Extracting recipe ingredients from cookbooks

Erste Schritte

von

Torsten Knauf

Der Beginn einer Master-Arbeit

Tag-Schema

```
Mom's World Famous Banana Bread
By John Smith, May 8, 2009
<img src="bananabread.jpg" alt="Banana bread on a plate" />
This classic banana bread recipe comes from my mom -- the
walnuts add a nice texture and flavor to the banana bread.
Prep Time: 15 minutes
Cook time: 1 hour
Yield: 1 loaf
Tags: Low fat
Nutrition facts:
240 calories, 9 grams fat
Ingredients:
- 3 or 4 ripe bananas, smashed
- 1 egg
- 3/4 cup of sugar
...
Instructions:
Preheat the oven to 350 degrees. Mix in the ingredients in
a bowl. Add the flour last. Pour the mixture into a loaf
pan and bake for one hour.
140 comments:
From Janel, May 5 -- thank you, great recipe!
...
```

(Schema.org 2016)

```
<div itemscope itemtype="http://schema.org/Recipe">
  <span itemprop="name">Mom's World Famous Banana Bread</span>
  By <span itemprop="author">John Smith</span>,
  <meta itemprop="datePublished" content="2009-05-08">May 8, 2009
  <img itemprop="image" src="bananabread.jpg"
       alt="Banana bread on a plate" />
  <span itemprop="description">This classic banana bread recipe comes
  from my mom -- the walnuts add a nice texture and flavor to the banana
  bread.</span>
  Prep Time: <meta itemprop="prepTime" content="PT15M">15 minutes
  Cook time: <meta itemprop="cookTime" content="PT1H">1 hour
  Yield: <span itemprop="recipeYield">1 loaf</span>
  Tags: <link itemprop="suitableForDiet" href="http://schema.org/LowFatDiet" />Low fat
  <div itemprop="nutrition"
    itemscope itemtype="http://schema.org/NutritionInformation">
    Nutrition facts:
    <span itemprop="calories">240 calories</span>,
    <span itemprop="fatContent">9 grams fat</span>
  </div>
  Ingredients:
  - <span itemprop="recipeIngredient">3 or 4 ripe bananas, smashed</span>
  - <span itemprop="recipeIngredient">1 egg</span>
  - <span itemprop="recipeIngredient">3/4 cup of sugar</span>
  ...
  Instructions:
  <span itemprop="recipeInstructions">
  Preheat the oven to 350 degrees. Mix in the ingredients in a bowl. Add
  the flour last. Pour the mixture into a loaf pan and bake for one hour.
  </span>
  140 comments:
  <div itemprop="interactionStatistic" itemscope itemtype="http://schema.org/InteractionCounter">
    <meta itemprop="interactionType" content="http://schema.org/CommentAction" />
    <meta itemprop="userInteractionCount" content="140" />
  </div>
  From Janel, May 5 -- thank you, great recipe!
  ...
</div>
```

Pfannekuchen (Rezept mit Bild) von 06onkel | Chefkoch.de

www.chefkoch.de › Rezepte › Kategorien › Menüart › Frühstück ▾

★★★★½ Bewertung: 4,3 - 93 Rezensionen - 30 Min.

20.07.2005 - **Pfannekuchen**, ein gutes Rezept mit Bild aus der Kategorie Frühstück. 93 Bewertungen: Ø 4,3. Tags: Braten, Camping, einfach, Frühstück, ...

Pfannekuchen Rezepte - kochbar.de

www.kochbar.de › Rezepte ▾

Die besten **Pfannekuchen** Rezepte - **Pfannekuchen** Rezepte und viele weitere beliebte Kochrezepte finden Sie bei kochbar.de.

