Extracting recipe ingredients from cookbooks

Ritter und Fabelwesen

von
Torsten Knauf

Der Beginn einer Master-Arbeit

Irrlichter

Heraus aus dem Sumpf

Contents

1. Introduction
2. Making a cookbook machine readable
3. Related Work
4. CRF-based extraction
5. Dictionary- and Rule-based extraction
6. Discussion
7. Summary

2. Making a cookbook machine readable

1. Digitalisation
2. CueML ontology
3. Need for automation

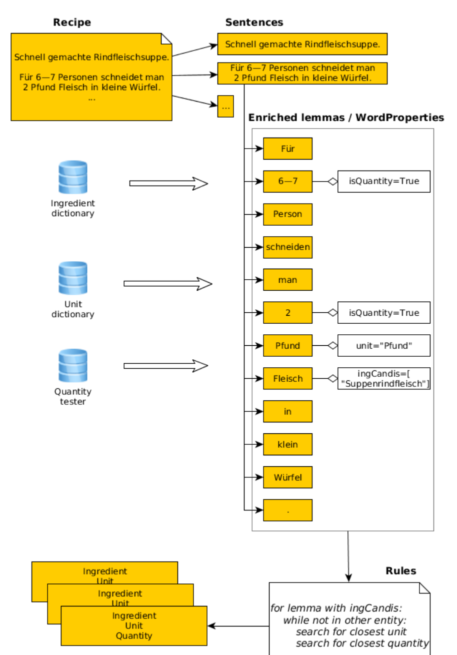
3. Related Work

1. Skip The Pizza
2. Extracting Structured Data From Recipes Using Conditional Random Fields
    1. CRF
    2. Implementation of NYT
3. Domain Specific Information Extraction for Semantic Annotation
4. Data-driven Knowledge Extraction for the Food Domain
5. Lessons for this work

4. CRF-based extraction

1. CRF prototype
2. Evaluation

5. Dictionary- and Rule-based extraction



Recipe

Schnell gemachte Rindfleischsuppe.

Für 6—7 Personen schneidet man
2 Pfund Fleisch in kleine Würfel.
...

Sentences

Schnell gemachte Rindfleischsuppe.

Für 6—7 Personen schneidet man
2 Pfund Fleisch in kleine Würfel.

...

Enriched lemmas / WordProperties

Für	
6—7	isQuantity=True
Person	
schneiden	
man	
2	isQuantity=True
Pfund	unit="Pfund"
Fleisch	ingCandis=["Suppenrindfleisch"]
in	
klein	
Würfel	
.	

Ingredient
dictionary

Unit
dictionary

Quantity
tester

Rules

```
for lemma with ingCandis:
    while not in other entity:
        search for closest unit
        search for closest quantity
```

Ingredient
Unit

Ingredient
Unit

Ingredient
Unit
Quantity

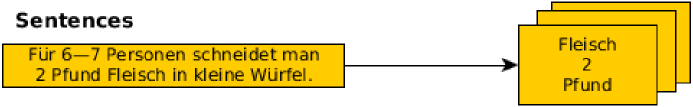
5. Dictionary- and Rule-based extraction

```
<cue:ingredient xml:id="Midder"
     BLSref="V582100">
    <cue:prefBasicForm>
       Midder
    </cue:prefBasicForm>
    <cue:altBasicForm>
       Kalbsmidder
    </cue:altBasicForm>
    <cue:altBasicForm>
       Bries
    </cue:altBasicForm>
    <cue:altBasicForm>
       Kalbsmilch
    </cue:altBasicForm>
    <cue:note>
       "Kalbsmidder ist auch
       unter dem Synonym [...]"
       (http://www.[...])
    </cue:note>
</cue:ingredient>
<cue:ingredient
     xml:id="Rindkochfleisch"
     BLSref="U180100">
    <cue:prefBasicForm>
       Rindfleisch
    </cue:prefBasicForm>
<cue:ingredient
     xml:id="Hammelfleisch"
     BLSref="Y400003">
    <cue:prefBasicForm>
       Hammelfleisch
    </cue:prefBasicForm>
</cue:ingredient>
```

