Multimodal Event Recognition with an Ontology For Cooking Recipes

Alison Emily CHOW [a] and Michael GRÜNINGER [a]

[a] *University of Toronto, Department of Mechanical and Industrial Engineering, Canada*

Abstract. The tasks of recognizing events and following instructions are ubiquitous in everyday life. This paper focuses on developing an ontology-based approach to understand the semantics behind cooking instructions; the methodology and analysis of process descriptions across the auditory, visual and textual representations of a typical cooking recipe were generated and performed. Specifically, formalized relations, using first order logic from the Process Specification Language (PSL), across the three modalities in the domain of cooking were contrasted and helped uncover several insights and comparisons, such as the difference in sequential ordering, activity complexity, and implicit constraints. Altogether, the analysis of relationships between modalities in the cooking domain led us to further our understanding on how to represent steps in a recipe, and explore future work, such as automatic process description generators and a cooking-based ontology.

Keywords. multimodal event recognition, cooking, ontology, process specification language, process description

1. Introduction

Although we live in a dynamic world, with many events occurring around us, our understanding of events is often quite difficult to articulate. We all recognize when change occurs, yet we do not always agree on the processes that cause the change. Even when we agree on the underlying ontology of processes, there may be disagreement about how the composition or granularity of processes is described, even when two people are observing the same event. Previous work on event recognition has focused on a single modality, such as video or text; however, people typically use all modalities in conjunction with one another. Furthermore, the event descriptions that are generated from each modality are often conflicting with each other, even though they are supposedly describing the same process. How can we possibly harmonize event recognition across all three modalities?

The recognition of events occurring in the world presumes an underlying representation of events. A major drawback in current work is the lack of expressive process representations. As highlighted in the Preface of the CREOL Workshop in 2017, "Additional properties of events are currently missing: duration of events, event internal substructure, event pre- and post-situations, relations to other events in terms of explanatory/causal and temporal relations. These properties are essential to promote reasoning on events and their participants, and they may vary according to the specific context of occurrence in a text/document." [1] To address these problems in process representations, we will be using the Process Specification Language (PSL) ontology as the basis for the process modelling in this paper.

We address the multimodality problem through uncovering the relationships between events across a variety of modalities with an ontological approach. This paper will explore the role that ontologies play across all three modalities of video, audio, and text in the specific domain of cooking. Recipes in the cooking domain are ideal sample sets of data for multimodal event recognition; this is mainly due to the standardization across recipes, the rich sources of data in all three modalities of video, audio, and text, and the abundance of resources in the domain.

1.1. Related Work

Despite cooking recipes being a popular domain for semantics research, there has not been significant research done to compare the relationships between the three different modalities using an ontology-based approach. Related work has been done mainly for cooking instructions in the textual recipe modality, as seen in Ribeiro et al. and Malmaud et al's work [13] [9], instead of a multimodal approach, and for the sole purposes of automating the process of recipe or text extraction of key categories, such as resources and activities. Malmaud et al. also highlight the problems in textual cooking recipes, including the lack of explicit details and ambiguous object references. The objects used in recipes are implicitly available based on: domain-specific expectations of the initial context or, results from preceding steps in the recipe.

The goal of our research is two-fold: to evaluate the adequacy of the process ontology of PSL (Process Specification Language) to formalize the process descriptions for each modality of a recipe, and then compare the process descriptions to uncover the relationships between activities and objects in the recipes across modalities. We are particularly interested in two questions:

- Do different modalities give rise to different process descriptions?
- What are the relationships between the process descriptions generated from each modality?

2. Event Modelling

2.1. Existing Approaches

Several efforts have proposed ontologies for activities and temporal concepts to support the annotation and recognition of events in video ([11], [6], [14], [8], [16]). Existing approaches lack the expressiveness needed for the representation of the processes that underly recipes. In particular, recipes require the representation of composition of activities, partial orderings of activities within complex activities, duration, and a specification of preconditions and effects (i.e. how activity occurrences depend on the state of the world and how they change the state of the world). Although these approaches may contain constructs corresponding to these concepts, the axiomatization is too weak to capture the intended semantics of the concepts. The subsequent existence of unintended models means that the ontologies cannot support automated reasoning (without the supplementation of extralogical implementations).

