Extracting recipe ingredients from cookbooks

by
Torsten Knauf

A thesis presented for the degree of
Master of Science

Research Group for Communication Systems
at
Faculty of Engineering
Christian-Albrechts-Universität zu Kiel
Germany
31.03.2017

Supervisor: Prof. Dr.-Ing.Norbert Luttenberger
Dr.-Ing. Jesper Zedlitz

Abstract

Always do this one last, when knowing the things to praise yourself for :P

Acknowledgements

If I don't profit from nice people in these thesis, I have done something horrible wrong. So try to remember most of them here at the end... :)

Contents

Chapter 1

Introduction

A recipe parser, which can tag old and rather unstructured cookbooks according to the ontology of (Schema.org 2016), is developed in this thesis. Once it is tagged, it is easy to extract the tagged entities. The parser is developed and tested with *Davidis, Henriette: Praktisches Kochbuch für die gewöhnliche und feinere Küche. 4. Aufl. Bielefeld, 1849* but can be adapted easily to every German cookbook. For other languages new training data and dictionaries for the machine learning preparation of the parser have to be provided, but the general algorithm can be inherited.

This effort is motivated by nutritional science. Being able to extract the ingredients of a recipe automatically simplifies research according healthy food as well as historical analysis.

The thesis is structured as follows:

Chapter 2

Making a cookbook machine-readable

This chapter covers shortly how we transform a cookbook in a machine-readable XML file, which can be processed further arbitrarily. To achieve this, the cookbook has to be digitalized first and afterwards enriched through an ontology.

2.1 Digitalisation

In general there are two different ways, how to digitalize a book. The first one is to scan each side and let an *optical character recognition*-program translate the scanned pictures into text. The second one is to *type it manual* into a computer.

The German Text Archive provides a collection of German texts from 16th to 19th century including *Davidis, Henriette: Praktisches Kochbuch für die gewöhnliche und feinere Küche. 4. Aufl. Bielefeld, 1849* in (Deutsches Textarchiv 2016). They digitalized it through double keying, meaning that two people manual typed the book into the computer. Differences in their versions were revised by a third person.

The digitalized version can be downloaded from (Deutsches Textarchiv 2016). It is already enriched through *TEI: Text Encoding Initiative*-standard[1]. TEI is a standard for representing text in digital form, which should be analysed later in humanities, social sciences or linguistics.

Figure XXX depicts the digitalized version of our cookbook.

Because we are only interested in extracting certain data from the recipes and not in linguistic analysis or something else, we transformed the digitalized version as depicted in figure XXX. The essence of this version is, that every recipe can be easily transformed into this form and that every recipe can be easily read by human for manual tagging.

2.2 CueML ontology

For general automatic extraction and further processing of information an ontology is needed. In computer science an ontology is a vocabulary with defined meaning. A very nice description for the use of ontologies, which are a requirement for the Semantic Web, can be found in (Berners-Lee et al. 2000).

[1] http://www.tei-c.org/index.xml

(Schema.org 2016) is an existing ontology for recipes. Figure XXX depicts its usage. It works fine for well structured recipes, which already have a list of used ingredients. However it is not well suited for tagging our cookbook as the following example demonstrates.

Therefore we invented *cueML*...

Now this format can be easily transformed into a well structured recipe and afterwards be tagged with the standard ontology from (Schema.org 2016)

2.3 Need for automatisation

(Erdmann et al. 2000) error prone. Brauche so lange pro Rezept

Chapter 3

Related Work

In general there exists many effort about extracting useful information from textual and unstructured resources. The superordinate term for this field of research is *Text Mining*. It was first mentioned in (Feldman and Dagan 1995) and on overview can be found in (Hotho et al. 2005).

The algorithms for extracting useful information depend highly on existing semi-structures, which can be taken advantage of. Here we present existing algorithm, which we found in the domain of cooking, and distinct their effort from this thesis.

3.1 Skip The Pizza

(http://skipthepizza.com/ 2012) is a project described on WordPress.org. The author wants to combine his two hobbies cooking and software engineering. For being able to answer questions like "How many ingredients does a typical recipe consist of?" or "Which are the most frequent ingredients?", he extracts the ingredients of recipes from http://recipes.wikia.com/wiki/Recipes_Wiki.

