VERB SEMANTICS AS DENOTING CHANGE OF STATE IN THE PHYSICAL WORLD

By

Malcolm Doering

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Computer Science - Master of Science

2015

ABSTRACT

VERB SEMANTICS AS DENOTING CHANGE OF STATE IN THE PHYSICAL WORLD

By

Malcolm Doering

In the not too distant future we anticipate the existence of robots that will help around the house, in particular, the kitchen. Thus, it is critical that robots can understand the language commonly used within this domain. Therefore, in this work we explore the semantics of verbs that frequently occur when describing cooking activities. Motivated by linguistic theory on the lexical semantics of concrete action verbs and data collected via crowdsourcing, an ontology of the changes of states of the physical world as denoted by concrete action verbs is presented. Furthermore, additional datasets are collected for the purpose of validating the ontology, exploring the effects of context on verbal change of state semantics, and testing the automatic identification of changes of state denoted by verbs. In conclusion, several areas of further investigation are suggested.

ACKNOWLEDGEMENTS

Writing a master's thesis has been a much more difficult undertaking than originally anticipated. I would not have been successful without the help of several people. Foremost, I would like to thank my advisor, Dr. Joyce Y. Chai, for her guidance through the arduous process of formulating a thesis topic and along every step of the way to its completion. I would also like to thank Dr. Rui Fang for taking me under his wing during my time as a master's student, Shaohua Yang for his input regarding visual features and classification results, my thesis committee members Dr. Cristina Schmitt and Dr. Xiaoming Liu for their insightful questions and feedback, Zach Richardson for his annotation efforts, and all the remaining members of the Language and Interaction Research Laboratory who made my time at MSU enjoyable. Lastly, I would like to thank my family and friends for their continued moral support throughout.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction and Motivation

In the future robots will work closely with humans in the home to aid in various domestic tasks. Foremost are tasks in the kitchen. For a robot to help out and learn new abilities in the kitchen domain it must be able to understand a human's instructions. This work focuses in particular on how a robot may represent concrete verbs in the kitchen domain.

Concrete action verbs are verbs that denote concrete activities performed by an agent in the world. These actions are visually perceivable events that can potentially be understood by computer vision algorithms. Furthermore, they can be categorized into two classes based on their semantics: Result Verbs and Manner Verbs [14, 13]. Manner verbs typically denote the Form or the Manner in which the action denoted by the verb is performed, whereas Result verbs (the focus of this work) denote the Change of State (CoS) that the object of the verb undergoes as a result of the action denoted by the verb. In order to ground these verbs to the environment, the robot must have a rich representation of the changes of state associated with the verb. Existing verb resources such as VerbNet [20] do not contain this rich information. In VerbNet, although its semantic representation for various verbs may indicate that a change of state is involved, it does not always provide the specifics associated with the verb's meaning. For example, the change will occur to some attribute of the verb's direct object such as color, number of pieces, speed, etc.

This work presents an ontology of different types of CoS to fill out the VerbNet representation. Specifically, this work is focused on verbs in the cooking domain, with the

forethought that this method can be generalized to other domains. We carried out a series of data collection experiments via crowdsourcing designed to elicit natural language descriptions of changes of state as denoted by various verbs. These descriptions provided insight into how mid-level visual features such as state can be realized linguistically at the surface level and guided the design of the CoS ontology which categorizes different types of change of state.

This data allows us to ask some questions: How does the object of a verb affect the meaning of the verb? I.e., does the verb denote different changes of state depending on the type of direct object? Furthermore, how does the presence or absence of a scene in combination with the verbal description affect CoS which the viewer attributes to the verb? Do ambiguities in the CoS for a single sense of a verb indicate sub-categorizations of verb senses? Answers to these questions are important for a robot whose understanding of a verb is based on its context of use. By determining on-line the CoS indicated by a verb, the robot can focus its sensing resources on the indicated CoS (active sensing).

In the end, I hope that this will be a useful resource for Situated NLP researchers.

Chapter 2

Related Work

The related work can be divided into sections including theoretical linguistics work on the lexical semantics of verbs and computer vision work on recognizing visual attributes of objects and understanding events in videos.

Previous work in theoretical linguistics has defined the types of concrete action verbs we are interested in – mainly Result verbs, which indicate their object's change of state [14, 13]. Kennedy provides a more detailed analysis of gradable predicates in terms of scale structure, which is applicable to Result verbs [6]. And lastly [20] presents digital dictionary of verbs (VerbNet), and their categorizations from [7]. Our work supplements the semantic specifications of some of the verbs contained VerbNet by categorizing the types of changes of state in the world which the verbs may denote. More information on verb semantics is contained in Section 3.

In traditional verb semantics the meaning of a verb may be ambiguous. There is a static number of senses (denoting subtleties of meaning) for a verb, one of which must be selected based on the verb's context of use. Alternatively, Generative Lexicon argues against the traditional fixed number of senses for a word, as they may not apply to novel uses [12]. GL proposes that the meanings of verbs are dependent not only on the lexical specification of the verb itself, but through interactions with the complex semantic representations of its arguments. That is, the properties of nouns affect the meaning of the verb. Thus, "the semantic load is spread evenly across the lexicon". The current work provides examples of how

a verb's meaning may depend on its patient argument, i.e. examples of sub-categorizations of verb senses. Also, we show how visual context affects the human's interpretation of the verb sense in terms of change of state.

Gillette et al. carried out studies of vocabulary learning using the Human Simulation Paradigm [4]. College students were tasked with identifying verbs when presented with different information including the nouns that appeared in the sentence with the verb, the syntactic information of the sentence containing the verb, and extra-linguistic information (i.e., a video which the verb describes). The verb itself was not presented. Experimenters found that verbs with a high degree of 'imigability' or 'concreteness' were more easily identified from the extra-linguistic context, whereas syntax was a more useful cue for abstract verbs. This suggests that the visual context may be important to identifying the changes of state in the world which these concrete verbs denote. In our work we focus primarily on verbs from the cooking domain with a high degree of imagability (e.g., *cut*, *rinse*, etc.).