2.2. PSL Ontology

The Process Specification Language (PSL) ([4], [5]) has been designed to facilitate correct and complete exchange of process information among manufacturing systems. Included in these applications are scheduling, process modelling, process planning, production planning, simulation, project management, workflow, and business process reengineering[1]. PSL is a modular, extensible ontology capturing concepts required for process specification. There are currently 300 concepts across 50 extensions of a common core theory (PSL-Core) all axiomatized in first-order logic using Common Logic (ISO 24707).

2.2.1. PSL Core

The minimal module is referred to as PSL-Core[2], which axiomatizes the fundamental ontological categories, while extensions to PSL-Core axiomatize additional relations and properties. There are four kinds of entities required for reasoning about processes – activities, activity occurrences, timepoints, and objects. Activities may have multiple occurrences, or there may exist activities which do not occur at all. Activity occurrences and objects are associated with unique timepoints that mark the begin and end of the occurrence or object. Timepoints are linearly ordered, forwards into the future, and backwards into the past. Objects may participate in activity occurrences at specific timepoints.

2.2.2. Subactivities

A ubiquitous feature of process formalisms is the ability to compose simpler activities to form new complex activities (or conversely, to decompose any complex activity into a set of subactivities). The PSL Ontology incorporates this idea while making several distinctions between different kinds of composition that arise from the relationship between composition of activities and composition of activity occurrences.

The PSL Ontology uses the $subactivity(a_1, a_2)$ relation to capture the basic intuitions for the composition of activities[3]. The core theory $T_{subactivity}$ axiomatizes this relation as a discrete partial ordering, in which primitive activities are the minimal elements. In this way, activities can be composed together to construct complex activities.

2.2.3. Complex Activities

The theory of subactivities alone does not specify any relationship between the occurrence of an activity and occurrences of its subactivities. Occurrences of complex activities correspond to sets of occurrences of their subactivities[4]. Different occurrences of complex activities may contain occurrences of different subactivities or different orderings on the same subactivity occurrences. There are different ordering relations on

[1] The PSL Ontology has been adopted as part of the ISO 18629 International Standard.

[2] `colore.oor.net/psl_core/psl_core.clif`

[3] `colore.oor.net/psl_subactivity/subactivity.clif`

[4] `colore.oor.net/psl_complex/complex.clif`

activity occurrences – the ordering relation $precedes(s_1, s_2)$ on possible activity occurrences, a linear ordering relation $min\_precedes(s_1, s_2, a)$ of subactivity occurrences of a complex activity occurrence, and a partial ordering $soo\_precedes(s_1, s_2, a)$ of subactivity occurrences for a set of complex activity occurrences Classes of complex activities are therefore defined with respect to the following two criteria:

- the relationship between the occurrence of the complex activity and occurrences of its subactivities;
- the conditions under which a complex activity occurs.

2.2.4. Activities and State

Properties in the domain that can change are called *fluents*. Similar to the representation of activities, fluents can also be denoted by terms within the language. Intuitively, a change in state is captured by the set of fluents that are either achieved or falsified by an activity occurrence[5]. The $prior(f, o)$ relation specifies that a fluent f is intuitively true prior to an activity occurrence o and the $holds(f, o)$ relation specifies that a fluent f is intuitively true after an activity occurrence o.

There are constraints on which activities can possibly occur in some domain (preconditions). State is changed by the occurrence of activities, and state can only be changed by the occurrence of activities. State does not change during the occurrence of a primitive activity.

3. Approach to Ontology Development

The following section discusses the background of how and why we formalized each modality, and further expands on the details behind the approach for each visual, auditory and textual modality.