Pfannekuchen Muttis Rezept :) - Rezept mit Bild - kochbar.de

www.kochbar.de › Rezept ▾

★★★★★ Bewertung: 4,8 - 16 Abstimmungsergebnisse - 30 Min. - Kalorien: 171

03.10.2008 - Das perfekte **Pfannekuchen** Muttis Rezept :)-Rezept mit Bild und einfacher Schritt-für-Schritt-Anleitung: Butter, Vanille Zucker und Zucker ...

```
<script type="application/ld+json">
{
 "@context": "http://schema.org",
 "@type": "Recipe",
   "datePublished": "2005-07-20",
 "description": "Pfannekuchen, ein gutes Rezept mit Bild aus der Kategorie Fr&amp;uuml;hst&amp;uuml;
 "image": "http://static.chefkoch-cdn.de/ck.de/rezepte/36/36386/875411-960x720-pfannekuchen.jpg",
 "recipeIngredient": [
                 "250 g Mehl ",                          "4  Ei(er) ",                          "  Salz
  "name": "Pfannekuchen",
    "prepTime": "PT30M",
  "recipeInstructions": "Den ganzen Teig mit dem Mixer zubereiten (Achtung es spritzt), so dass es ei
Eine Pfanne mit etwas Öl heiß werden lassen und dann eine Suppenkelle Teig in die Pfanne geben, dabei

Dazu passen süße sowie auch herzhafte Sachen, z. B. Nutella, Leberwurst, Konfitüren, Lachs usw.",
  "recipeYield": "4"
  ,
  "aggregateRating": {
    "@type": "AggregateRating",
    "ratingValue": "4.27",
    "reviewCount": "93",
    "worstRating": 0,
    "bestRating": 5
  }
</script>
<script type="text/javascript">
    var _simplora_params = { recipes: [] };
    _simplora_params.recipes.push({
        id: '363861121870267',
        name: 'Pfannekuchen',
        servings: 4,
        ingredients: [
                        { id: '71', base_amount: '62.5', amount_unit: 'g', name: 'Mehl'},
                        { id: '36', base_amount: '1', amount_unit: '', name: 'Ei(er)'},
                        { id: '6', base_amount: '0', amount_unit: '', name: 'Salz'},
```

```
Ingredients:
- <span itemprop="recipeIngredient" quantity="3-4" name="banana">3 or 4 ripe bananas, smashed</span>
- <span itemprop="recipeIngredient" quantity="1" name="egg">1 egg</span>
- <span itemprop="recipeIngredient" quantity="3/4" unit="cup" name="sugar">3/4 cup of sugar</span>
...
Instructions:
```

```
<recipe type="Suppen." rcp-id="B-16">
  <head>Mock Turtle Suppe.</head>

  <p>Es wird hierzu für 24-30 Personen eine kräftige Bouillon von 8-10 Pfund
       Rindfleisch mit Wurzelwerk gekocht. Zugleich bringt man einen großen
       Kalbskopf, eine Schweineschnauze und Ohren, einen Ochsengaumen und eine
        geräucherte Ochsenzunge zu Feuer und kocht dies Alles gahr, aber nicht
        zu weich. Kalt, schneidet man es in kleine, länglich viereckige Stü
       ckchen, gibt das Fleisch in die Bouillon, nebst braunem Gewürz, ein
       Paar Messerspitzen Cayenne-Pfeffer, einige Kalbsmidder in Stückchen
       geschnitten (siehe Vorbereitungsregeln), kleine Saucissen, so viel
       Kalbskopfbrühe, daß man hinreichend Suppe hat, und macht dies mit in
       Butter braun gemachtem Mehl gebunden. Nachdem dies Alles 1/4 Stunde
       gekocht hat, kommen noch Klöße von Kalbfleisch, einige hart gekochte
       Eier in Würfel geschnitten, ein Paar Eßlöffel Engl. Soja hinzu, und
       wenn die Klößchen einige Minuten gekocht haben, 1/2 Flasche Madeira und
        auch Austern, wenn man sie haben kann. Dann wird die Suppe sogleich
       angerichtet.</p>

  <note>Anmerk. Der Soja macht die Suppe gewürzreicher, kann jedoch gut
       wegbleiben, und statt Madeira kann man weißen Franzwein und etwas Rum
       nehmen. Sowohl die Bouillon als Kalbskopf können schon am
       vorhergehenden Tage, ohne Nachtheil der Suppe, gekocht werden. </note>
</cue:recipe>
```

