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

http://skipthepizza.com/ 2016

3.1 Masterarbeit

3.2 CRF

Chapter 4

Development of recipe parser

4.1 Test set-up

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) - Power of Strucutred Data / Semantic web - soziologische Forschung (Globalisation, Wohlstand, Kriegsküche) - Was wir über uns preisgeben (Hilary's Gesundheitsstatus wichtiger Faktor im US-Wahlkampf) - 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 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)

[1] A well known proverb with uncertain origin (*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

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*.

http://skipthepizza.com/ (2016). *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).