3.1. Approach

Inspired by Ribeiro et al.'s methodology of knowledge acquisition, conceptualization and formalization through in-person brainstorming sessions and weekly meetings, the process of transitioning from knowledge acquisition to formalization is also used in our research [13]. We also employed the process of transcribing recipes using hand annotation as the ground truth, similar to Malmaud et al.'s manual process [9].

First, we sampled a set of five English instructional videos of approximately 3-5 minutes in length from the Food Network, with corresponding recipes in text format by the same instructing chef. The sampled videos vary in chefs responsible and recipe types (baking, frying, etc); however, it is important to note that the focus of this paper is primarily on the differences between modalities by the same instructor, not across a broad number of recipes. Thanks to the consistency in authoring across modalities, the multimodal recipes allowed us to dive deeper into the analysis of one recipe to uncover the significant takeaways between its modalities.

Afterwards, we analyzed each recipe's content according to its three corresponding modality methodologies, in the following order: (1) the visual instructional video without

[5] `colore.oor.net/psl_disc_state/disc_state.clif`

audio playing, (2) the audio transcript solely derived from the instructional video, and finally, (3) the recipe in its textual form. This order was consciously constructed to ensure independence between the analysis of different modalities and avoid any bias that may have been introduced by the identification of specific activities in the textual or auditory representations of the recipe. For simplicity's sake, specific ingredient measurements were excluded from the analysis of each recipe.

The following three modality-specified methodologies were formulated to identify the visual, audio and textual cues in the recipe, and help generate and formalize a recipe's corresponding process description of axioms in PSL from its original transcribed narrative perspective.

3.2. Textual Recipe

We adopted Malmaud et al.'s procedure [9] using linguistic cues to analyze textual recipes, particularly for identifying activities as verbs in textual recipes. Each textual recipe is composed of a numbered set of steps, and similarly to the audio transcript, each step is composed of one or more English sentences and subactivities of a step were identified with the presence of conjunctions in a sentence. Unlike the auditory modality, the textual recipe has a predefined numbered syntactic structure consistent across the majority of recipes in the textual form; therefore, conjunctive adverbs, serving as a linguistic cue, are not necessary to indicate when a step occurs.

The detailed methodology for textual formalization is as follows:

1. Each numbered step in the textual recipe corresponds to a discrete subactivity in the overall process.
2. The sequencing of numbered steps implies a set of ordering constraints to be followed by the human reader, as well as the implied ordering of actions in each sentence. For example, if an action, *a1*, precedes another action, *a2*, an ordering constraint will be formalized using *next_subacc(a1,a2)* to remain consistent.
3. Each identified verb corresponds to an activity, each noun corresponds to an object that participates in an activity, and each preposition, such as *"until melted"*, corresponds to a constraint.
4. Time details, such as *"bake for 30 minutes"*, provided in the recipe instructions correspond to temporal constraints.
5. Activities identified can also be done in parallel, as indicated by adverbs describing simultaneous actions, such as *"while"*.
6. Units of measure for quantity are ignored (although future work will incorporate such constraints).

Following the comparison of the three recipe modalities, we defined primitive activities to be repeatable patterns of behaviours [4] at the lowest level of granularity that are responsible for all the physical change in the world and cannot be decomposed into further subactivities [12]. Therefore, we decided to remain agnostic throughout the formalization approach to explore the differences after comparing the process descriptions.

3.3. Audio Transcript from Instructional Video

Each audio transcript does not follow a clearly defined set of steps; instead, it contains a continuous sequence of all sentences as steps together, each step separated by conjunctive adverbs, such as ”then”, ”finally” and ”now”. Conjunctive adverbs read aloud serve to inform the subject when the next step occurs and each step is composed of one or more English sentences to help guide the reader through the cooking process. To identify the subactivities of a step, conjunctions (such as ”and”, ”or” and ”but”) were used. Based on the number of conjunctions in a sentence, each step would be divided into the corresponding number of subactivities.

The methodology to formalize each sentence is consistent with the methodology outlined by the textual modality, as outlined in detail in Section 3.2, as they both contain linguistic cues to signify the activities, resources and duration of a step.