Listing 1: Beispiel Rezept von Frau Davidis

```
<div itemscope itemtype="http://schema.org/Recipe http://cueML.org">
  <recipe type="Suppen." rcp-id="B-16">
    <head><span itemprop="name">Mock Turtle Suppe</span></head>

    <meta>
      <span itemprop="recipeYield" quantity="24-30" unit="Personen">24-30
           Personen</span>
      <span itemprop="recipeIngredient" name="Bouillon" reference="#Bouillon">
           Bouillon</span>
      <span itemprop="recipeIngredient" name="Rindfleisch" quantity="8-10" unit=
           "Pfund">8-10 Pfund Rindfleisch</span>
      <span itemprop="recipeIngredient">Wurzelwerk</span>
      <span itemprop="recipeIngredient" name="Cayennepfeffer" quantity="vague"
           unit="Messersptize">ein Paar Messerspitzen Cayenne-Pfeffer</span>
      <span itemprop="recipeIngredient" name="Ei" quantity="vague">einige hart
           gekochte Eier</span>
      ...
      <span itemprop="recipeIngredient" name="Auster" isOptional="True">Austern,
           wenn man sie haben kann</span>
      <span itemprop="recipeIngredient" quantity="vague" unit="EL" isOptional="
           True">ein Paar Eßlöffel Engl. Soja</span>
      <recipeIngredientAlternations>
        <alt>
          <span itemprop="recipeIngredient" name="Madeira" quantity="0.5" unit="
               Flasche">Madeira</span>
        </alt>
        <alt>
          <span itemprop="recipeIngredient" name="weißen Franzwein">weißen
               Franzwein</span>
          <span itemprop="recipeIngredient" name="Rum" quantity="vague">etwas
               Rum</span>
        </alt>
      </recipeIngredientAlternations>
    </meta>
```

```
<div itemscope itemtype="http://cueML.schema.org/Recipe">
  <recipe type="Suppen." rcp-id="B-16">
    <meta>
      <span itemprop="name" content="Mock␣Turtle␣Suppe"/>

      <span itemprop="recipeYield" quantity="24-30" unit="Personen"/>

      <span itemprop="recipeIngredient" name="Bouillon" reference="#Bouillon"/>
      <span itemprop="recipeIngredient" name="Rindfleisch" quantity="8-10" unit=
          "Pfund"/>
      <span itemprop="recipeIngredient" uncertainName="Wurzelwerk"/>
      <span itemprop="recipeIngredient" name="Cayennepfeffer" quantity="vague"
          unit="Messersptize"/>
      <span itemprop="recipeIngredient" name="Ei" quantity="vague"/>
      ...
      <span itemprop="recipeIngredient" name="Auster" isOptional="True"/>
      <span itemprop="recipeIngredient" uncertainName="Engl.␣Soja" quantity="
          vague" unit="EL" isOptional="True"/>
      <recipeIngredientAlternations>
        <alt>
          <span itemprop="recipeIngredient" name="Madeira" quantity="0.5" unit="
              Flasche"/>
        </alt>
        <alt>
          <span itemprop="recipeIngredient" name="weißen␣Franzwein"/>
          <span itemprop="recipeIngredient" name="Rum" quantity="vague"/>
        </alt>
      </recipeIngredientAlternations>
    </meta>

    <span itemprop="recipeInstructions">...</span>
  <recipe>
</div>
```

- Das Tag-Schema ist nicht trivial, da 3 unterschiedliche Themenbereiche:
  - TEI
  - Schema.org/Recipe
  - Information Extraction

SoP (Separation of Concerns)

Algorithmen um Informationen zu extrahieren

(*Skip The Pizza* 2012)

```
* 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:Garlic Recipes]]
[...]
```

Linear-chain Conditional Random Field

(Sutton and McCallum 2012), (The New York Times 2015)

