Extracting recipe ingredients from cookbooks

Heraus aus dem Sumpf

von
Torsten Knauf

Der Beginn einer Master-Arbeit

Irrlichter

Beispiel cueML

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

    <p>Es wird hierzu für <cue:yield atLeast="24" atMost="30">24-30
        Personen</cue:yield> eine kräftige <ref
        target="#Bouillon">Bouillon</ref> von 8-10 Pfund <cue:recipeIngredient
        ref="#Rindkochfleisch" atLeast="8" atMost="10"
        unit="Pfund">Rindfleisch</cue:recipeIngredient> mit
        <cue:recipeIngredient ref="#Wurzelwerk">Wurzelwerk
        </cue:recipeIngredient> gekocht. [...]</p>

    <note>Anmerk. Der <cue:recipeIngredient ref="#Englische_Soja"
        optional="True">Soja</cue:recipeIngredient> macht die Suppe
        gewürzreicher, kann jedoch gut wegbleiben, und statt
        <cue:recipeIngredient ref="#Madeira"
        altGrp="1">Madeira</cue:recipeIngredient> kann man
        <cue:recipeIngredient ref="weißen_Franzwein" altGrp="2">weißen
        Franzwein</cue:recipeIngredient> und etwas <cue:recipeIngredient
        ref="#Rum" altGrp="2" quantity="etwas">Rum</cue:recipeIngredient>
        nehmen<cue:alt target="1␣2"/>. Sowohl die Bouillon als Kalbskopf
        können schon am vorhergehenden Tage, ohne Nachtheil der Suppe, gekocht
        werden.</note>
</cue:recipe>
```

Wie man es hätte auch machen können I

```
<...>
und statt <cue:recipeIngredient ref="#Madeira"
     altGrp="1">Madeira</cue:recipeIngredient> kann man
     <cue:recipeIngredient ref="weißen_Franzwein" altGrp="2">weißen
     Franzwein</cue:recipeIngredient> und etwas <cue:recipeIngredient
     ref="#Rum" altGrp="2" quantity="etwas">Rum</cue:recipeIngredient>
     nehmen <cue:alt target="1␣2"/>.

    VS

und statt <recipeIngredient ref="#Madeira"
     xml:id="B-16-Madeira">Madeira</recipeIngredient> kann man
     <recipeIngredient ref="weißen_Franzwein"
     xml:id="B-16-Franzwein">weißen Franzwein</recipeIngredient> und etwas
     <recipeIngredient ref="#Rum" xml:id="B-16-Rum">Rum</recipeIngredient>
     nehmen. Sowohl die Bouillon als Kalbskopf können schon am
     vorhergehenden Tage, ohne Nachtheil der Suppe, gekocht werden.

<recipeIngredientGrp xml:id="B-16-alt1" target="#B-16-Madeira"/>
<recipeIngredientGrp xml:id="B-16-alt2" target="#B-16-Franzwein␣#B-16-Rum"/>
<alt target="#B-16-alt1␣#B-16-alt2" mode="excl"/>
```

Wie man es hätte auch machen können II

```
<...>
8-10 Pfund <cue:recipeIngredient ref="#Rindkochfleisch" atLeast="8"
     atMost="10" unit="Pfund">Rindfleisch</cue:recipeIngredient>

VS

<atLeast target="#Suppenrindfleisch">8</atLeast>-<atMost
     target="#Suppenrindfleisch">10</atMost> <unit
     target="#Suppenrindfleisch">Pfund</unit> <recipeIngredient
     ref="#Suppenrindfleisch">Rindfleisch</recipeIngredient>
```

Für quantity-Attribut *ref="#EINS"* für *eine*
Für unit-Attribut *ref="#MaßDef"* für *Maß*

1. CRF Prototyp :)

```
für             ... O
24              ... B-Yield
-               ... I-Yield
30              ... I-Yield
Personen        ... I-Yield
kräftige        ... O
Bouillon        ... O
von             ... O
8               ... B-Quantity
-               ... I-Quantity
10              ... I-Quantity
Pfund           ... B-Unit
Rindfleisch     ... B-Ingredient
mit             ... O
Wurzelwerk      ... B-Ingredient
gekocht         ... O
.               ... O
```