3.4. Visual Instructional Video Without Audio

Each instructional video was viewed in the absence of audio, in order to capture the visual cues in the instructional video independently. Despite continuous efforts by the semantic community to bridge the gap between visual representations and linguistic cues [2], there has not been sufficient work done in the cooking domain to map visual cues to events occurring in cooking videos. To add, visual and auditory cues in cooking videos are often used in conjunction with one another, making it difficult to analyze the two modalities independently and formalize the instructions.

Next steps for a more holistic and complete methodology for visual cues are outlined in Section 6.2. The visual cues employed in our research to identify the activities were mainly deduced from observing distinct movements and actions performed by the chef and recording actions in a continuous, freeflow structure. Overall, the aim was not to achieve uniformity of description but rather to elicit the potential diversity of process descriptions that can arise from the same video.

3.5. Limitations

A limitation that arose with developing a complete process description for the audio transcript was the occasional absence of verbs to convey actions in a recipe. The narrating chef commonly left out verbs when describing an activity performed, and instead relied on performing the activity visually. To address this limitation, we chose to represent the missing verb as *Unknown_Activity*. In similar fashion to Skolemization, a technique used to remove existential quantifiers from formulas [7], the original axiom of ”for every *o*, the occurrence of the step, there exists an activity, *Activity* such that *occurrence_of(o,Activity)*” will be transformed to ”there exists a function *Unknown_Activity* mapping every *o* into a *Unknown_Activity* such that, for every *o* it holds *occurrence_of(o,Unknown_Activity(o))*”.

4. Formalization

Using the general guidelines for analysis of recipes and keeping limitations in mind from Section 3.1, a set of axioms was constructed for each recipe’s modality using the Process Specification Language (PSL). One annotator formalized the set of axioms for each modality from the *Food Network's Lynn Crawford's Buttermilk Fried Chicken* [3] recipe, also known as Recipe1[6].

The textual modality of Recipe1, where each ordered step from the numbered list is represented as an occurrence, can be found below:

$$\begin{gathered}(\forall o)\ \mathit{occurrence\_of}(o, \mathit{Recipe}1) \supset \\ (\exists o1, o2, o3, o4, o5, x) \mathit{occurrence\_of}(o1, \mathit{Step}1(x)) \wedge \mathit{occurrence\_of}(o2, \mathit{Step}2(x)) \wedge \\ \mathit{occurrence\_of}(o3, \mathit{Step}3(x)) \wedge \mathit{occurrence\_of}(o4, \mathit{Step}4(x)) \wedge \mathit{occurrence\_of}(o5, \mathit{Step}5(x)) \wedge \\ \mathit{next\_subacc}(o1, o2) \wedge \mathit{next\_subacc}(o2, o3) \wedge \mathit{next\_subacc}(o3, o4) \wedge \mathit{next\_subacc}(o4, o5)\end{gathered}$$

After defining the overall sequencing of the recipe, each step was formalized. Each verb corresponds to a subactivity in each step, and each step is denoted as an activity. For example, the first step from the original text recipe, *In a large resealable plastic bag set over a large bowl, combine 2 cups of buttermilk, Dijon mustard, 1 tbsp hot sauce, 2 tbsp onion powder, 1 tsp salt, hot sauce, 1 tsp black pepper, fresh thyme, and chicken pieces. Press air from bag, seal, and refrigerate for at least 12 hours.*, is formalized in PSL below:

$$\begin{gathered}(\forall o)\ \mathit{occurrence\_of}(o, \mathit{Step}1) \supset \\ (\exists o1, o2, o3, o4, a)\ \mathit{combine}(a)\ \wedge\ \mathit{occurrence\_of}(o1, a) \wedge \mathit{occurrence\_of}(o2, \mathit{PressAir}(x) \wedge \\ \mathit{Bag}(x)) \wedge \mathit{occurrence\_of}(o3, \mathit{Seal}(x)) \wedge\ \mathit{occurrence\_of}(o4, \mathit{Refrigerate}(x)) \wedge \\ \mathit{greaterEq\_duration}(\mathit{duration\_of}(o4), \mathit{multduration}(12, \mathit{hour})) \wedge \mathit{next\_subacc}(o1, o2) \wedge \\ \mathit{next\_subacc}(o2, o3) \wedge \mathit{next\_subacc}(o3, o4)\end{gathered}$$