Ingredient dictionary

```
{
      Midder         :  V582100
      Kalbsmidder    :  V582100
      [...]
      Rindfleisch    :  U180100
      Hammelfleisch  :  Y400003
      [...]
      Fleisch        :  [U180100,
                          Y400003,
                          ...
                         ]
}
```

5. Dictionary- and Rule-based extraction

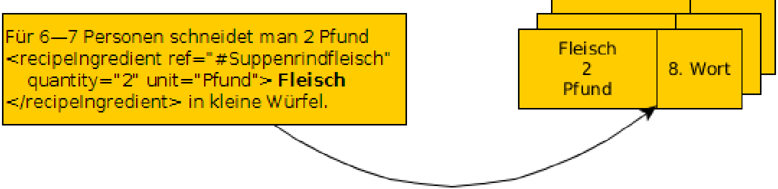


Sentences

Für 6—7 Personen schneidet man 2 Pfund Fleisch in kleine Würfel.

Fleisch
2
Pfund

$$Recall = \frac{\#(retrieved \cap relevant)}{\#relevant}$$

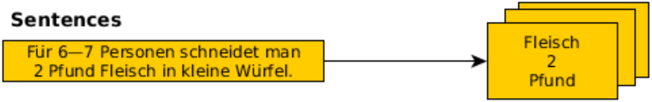


Für 6—7 Personen schneidet man 2 Pfund <recipeIngredient ref="#Suppenrindfleisch" quantity="2" unit="Pfund"> **Fleisch** </recipeIngredient> in kleine Würfel.

Fleisch
2
Pfund

8. Wort

5. Dictionary- and Rule-based extraction

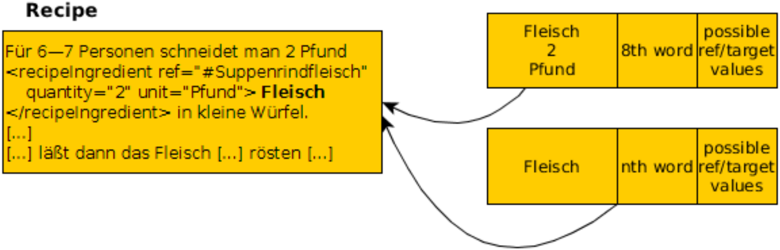


Sentences

Für 6—7 Personen schneidet man 2 Pfund Fleisch in kleine Würfel.

Fleisch
2
Pfund

$$Precision = \frac{\#(retrieved \cap relevant)}{\#retrieved}$$



Recipe

Für 6—7 Personen schneidet man 2 Pfund
<recipeIngredient ref="#Suppenrindfleisch"
quantity="2" unit="Pfund"> **Fleisch**
</recipeIngredient> in kleine Würfel.
[...]
[...] läßt dann das Fleisch [...] rösten [...]

Fleisch 2 Pfund	8th word	possible ref/target values
Fleisch	nth word	possible ref/target values

5. Dictionary- and Rule-based extraction

Evaluation with recipes B-1 to B-50:

(Only considering ingredients)

- Recall: 0.807 (394/488)
- Precision: 0.833 (434/521)
- Time: 144.7 seconds
- Flaws:
  - Lemmatization (*Saucissen* $\not\Rightarrow$ *Saucisse*)
  - Is *Brühe* an ingredient?
  - Ingredients within title not tagged

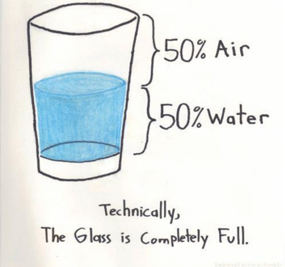
5. Dictionary- and Rule-based extraction

1. Dictionary- and Rule-based prototype
    1. Conceptual idea
    2. Evaluation
2. Refinement of prototype
    1. Illustrative enhanced rules
    2. Evaluation
3. GermaNet

6. Discussion

1. Usefulness of automatic extraction of ingredients
2. Quality of cueML and the obtained data
3. The development process
4. Knowledge is power

(8.) Ich bin Realist



Ich glaube, ich kann:

- Eine makellose Zutatenliste für jedes Rezept automatisch extrahieren
- Alle Informationen aus dem Buch extrahieren
- Eine wunderschöne Webseite zum Kochbuch erstellen und mit Inhalt füllen
- Einen Nutellabaum pflanzen

*XML-Tagger