```
O                 O                 Hammelfleischsuppe
O                 O                 Das
B-Ingredient      B-Ingredient      Fleisch
O                 O                 wird
O                 O                 in
O                 O                 Stückchen
O                 O                 gehauen
O                 O                 ,
O                 O                 gut
O                 O                 gewaschen
O                 O                 ,
O                 O                 mit
B-Ingredient      B-Ingredient      Salz
O                 O                 ausgeschäumt
O                 O                 ,
B-Ingredient      B-Ingredient      Wurzelwerk
O                 O                 ,
B-Quantity        O                 eine
O                 O                 fein
O                 O                 geschnittene
B-Ingredient      B-Ingredient      Zwiebel
O                 O                 und
B-Quantity        O                 einige
O                 O                 in
O                 O                 Würfel
O                 O                 geschnittene
B-Ingredient      O                 Kartoffeln
```

```
B-Quantity          O                   eine
B-Quantity          O                   einige
B-Ingredient        O                   Kartoffeln
B-Ingredient        O                   Kartoffeln
B-Ingredient        O                   Kartoffel-Klöße
Insgesamt 96 Wörter, 13 labels != 'O' und 5 Unterschiede
```

Was noch fehlt

- ↯ *optional="True"*
- ↯ *altGrp="x"*
- "Das Kalbfleisch wie in No. 1, nach der Personenzahl, doch etwas reichlicher genommen, da solches weniger Kraft gibt, als Rindfleisch." ↯ *dontUse="True"*

$\Rightarrow$ Label & Feature Engineering

- Neue Labels: OptionalIngredient, AlternativeIngredient, DontUseIngredient
- Features: w[-1], w[-2], w[1], w[2], ...
- Noch mehr Labels und bigram-Features: "statt"/"oder" $\rightarrow$ IndicatorForAlternativeIngredient, "kann" $\rightarrow$ IndicatorForOptionalIngredient, ...

Was dann noch immer fehlt

- ist altGrp $\lightning$ *aber altGrp* $\overset{!}{=}$ *x*
- "Rindfleischsuppe [...] das Fleisch" $\lightning$ *implizite Informationen; ref* $\overset{!}{=}$ *x*
- "8-10 Pfund Rindfleisch"
  $\rightarrow$ "B-Quantity O B-Quantity B-Unit B-Ingredient"
  $\lightning$ *Zuordnung von Unit und Quantity*

Ist CRF überhaupt der richtige Ansatz? :/

- NYT betrachten nur Zeilen aus Zutatenliste & hatten bereits über 130.000 gelabelte Trainingsdaten
- Labeln ist sehr aufwändig (10 Rezepte $\rightarrow$ 1514 Zeilen $\rightarrow$ 1514/60*3 $\rightarrow$ ca. 1.5h bis zu 3h)
- Wie gut wird das CRF sein?

Alternativer Ansatz

1. Lemmatisierung
2. Dictionary-based Extraktion von Ingredients, Quantities, Units
3. Rule-based Entity Processing

```
if "kann" in sentence:
    # add "optional='True'" to Ingredient

if "statt" in sentence:
    # add "altGrp='id1'" to first ingredient
    # add "altGrp='id2'" to following
        ingredients

if ingredient == "Fleisch" and
        recipe.Name.find("Rind") != -1:
    # ingredient = Rindfleisch
if ingredient == "Rindfleisch" and
        recipe.rcpId == "Suppe":
    # ingredient = Suppenrindfleisch
```

Brainstorming CRF vs. Rule-based Entity Processing

- 8-10
- optional Ingredient

- Link target?
- Zuordnungen von altGrps, Quantity / Unit / Ingredients nicht via CRF
- Rule-based schneller zu entwickeln und gezielter testbar / steuerbar (Feature, Labels, Trainingsdaten -> ?, Rules ändern -> Unit tests :))

Happy about ALL kind of feedback