5. Analysis of Process Descriptions

Through the comparison of formalized process descriptions between the three modalities from above Section 4 in PSL, the following insights were uncovered: the sequence, presence, implicitness, complexity, substitution, timing, and narration of activities in cooking instructions.

5.1. Sequence and Presence of Activities

By cross-referencing the ordering constraints between modalities, it can be seen that several activities did not have a defined order between modalities or simply omitted certain activities, if deemed unnecessary to show or include.

5.1.1. Sequence of Activities

The differences in the sequence of steps is interesting to note; in the visual and auditory modalities, *Refrigerate(x)* is not performed until after mixing the chicken:

$$(\forall o)\ \mathit{occurrence\_of}(o, \mathit{Step}7) \supset (\exists o1) \wedge \mathit{occurrence\_of}(o1, \mathit{Mix}(x))$$

$$\begin{gathered}(\forall o)\ \mathit{occurrence\_of}(o, \mathit{Step}8) \supset (\exists o1) \wedge \mathit{occurrence\_of}(o1, \mathit{Refrigerate}(x)) \wedge \\ (\mathit{duration}(\mathit{beginof}(\mathit{Refrigerate}(x)), 240) \vee \mathit{duration}(\mathit{beginof}(\mathit{Refrigerate}(x)), 360))\end{gathered}$$

[6] The full set of axioms for all three modalities can be found on our Github repository at `https://github.com/gruninger/colore/tree/master/ontologies/cooking`.

However in the textual modality, *Refrigerate(x)* is immediately performed in Step 1, right after *Add(x), Press(x)* and *Seal(x)*.

This shows the any-order constraint that is implied for many of the steps across different modalities, and further highlights how each process description may simply be an occurrence of the "ground truth" behind a recipe. With a difference in sequencing between modalities, yet a consistent outcome for the ingredients to create the final recipe result, it goes to show how it is possible that a decision to choose which step to follow can result in the same outcome in cooking processes.

5.1.2. Presence of Activities

The presence and absence of activities differed across modalities for the same recipe; for example, in the visual and auditory modalities, the activity of *Preheat(x)* is excluded.

Despite it being a step for preparation, *Preheat(x)* is an important step in the cooking process; yet, it is neglected in the visual and auditory modalities and mentioned specifically in the textual modality in Step 2.

$$\begin{gathered}(\forall o)\ occurrence\_of(o, Step2) \supset (\exists o1, o2, o3, o4, o5, o6) \wedge occurrence\_of(o1, Preheat(x)\ \wedge \\ Oven(x)) \wedge\ occurrence\_of(o2, Remove(x)) \wedge\ occurrence\_of(o3, Arrange(x)) \wedge \\ occurrence\_of(o4, Discard(x) \wedge Marinade(x) \wedge occurrence\_of(o5, Roast(x)) \wedge \\ (duration(beginof(Roast(x)), 30) \vee duration(beginof(Roast(x)), 40)) \wedge \\ occurrence\_of(o6, Cool(x) \vee (Wrap(x) \wedge Refrigerate(x)) \wedge next\_subacc(o1, o2) \wedge \\ occurrence\_of(o6, Cool(x) \vee (Wrap(x) \wedge Refrigerate(x)) \wedge next\_subacc(o2, o3) \wedge \\ next\_subacc(o3, o4) \wedge next\_subacc(o4, o5) \wedge next\_subacc(o5, o6)\end{gathered}$$

The presence of this activity in the textual modality showcases how a detail such as preheating may be ignored in the presentation of the recipe in the cooking video due to the domain-specific knowledge required, not because it is trivial to the overall process description.