Listing 3.1: Shortened example recipe from http://recipes.wikia.com/wiki/Recipes_Wiki

```
    * Makes 6 to 8 servings

== Ingredients ==
* 2 tbsp extra virgin [[olive oil]]
* 3 cloves [[garlic]], finely chopped
[...]

== Directions ==
Heat olive oil and garlic in large skillet over low heat until
garlic begins to sizzle.
Add tomatoes, [...]

[[Category:Cathy's Recipes]]
[[Category:Garlic Recipes]]
[...]
```

The recipes have the internal structure of listing 3.1. The semi-structure, that after `== Ingredients ==` comes a list of ingredients, can be recognized easily. Per line is one ingredient enclosed within `[[ingredient name]]`. Using this semi-structure

a regular expression is already good enough for extracting the ingredients from these recipes.

3.2 Domain Specific Information Extraction for Semantic Annotation

(Ahmed 2009) is a diploma thesis about extracting ingredients and their further processing from recipes. Their algorithm could be divided into two main parts.

One main part is to check every word if it is occurs in a dictionary of ingredients respectively a dictionary of actions and tag it accordingly. For keeping the dictionaries as small as possible they do a morphological Analysis and only store the lemmas of the words. The second main part, and more sophisticated task, is to identify which action should be applied to which ingredient. They have two different approaches for that.

The first one is to do a part of speech tagging and afterwards trying to apply a small set of rules. The example rule 1 means apply the action to all following ingredients and gets matched. Therefore it is extracted that buttermilk and bananas should be extracted.

Listing 3.2: Rule based example

```
add buttermilk and bananas
->
add[ACT] buttermilk[ING] and[CC] bananas[ING]
->
rule 1: VP -> ACT NP (,NP)* (CC NP)?
   -> ACT NP CC NP                          und Hilfsregel:  NP -> DT? JJ* ING
   -> ACT ING CC ING
```

The second one is to do a dependency based parsing, which represent the semantic structure of a sentence in a tree like format.

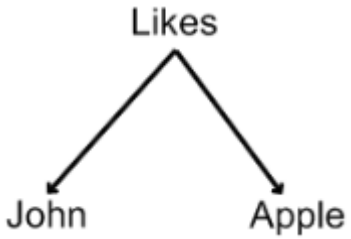

Figure 3.1: John likes Apple (Ahmed 2009)

For example fig. 3.1 represents, that the subject of like is John and the liked object is apple. The format as well as building the tree is way more complex than the previous simple rules, but the tree can be build by already existing tools like the Standfod Parser[1]. Having the semantic structure of the sentences it is trivial to extract which action should be applied to which ingredient.

[1] http://nlp.stanford.edu/software/lex-parser.shtml

3.3 Extracting Structured Data From Recipes Using Conditional Random Fields

3.4 Distinction to this work

Chapter 4

Development of recipe parser

4.1 Overall picture

4.1.1 workflow

4.1.2 preparation

4.1.3 evaluation

-Precision -Recall F-measure

4.2 First Iteration: Basis CRF

Git-tag

4.2.1 Idea

4.2.2 Evaluation

4.3 Final recipe parser

4.3.1 Workflow

4.3.2 Evaluation

Chapter 5

Discussion

- Übertragung auf beliebige Bücher (Wenn Buch hat bereits Zutatenliste erste Phase entfällt) - TDD / früh Plausibilitäts-Überprüfungen (insbesondere fürs Schemata, Mengenangaben historische Recherche nötig)

5.1 Power of machine readable data

Machine readable data are very powerful. But *with great power comes great responsibility*[1]. In the context of a recipe parser this might be a little bit exaggerated. But specially in mind of the global surveillance disclosures denounced by Edward Snowed with still uncertain dimension, I think it is important to have a consciousness for what can be done. Therefore I want to think about, what can be done through innocent looking machine readable tags.

For the good there exists already much effort for services, which require being able to extract ingredient from recipes.

(e.g. Teng et al. 2012 or Ueda et al. 2011).