Q=Quantity, U=Unit, I=Ingredient, P=Punctuation, OPT=Optional, N=Negation, O=Other

```
2 EL Zucker                              Q U I
Einen Schuss Zucker                      Q U I
Zucker, wenn nicht süß genug             I P OPT N O O
```

Listing 4: training data

- **Model** $p(X, Y) = \prod_{t=1}^{T} p(y_t|y_{t-1}) * p(x_t|y_t)$
- Trainings data + Magie mit Model

$$\Downarrow$$

```
2 TL Zucker                              Q U I
```

Listing 5: Labeling

Model $p(X, Y) = \prod_{t=1}^{T} p(y_t|y_{t-1}) * p(x_t|y_t)$

$$p(X, Y) = \frac{1}{Z(X, Y)} \prod_{t=1}^{T} exp(\sum_{i,j \in S} \Theta_{i,j} * 1_{y_t=i} * 1_{y_{t-1}=j} + \sum_{i \in S} \sum_{j \in O} \mu_{o,i} * 1_{y=i} * 1_{x_t=o})$$

$Z(X,Y) = \sum_X \sum_Y \prod_{t=1}^{T} exp(\sum_{i,j \in S} \Theta_{i,j} * 1_{y_t=i} * 1_{y_{t-1}=j} + \sum_{i \in S} \sum_{j \in O} \mu_{o,i} * 1_{y=i} * 1_{x_t=o})$,

S=all possible labels, O=all possible words

$$p(X, Y) = \frac{1}{Z(X, Y)} \prod_{t=1}^{T} exp(\sum_{k=1}^{K} \Theta_k * f_k(y_t, y_{t-1}, x_t))$$

Bemerkung: Bestimmung der Θ_k mathematisches Optimierungsproblem mit sehr wahrscheinlich keiner exakten Lösung

$$p(Y|X) = \frac{p(X,Y)}{\sum_{Y' \in S} p(Y',X)}$$

$$\boldsymbol{prediction(X) = argmax_Y(P(Y|X))}$$

- Alles bis jetzt ist Hidden Markov Model
- Linear-chain CRF
  - Teilmenge der f_k
  - Zusätzliche custom feature functions:

    $f_{k+1}(y_t, y_{t-1}, X) = 1_{x_t \text{ startswith upper case}}$
    $f_{k+2}(y_t, y_{t-1}, X) = 1_{x_t \text{ is in a domain specific dictionary}}$
    $f_{k+3}(y_t, y_{t-1}, X) = 1_{x_t \text{ is in a domain specific dictionary}} * 1_{x_{t-1} \text{ is an article}}$
- Aufwand einer Auswertung von *prediction*: $(\#S)^2 * \#X$

Implementierung der New York Times

```
3/4          I1   L12   NoCAP   NoPAREN   B-QTY
pound        I2   L12   NoCAP   NoPAREN   OTHER
shiitake     I3   L12   NoCAP   NoPAREN   B-NAME
mushrooms    I4   L12   NoCAP   NoPAREN   I-NAME
,            I5   L12   NoCAP   NoPAREN   OTHER
stemmed      I6   L12   NoCAP   NoPAREN   B-COMMENT
and          I7   L12   NoCAP   NoPAREN   I-COMMENT
quartered    I8   L12   NoCAP   NoPAREN   I-COMMENT
```

Listing 6: Extract of the training data for New York Times CRF

```
U01:%x[-1,0] U02:%x[0,0] U03:%x[1,0] U07:%x
    [0,3] U14:%x[0,1]/%x[0,2]
B
```

Listing 7: Feature templates

- Über 130.000 labelled ingredient phrases
- 89% sentence-level accuracy
- Getestet mit 481 ingredient phrases
  - Nicht eindeutig:
    1 garlic clove, minced (optional)
  - Mehrere Zutaten in einer ingredient phrase nicht vorgesehen:
    4 tablespoons melted non-hydrogenated, melted coconut oil or canola oil
  - Fehler vom Algorithmus oder vom Menschen?