5.2. Implicit Steps, Preconditions and Constraints

The idea of implicit instructions is mainly present in the textual modality of the recipes, with certain activities involving commonsense reasoning or domain knowledge. Examples of referential pronoun ambiguity were observed in the following implicit instructions. To start, the instruction of *"Transfer chicken to a baking sheet and keep warm in oven"* fails to mention explicitly the subtle yet crucial details of how the human subject performing this step in the recipe should be placing the chicken, tray and baking sheet all in the oven, not only the chicken and baking sheet. Another example found in the textual modality is *"Carefully add chicken pieces skin side down to hot oil"*. This instruction fails to explicitly state the presence of a skillet, and how the chicken should be placed in the hot oil, which is located inside the skillet. Altogether, these small, yet meaningful, resources and activities are implied indirectly in cooking recipes, but need to be explicitly mentioned in order to provide the most accurate process description for a recipe or for the task of future recipe automation.

5.3. Activity Complexity

The complexity of each activity in the recipe was not defined at the start of the process, as described in Section 3.1. As a result, the formalization and analysis was performed agnostic to activity complexity, and based solely on the number of subactivities identified for each activity across modalities. It was interesting to note how certain activities, such as *combine* in the textual modality and *add* in the visual modality, were deduced consistently as complex activities interacting with multiple resources across all three modalities. This can be seen in the textual modality for the combine activity:

$$(\forall a)\, combine(a) \wedge occurrence\_of(o1,a) \supset$$
$$(\exists x1..x8)\ \ buttermilk(x1) \wedge mustard(x2) \wedge hotsauce(x3) \wedge onionpowder(x4) \wedge salt(x5) \wedge$$
$$blackpepper(x6) \wedge thyme(x7) \wedge chicken(x8) \wedge participates(x1,o1) \wedge participates(x2,o1) \wedge$$
$$participates(x3,o1) \wedge participates(x4,o1) \wedge participates(x5,o1) \wedge participates(x6,o1) \wedge$$
$$participates(x7,o1) \wedge participates(x8,o1) \wedge DryMixture(y)$$

However, the naming of the same activity was not referred to consistently across the modalities, and in the auditory modality, the activity referring to *combine* or *add* was omitted from the audio transcript. The chef simply referred to this activity informally without a prescribed action as: *"Two cups of buttermilk, a little tang, dijon mustard, teaspoon of onion powder, teaspoon of salt and black pepper"*. Overall, the activity complexity can be verified in one way by comparing it across modalities, and seemingly primitive activities may actually be complex.

5.4. Temporal Constraints

Temporal constraints help represent the element of time in the formalization of process descriptions, and duration is an element that affects multiple steps in recipes. One takeaway on temporal constraints is how specific time duration values cannot be directly observed in the visual modality, so the element of time is consequently lost in the formalization and analysis for the video. In the auditory and textual components, time is described mainly as a linguistic cue. This can be in the form of adverbs, such as "while", numeric intervals, and numbers, such as *"Roast chicken in oven for 30-40 minutes"* in the textual modality and *"Bake for 30-40 minutes"* in the auditory modality. These were represented in the formalization with *duration(start, end)*, and occasionally, these intervals would not match exactly with the corresponding activity performed in a different modality. This was observed in the conflicting duration for *Refrigerate* of 4-6 hours in the auditory modality and at least 12 hours in the textual modality. Despite the same author across all modalities, contradicting information regarding temporal constraints still remained, which begs the question: which duration should be followed for the optimal result and do both lead to the same result? The presence of "while" also introduces the idea of activity co-occurrence, since these are activities that can occur simultaneously.

6. Future Work

6.1. Linguistic Exploration

Building upon potential research questions for the textual modality, it would be interesting to explore whether or not the textual recipes impose additional complex axioms for more complex activities in its process description. Since the visual and auditory modalities are less common than the textual modality across recipes and different textual recipes share a common structure, the impact of linguistic cues is much more significant in the textual modality. Therefore, more research should be done in understanding the relationship between textual cues and formalization, with a larger emphasis placed on the linguistic sentence structure.

