguistic analyses, were discussed. Finally, we also addressed potential problems that NLP methods may cause on the tasks we have proposed.

Acknowledgements

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

REFERENCES

[1] Fadi Badra, Rokia Bendaoud, Rim Bentebibel, Pierre-Antoine Champin, Julien Cojan, Amélie Cordier, Sylvie Després, Stéphanie Jean-Daubias, Jean Lieber, Thomas Meilender, Alain Mille, Emmanuel Nauer, Amedeo Napoli, and Yannick Toussaint, ‘TAAABLE: Text Mining, Ontology Engineering, and Hierarchical Classification for Textual Case-Based Cooking’, in *European Conference on Case-Based Reasoning - ECCBR 2008, Workshop Proceedings*, (2008).

[2] Brandon Brown, Marshini Chetty, Andrea Grimes, and Ellie Harmon, ‘Reflecting on health: A system for students to monitor diet and exercise’, in *Proceedings of CHI Extended Abstracts*, (2006).

[3] Wendy Webber Chapman, Will Bridewell, Paul Hanbury, Gregory F. Cooper, and Bruce G. Buchanan, ‘A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries’, *Journal of Biomedical Informatics*, **34**, 301 – 310, (2001).

[4] Pei-Yu Chi, Jen-Hao Chen, Hao-Hua Chu, and Bing-Yu Chen, ‘Enabling Nutrition-Aware Cooking in a Smart Kitchen’, in *Proceedings of CHI Extended Abstracts*, (2007).

[5] Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan, ‘Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns’, in *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP)*, Vancouver, BC, Canada, (2005).

[6] Kenneth Ward Church and Patrick Hanks, ‘Word Association Norms, Mutual Information, and Lexicography’, *Computational Linguistics*, **16**(1), 22–29, (1990).

[7] Amélie Cordier, Jean Lieber, Pascal Molli, Emmanuel Nauer, Hala Skaf-Molli, and Yannick Toussaint, ‘Knowledge Acquisition and Discovery for the Textual Case-Based Cooking system WIKITAABLE’, in *International Conference on Case-Based Reasoning - ICCBR 2009, Workshop Proceedings*, (2009).

[8] Jennifer Foster, “‘cba to check the spelling”: Investigating Parser Performance on Discussion Forum Posts’, in *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*, (2010).

[9] Jennifer Foster, Ozlem Cetinoglu, Joachim Wagner, Joseph Le Roux, Stephen Hogan, Joakim Nivre, Deirdre Hogan, and Josef van Genabith, ‘#hardtoparse: POS Tagging and Parsing the Twitterverse’, in *Proceedings of AAAI Workshop on Analysing Microtext*, (2011).

[10] Jennifer Foster, Ozlem Cetinoglu, Joachim Wagner, Joseph Le Roux, Joakim Nivre, Deirdre Hogan, and Josef van Genabith, ‘From News to Comment: Resources and Benchmarks for Parsing the Language of Web 2.0’, in *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*, (2011).

[11] Jeana Frost and Brian K. Smith, ‘Visualizing Health: Imagery in Diabetes Education’, in *Proceedings of Designing for User Experiences (DUX)*, (2003).

[12] Andrea Grimes and Richard Harper, ‘Celebratory Technology: New Directions for Food Research in HCI’, in *Proceedings of the Annual SIGCHI Conference on Human Factors in Computing Systems (CHI)*, (2008).

[13] Reiko Hamada, Jun Okabe, Ichiro Ide, Shin’ichi Satoh, Shuichi Sakai, and Hidehiko Tanaka, ‘Cooking Navi: Assistant for Daily Cooking in Kitchen’, in *Proceedings of the annual ACM international conference on Multimedia (MULTIMEDIA)*, (2005).

[14] Ichiro Ide, Yuka Shidochi, Yuichi Nakamura, Daisuke Deguchi, Tomokazu Takahashi, and Hiroshi Murase, ‘Multimedia Supplementation to a Cooking Recipe Text for Facilitating Its Understanding to Inexperienced Users’, in *Proceedings of the Workshop on Multimedia for Cooking and Eating Activities (CEA)*, (2010).

[15] Nancy Ide and Catherine Macleod, ‘The American National Corpus: A Standardized Resource of American English’, in *Proceedings of Corpus Linguistics*, (2001).

[16] Nitin Khann, Carol J. Boushey, Deborah Kerr, Martin Okos, David S. Ebert, and Edward J. Delp, ‘An Overview of The Technology Assisted Dietary Assessment Project at Purdue University’, in *Proceedings of the Workshop on Multimedia for Cooking and Eating Activities (CEA)*, (2010).

[17] Jennifer Mankoff, Gary Hsieh, Ho Chak Hung, Sharon Lee, and Elizabeth Nitao, ‘Using Low-Cost Sensing to Support Nutritional Awareness’, in *Proceedings of Ubicomp*, (2002).

[18] Michael McCandless, Erik Hatcher, and Otis Gospodnetić, *Lucene in Action*, Manning Publications, 2nd edn., 2010.

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

[20] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky, ‘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, (2009).

[21] Kenzaburo Miyawaki, Satoshi Nishiguchi, and Mutsuo Sano, ‘Extraction of Mastication in Diet Based on Facial Deformation Pattern Descriptor’, in *Proceedings of the Workshop on Multimedia for Cooking and Eating Activities (CEA)*, (2010).

[22] Kenzaburo Miyawaki and Mutsuo Sano, ‘A Cooking Support System for People with Higher Brain Dysfunction’, in *Proceedings of the ACM Multimedia 2009 Workshop on Multimedia for Cooking and Eating Activities (CEA)*, (2009).

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

[24] Pradeep G. Mutalik, Aniruddha Deshpande, and Prakash M. Nadkarni, ‘Use of General-purpose Negation Detection to Augment Concept Indexing of Medical Documents – A Quantitative Study Using the UMLS’, *Journal of the American Medical Informatics Association*, **8**(6), 598 – 609, (2001).

[25] Yuka Shidochi, Tomokazu Takahashi, Ichiro Ide, and Hiroshi Murase, ‘Finding Replacable Materials in Cooking Recipe Texts Considering Characteristic Cooking Actions’, in *Proceedings of the ACM Multimedia 2009 Workshop on Multimedia for Cooking and Eating Activities (CEA)*, (2009).

[26] Vladimir N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, 1995.

[27] Suzan Verberne, Lou Boves, Nelleke Oostdijk, and Peter-Arno Coppen, ‘What is not in the Bag of Words for Why-QA?’, *Computational Linguistics*, **36**(2), 229 – 245, (2010).

[28] Qing Wang and Kiduk Jie Yang, ‘Drinking Activity Analysis from Fast Food Eating Video Using Gernative Models’, in *Proceedings of the ACM Multimedia 2009 Workshop on Multimedia for Cooking and Eating Activities (CEA)*, (2009).

[29] Michael Wiegand, Benjamin Roth, and Dietrich Klakow, ‘Web-based Relation Extraction for the Food Domain’, in *Proceeding of the International Conference on Applications of Natural Language Processing to Information Systems (NLDB)*. Springer-Verlag, (2012).

[30] Michael Wiegand, Benjamin Roth, Eva Lasarcyk, Stephanie Köser, and Dietrich Klakow, ‘A Gold Standard for Relation Extraction in the Food Domain’, in *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, (2012).