Ingredient Action relationship

(Ahmed 2009)

1. Dictionary-based entity recognition
   - Morphologisch Analyse der Wörter (z.B. Eier $\rightarrow$ Ei)
   - *Wort* $\in$ *Domain specifiy dictionary*?
2. Welche entities gehören zusammen?
   - Rule based: POS + context free grammar
   - Dependency based

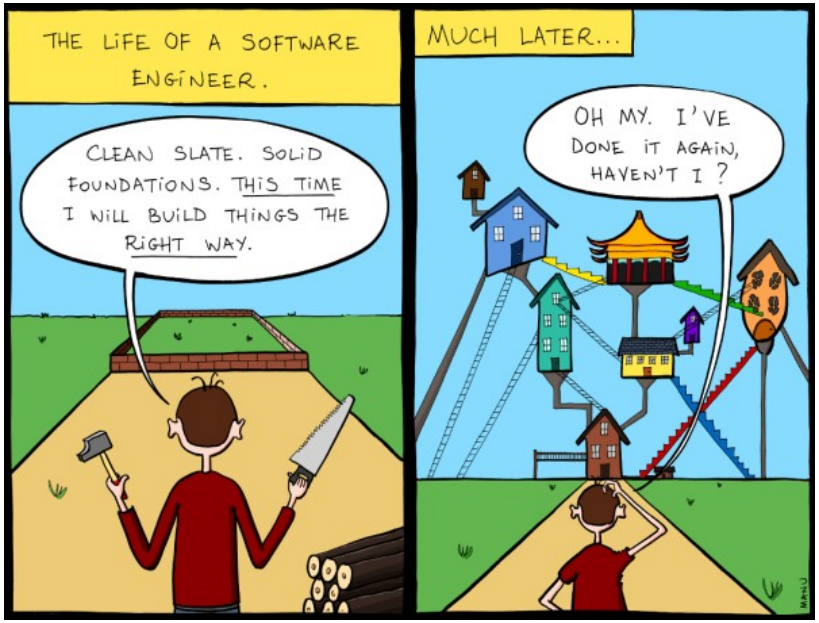
	Precision	Recall
Rule based	97.39%	51.54%
Dependency Based	95.4%	64.12%

Schwerpunkt setzen

- Schema.org/Recipe erweitern
  - Via GitHub
  - Offensichtlich Bedarf (Chefkoch, NYT)
- Algorithmus
  - Keine 130.000 Trainingsdaten vs. bessere feature functions (domain-specific dictionary check)
  - Decision tree (*if* $negationWord \in X$...)
- Webseite

- Ende Februar / im März nur noch schreiben

- SoP (Separation of Concerns)
- Martin Fowler: Duplizierter Code ist die Nummer eins der Gestanksparade
- Agile manifest: Simplicity –the art of maximizing the amount of work not done– is essential
- Praktiken von XP: Simple Design, consider the simplest thing that could possible work & YAGNI (You aren't gonna need it)



Danke für die Aufmerksamkeit

Literaturverzeichnis I

- 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.
- Schema.org (2016). *Schema.org*. URL: https://schema.org/ (visited on 10/30/2016).
- *Skip The Pizza* (2012). URL: http://skipthepizza.com/blog/analyzing-the-ingredients-of-29200-recipes (visited on 10/28/2016).
- Sutton, Charles and Andrew McCallum (2012). "An Introduction to Conditional Random Fields". In: *Found. Trends Mach. Learn.* 4.4, pp. 267–373. ISSN: 1935-8237. DOI: 10.1561/2200000013. URL: http://dx.doi.org/10.1561/2200000013.

Literaturverzeichnis II

The New York Times (2015). *Extracting Structured Data From Recipes Using Conditional Random Fields*. URL: http://open.blogs.nytimes.com/2015/04/09/extracting-structured-data-from-recipes-using-conditional-random-fields/?_r=1 (visited on 10/30/2016).