Another interesting linguistic focus in our research is on the presence and effect of part-whole relationships between objects and activities, as mentioned by Nanba et al. on the meronymy of terms in recipes [10]. This is a common linguistic cue found across all three modalities, such as *"Pour remaining buttermilk"*, *"Add rosemary leaves"*, and *"While frying remaining pieces"*. Rosemary leaves are a part of the rosemary herb, and must be removed from the stem; however, this step in the textual modality of the recipe is simply *"gently fry the rosemary"* followed by *"top with fried rosemary leaves"*. Clearly, the textual modality is missing an explicit action that is inferred to be commonsense knowledge or unimportant to the human user, which makes similar tasks difficult to interpret automatically by machines. The background knowledge and expertise required in the domain of cooking is a topic to be considered in the comparison of domain-specific linguistic cues, as well as a more extensive focus on the linguistics behind cooking recipes.

6.2. A More Robust Methodology for Visual Modality

The methodology for formalizing the visual modality has room for improvement; this can be done through leveraging current work done in the visual data classification area. By continuing to develop a more robust semantic representation model for visual data with image processing techniques on cooking instructions, similar to the one discussed in by Feng et al [2]. In Feng et al.'s research, the multimodal approach is considered and if more emphasis is placed on the cooking domain, a more robust methodology can be developed for not only the visual modality, but also the corresponding auditory and textual modalities.

Sun et al. proposed an interesting joint visual-linguistic model, VideoBERT, specifically in the domain of cooking videos on Youtube to learn high-level features in recipes and help classify unlabelled data [15]. Relevant applications include action classification and video captioning, which can aid in the process of transcribing the visual data in cooking videos for the purpose of our multimodal research. However, this simply makes it easier to construct ontologies for each modality, and does not help us understand the relationships between all three modalities.

6.3. Automatic Textual Process Description Generator

Since the approach outlined for textual process descriptions is comprised of a set of rules based on linguistic principles and past literature, there is potential in the area of developing an automatic process description generator for textual recipes.

This can be done by equating the English sentence and recipe structure outlined in Section 3.2, with its respective formalization in PSL. More research can be performed in this area through developing a more robust set of standards for translating English to PSL and classifying more recipes.

One main benefit of developing this automated process description generator is the scale of its impact: numerous textual recipes can be automatically generated with a simple rule-based system without a human translator or any ontology background. Most recipes on the Internet are text-based, and this rule-based system can map onerous amounts of text to their corresponding axioms in PSL to eventually build a fully-fledged cooking recipe ontology. The scalability of this generator transcends more than just the domain of cooking recipes; with the help of domain experts in process-based industries, this generator's modular structure of mapping text to formalized axioms can be applied in a wide array of the aforementioned process-based industries.

6.4. Psychology Experiments

A step further from the current research, composed of qualitative recipe annotations performed by two individuals, can include leading human factors experiments with two groups of human subjects: one group performing the textual recipe and one group describing their steps. This experiment can help distinguish whether or not groups have the same understanding of the recipe. In addition, it would be helpful to increase the number of recipe transcribers available to watch and analyze the visual modality, since the current methodology is limited to an individual perspective and visual cues are less comprehensive than textual or auditory cues. Visual cues depend on the viewer, and have a more ambiguous structure overall; they also take into account a human subject's perspective and expertise levels in the domain. Perhaps in future work, researchers will be able to generate formalized specification of recipes based on image recognition and video recordings of test subjects. A psychology experiment can also be performed to determine whether or not implicit constraints in a recipe modality are explicitly considered, as well as the corresponding impact on its process description. Commonsense reasoning is often implied in the auditory and visual modalities of the recipe; this experiment would enable us to uncover when certain actions are implied, such as steps of waste disposal. Altogether, more insights uncovered through human experiments can help us further comprehend the mismatch between the different modalities, and verify the key takeaways from the results in Section 5.

