Zusammenfassung & Ausblick

Zusammenfassung

Ausblick

Zusammenfassung

Das Ziel dieser Arbeit ist es, Frau Davidis' Kochbuch für eine kulinarische Analyse digital aufzubereiten. Die digitale Edition der Rezepte ist hier (Rezepte.html) zu finden.

Bis zu dieser Arbeit gab es noch kein frei verfügbares Vokabular in der Koch-Domäne, welches eine kulinarische Analyse von Rezepten ermöglicht. Daher haben wir Anforderungen an so ein Vokabular aufgestellt und darauf aufbauend culinary editions Markup Language (cueML) (cueML.html) entwickelt. CueML erweitert den Standard der Text Encoding Initiative (TEI) (http://www.tei-c.org/index.xml) und Schema.org/Recipe (https://schema.org/Recipe). Es führt unter anderem Möglichkeiten zur Auszeichnung von Mengenangaben und Einheiten von Zutaten ein sowie Unterscheidungen zwischen optionalen und alternativen Zutaten. Sowohl Verweise auf Teil-Rezepte als auch allgemeine Verweise auf andere Stellen können gesetzt werden.

Da das manuelle Auszeichnen zeitaufwendig und fehleranfällig ist, haben wir des Weiteren an Prototypen zum automatischen Auszeichnen geforscht. Dazu müssen die auszuzeichnenden Entities zuerst extrahiert werden. Für das Extrahieren haben wir zwei verschiedene Ansätze ausprobiert. Der Conditional Random Field-based Prototyp (AutomatischesAuszeichnenCueML.html#CRFPrototyp) hat sich nicht als zielführend erwiesen. Der dictionary- and rule-based Prototyp (AutomatischesAuszeichnenCueML.html#Dictbased) ist hingegen vielversprechend. Dieser lemmatisiert zu erst jedes Wort des Textes. Die Lemmatisierung baut auf TreeTagger (http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/) auf. Aufgrund der alten verwendeten deutschen Sprache von Frau Davidis entwickeln wir eine einfache Überarbeitung des Outputs von TreeTagger. Nach der Lemmatisierung können die Lemmata in Wörterbüchern nachgeschaut werden, welche alle gesuchten Entities enthalten. In der Koch-Domäne ist die Annahme der Existenz solcher Wörterbücher vertretbar. Nachdem die Entities extrahiert wurden, können diese mittels Regel weiterverarbeitet werden. Die *rule-based* Weiterverarbeitung ermöglicht z. B. das Zuordnen von Mengenangaben zu Zutaten oder eine Differenzierung, dass eine Zutat im Rezept optional ist.

Ausblick

Diese Arbeit war der Startschuss in ein neues Forschungsfeld. Dementsprechend stehen noch weitere Arbeiten an, die wir folgend kurz auflisten wollen:

- Die automatische Extraktion beschränkt sich zur Zeit noch auf Zutaten, sowie ihre Mengenangaben und Einheiten. Es fehlen noch folgende Entities: Vorbereitungszeit, Kochzeit, gesamte Zubereitungszeit, Zubereitungsmethode, Verweise auf allgemeine Stellen und Verweise auf inkludierte Teil-Rezepte.
- Eine weitere Differenzierung von Zutaten in *Beilagen* erscheint uns sinnvoll. Ansonsten würden z. B. beim Rezept *„Schmalz- oder Butterkohl“* aufgrund von *„Beilagen: Schinken, Rauchfleisch, Bratwurst.“* Bratwurst als Zutat in der Zutatenliste auftauchen.
- Verbesserung bzw. hinzufügen von weiteren Regeln in der rule-based Weiterverarbeitung des dictionary- and rule-based Prototypen (AutomatischesAuszeichnenCueML.html#Dictbased). Um die durch Regeln extrahierten Informationen nachvollziehen zu können, falls beispielsweise fehlerhafte Informationen extrahiert wurden, sollte die Historie von angewendeten Regeln transparent sein. Diese sollten nicht in der Auszeichnungssprache cueML integriert werden, sondern als Debugging-Informationen dem Software-Entwickler verfügbar gemacht werden.
- Das Finden bzw. das Aufbauen eines allgemeinen Zutaten-Wörterbuches; idealerweise als weltweit frei verfügbare, eindeutige Ressource, mit Nährwertangaben.
- Veröffentlichen und verbreiten von cueML.
- Und zu guter Letzt natürlich die eigentliche kulinarische Analyse.