Further more, having a huge machine-readable base of recipes and its ingredients, can also provide insights in sociological research. For example in (((Flavor network and the principles of food pairing : Scientific Reports))) is a comparison between American and Asian kitchen based on about 56.000 recipes.

There are many more interesting questions, which could be analysed like a historic analysis of the development and changes of cooking. Occurrences of non-local ingredients or meals are evidence for inter cultural exchange and globalisation. More expensive ingredients could be an indication for prosperity, while very simple kitchen for poverty or even wartimes...

In the bad (The Washington Post 2016) "schwacher vegetarier"

[1] A well known proverb which probably has its origin from the French National Convention during the period of French Revolution. (*With Great Power Comes Great Responsibility* 2015)

Chapter 6

Summary

Appendix A

Statutory Declaration

I declare that I have developed and written the enclosed Master Thesis completely by myself, and have not used sources or means without declaration in the text. Any thoughts from others or literal quotations are clearly marked. The Master Thesis was not used in the same or in a similar version to achieve an academic grading or is being published elsewhere.

Location, Date Signature

Appendix B

Something else

hi

Appendix C

Something else else

hello

Bibliography

Ahmed, Zeeshan (2009). "Domain Specific Information Extraction for Semantic Annotation". Diploma thesis. Charles University in Prague, Czech Republic and University of Nancy 2 in Nancy, France.

Berners-Lee, T., J. Hendler, and O. Lassila (2000). "Semantic web". In: *Scientific American, 1(1):68–88.*

Deutsches Textarchiv, Berlin-Brandenburgische Akademie der Wissenschaften (2016). *Davidis, Henriette: Praktisches Kochbuch für die gewöhnliche und feinere Küche. 4. Aufl. Bielefeld, 1849.* URL: http://www.deutschestextarchiv.de/book/show/davidis_kochbuch_1849 (visited on 10/30/2016).

Erdmann, M. et al. (2000). "From Manual to Semi-automatic Semantic Annotation: About Ontology-based Text Annotation Tools". In: *Scientific American, 1(1):68–88.*

Feldman, Ronen and Ido Dagan (1995). "Knowledge Discovery in Textual Databases (KDT)". In: *Proceedings of the First International Conference on Knowledge Discovery and Data Mining.* KDD'95. Montréal, Québec, Canada: AAAI Press, pp. 112–117. URL: http://dl.acm.org/citation.cfm?id=3001335.3001354.

Hotho, Andreas, Andreas Nürnberger, and Gerhard Paaß (2005). "A brief survey of text mining". In: *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology.*

http://skipthepizza.com/ (2012). *Skip The Pizza.* URL: http://skipthepizza.com/blog/analyzing-the-ingredients-of-29200-recipes (visited on 10/28/2016).

Schema.org (2016). *Schema.org/Recipe.* URL: https://schema.org/Recipe (visited on 10/30/2016).

Teng, Chun-Yuen, Yu-Ru Lin, and Lada A. Adamic (2012). "Recipe Recommendation Using Ingredient Networks". In: *Proceedings of the 4th Annual ACM Web Science Conference.* WebSci '12. New York, NY, USA: ACM, pp. 298–307. ISBN: 978-1-4503-1228-8. DOI: 10.1145/2380718.2380757. URL: http://doi.acm.org/10.1145/2380718.2380757.

The Washington Post (2016). *Hillary Clinton's health just became a real issue in the presidential campaign.* URL: https://www.washingtonpost.com/news/the-fix/wp/2016/09/11/hillary-clintons-health-just-became-a-real-issue-in-the-presidential-campaign/ (visited on 11/02/2016).

Ueda, Mayumi, Mari Takahata, and Shinsuke Nakajima (2011). "User's Food Preference Extraction for Personalized Cooking Recipe Recommendation". In: *Proceedings of the Second International Conference on Semantic Personalized Information Management: Retrieval and Recommendation - Volume 781.* SPIM'11. Aachen, Germany, Germany: CEUR-WS.org, pp. 98–105. URL: http://dl.acm.org/citation.cfm?id=2887675.2887686.

With Great Power Comes Great Responsibility (2015). URL: http://quoteinvestigator.com/2015/07/23/great-power/ (visited on 11/02/2016).