7. Conclusion

Through an analysis of a recipe across the three modalities of visual, auditory and textual knowledge representation, the following methodologies and semantic-driven insights were uncovered. For the textual and auditory modalities, we introduced and used a procedure to formalize the provided recipe based on linguistic cues and the overall ability to represent them in PSL. The procedure for the visual modality certainly has more room for improvement, with the future aid of image processing techniques and mapping of cooking images to its corresponding activities. With the formalization in place, the sequencing and presence of activities introduced the idea of any-order constraints and alternatives for a recipe, as well as how each process description formalized is simply one occurrence of the cooking recipe's generalized "ground truth". The idea of implicit steps and states mainly found in the textual modality of the recipes displayed a need for more explicit instructions to tie into the future work for automating process description generators and consistency across activities in the cooking domain. Furthermore, the consistency of activity complexity across modalities paired with the inconsistency of activity type, and contradiction between temporal constraints were identified. In the future, we can explore several potential areas to improve the consistency of process descriptions across the three modalities of cooking instructions, and extend our research further in prospective projects such as ontology-based voice assistants, psychology experiments, and a fully-fledged cooking-based ontology.

References

[1] S. Borgo, O. Kutz, F. Loebe, and F. Neuhaus. Preface. In *The Joint Ontology Workshops Episode 3: The Tyrolean Autumn of Ontology*, page 3, Sept 2017.

[2] Yansong Feng and Mirella Lapata. Visual information in semantic representation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 91–99, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[3] Food Network Canada. Lynn crawford's buttermilk fried chicken, 2019.

[4] Michael Grüninger. Using the PSL ontology. *Handbook on Ontologies*, page 423-443, 2009.

[5] Gruninger, M., Shapiro, S., Fox, M.S., and Weppner, H. (2010) Combining RFID with Ontologies to Create Smart Objects, *International Journal of Production Research* 48:2633-2654.

[6] Hakeem, A., Sheikh, Y. Shah, M. (2004) $CASE^{E}$: A Hierarchical Event Representation for the Analysis of Videos, *Nineteenth National Conference on Artificial Intelligence 2004*, 263-268. San Jose, California.

[7] Daniel Jackson. Automating first-order relational logic. *SIGSOFT Softw. Eng. Notes*, 25(6):130–139, November 2000.

[8] Lavee G., Rivlin, E., Rudzsky, M. (2009) Understanding Video Events: A Survey of Methods for Automatic Interpretation of Semantic Occurrences in Video. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 39:489-504.

[9] Jonathan Malmaud, Earl Wagner, Nancy Chang, and Kevin Murphy. Cooking with semantics. pages 33–38, 01 2014.

[10] Hidetsugu Nanba, Yoko Doi, Miho Tsujita, Toshiyuki Takezawa, and Kazutoshi Sumiya. Construction of a cooking ontology from cooking recipes and patents. pages 507–516, 09 2014.

[11] Nevatia, R., Hobbs J., Bolles, B. (2005) VERL: An Ontology Framework for Representing and Annotating Video Events, *IEEE MultiMedia* 12:76-86.

[12] Atalay Özgövde and Michael Grüninger. Foundational process relations in bio-ontologies. pages 243–256, 01 2010.

[13] Ricardo Ribeiro, Fernando Batista, Joana Paulo Pardal, Nuno J. Mamede, and H. Sofia Pinto. Cooking an ontology. *Artificial Intelligence: Methodology, Systems, and Applications Lecture Notes in Computer Science*, page 213-221, 2006.

[14] Sobhani, F. and Straccia, U. (2004) Towards a Forensic Event Ontology to Assist Video Surveillance-based Vandalism Detection. *Proceedings of the 34th Italian Conference on Computational Logic*. Trieste, Italy, June 19-21, 2019.

[15] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. VideoBERT: Joint Model for Video and Language Representation Learning. 04 2019.

[16] Yildirim, Y., Yazici, A. (2006) Ontology-Supported Video Modeling and Retrieval. *International Workshop on Adaptive Multimedia Retrieval*, 28-41.