Visual attributes are high level visual features of (objects in) scenes which have corresponding natural language descriptions. For example, 'green' refers to a specific range of RGB values. Visual attributes are semantically meaningful, discriminative, and generalizable across different object types and can be used for object recognition [11, 22, 27]. Attributes roughly correspond to adjectives (describing states – colors, properties, etc. – of objects), but can also describe other properties not named by adjectives (e.g. 'has wing' for birds); therefore, the semantics of a CoS verb may be grounded in changes of an object's visual attributes.

Chao et al. jointly models action and object categories made up of the synsets from WordNet [2, 10]. Specifically, they modeled the affordances of the object categories, which indicate the functions of objects, in order to improve action recognition. Whereas Chao et al. model the interactions between actions and objects, in this work we are interested in the interactions between verbs and direct objects (often nouns) and how the object affects the meaning of the verb. A verb/noun is different from an action/object because different senses

of the verb/noun may appear in separate action/object categories.

Siskind et al. demonstrate how an action in a video can be recognized automatically by using principles and constraints of human perception [24, 23]. Actions are represented via temporal predicate logic. Furthermore, they demonstrate that actions may be defined in terms of *relations* (e.g., support, contact, and attachment) between the entities rather than low level *kinematic representations* (e.g., joint angles and velocities). Beyond relations, our work explores *changes of state*, another high level feature integral to action representation/verb meaning.

Various methods for generating natural language descriptions of images and videos are presented by [9, 19, 28]

Chapter 3

Change of State in Verb Semantics

Lexical semantics is important to designing methods for robots to learn verbs because it indicates what must be learned as part of the verb representation. Verbs can be divided into two broad categories: stative verbs that denote states (such as *know*, *depend*, *loathe*) and action verbs which denote actions (such as *run*, *throw*, *cook*) In this work we are primarily interested in the latter.

A concrete action verb is one that, in combination with its arguments and modifiers, denotes an action in the world (as opposed to denoting a state or an abstract action not visible in the world). Hovav and Levin [14, 13] further divide the types of action verbs into *Manner* verbs, which "specify as part of their meaning a manner of carrying out an action", and *Result* verbs, which "specify the coming about of a result state". For example,

- **Manner verbs**: nibble, rub, scribble, sweep, flutter, laugh, run, swim...
- **Result verbs**: clean, cover, empty, fill, freeze, kill, melt, open, arrive, die, enter, faint...

In this work we focus specifically on result verbs, i.e. verbs of Change of State (CoS). A set of "canonical realization rules" specify how a particular change of state is incorporated into a verb's semantics. Semantics are determined based on the combination of a "root", which is particular to the verb (e.g., a result-state), and an "event schema" template as shown in Figure 3.1.

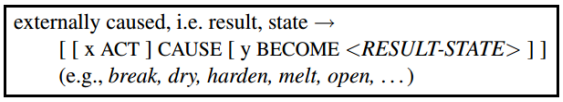


Figure 3.1: Event schema for verbs that denote externally caused state changes [14]

Previous work has further classified result verbs into three categories: Change of State verbs, which denote a change of state to a property of the verb's object (e.g. 'to warm'), Inherently Directed Motion verbs, which denote movement along a path in relation to a landmark object (e.g. 'to arrive'), and Incremental Theme Verbs, which denote the incremental change of volume or area of the object (e.g. 'to eat') [8]. In our work we propose a specific set of result-states that may be used to define the semantics of most concrete action verbs in the kitchen domain. Note that we use the term Change of State in a more general way throughout this paper such that the location and volume or area of an object are part of its state.

Hovav and Levin also claim that some verbs that denote change of state events lexically specify a scale [14, 13]. I.e., they are verbs of scalar change. A scale is "a set of points on a particular dimension (e.g. height, temperature, cost)". In the case of verbs, the dimension is an attribute of the object of the verb. For example, "John cooled the coffee" means that the temperature attribute of the object *coffee* has decreased. Kennedy and McNally give a very detailed description of scale structure and its variations [6]. Our work simplifies scale structure by representing change of state as a frame consisting of an object, one of its attributes, and the resulting value of the attribute.

Verbs of scalar change can be further split into two categories verbs with two points on the scale and verbs with multiple points on the scale [14]. Verbs on a two point scale are verbs of achievement, where the change of state is "conceptualized as instantaneous" (pg. 30). Examples of these verbs are 'crack' and 'arrive'. Verbs on a multiple point scale are associated with changes in attributes that can have multiple values. Examples of these verbs are 'advance', 'descend', 'fall', 'recede', 'rise', 'warm', 'cool', etc. These are called verbs of gradual change or degree achievement. In this work these subtle differences are not taken

into account, but they may be worth including in future work.

Manner verbs are verbs of non-scalar change. They are different from verbs of scalar change because they are more complex – they involve change in more than a single attribute or do not specify change in a single direction on a scale [14]. They may involve a combination of multiple changes. Some examples are 'walk' and 'jog', which specify a sequence of changes on several attributes. Additionally, verbs of non-scalar change do not always have to be specific about what changes are involved. E.g. exercise may denote any of several varieties of physical (and sometimes mental) activity (pg.33). We will not focus on manner verbs in this work.

Beavers and Koontz-Garboden introduce a set of diagnostics for judging whether a verb is a manner verb or result verb [1]. They also argues against Rappaport Hovav and Levin's claim that a verb cannot lexicalize both manner and result [14, 13]. Indeed, from our own observations it would seem that a verb can have both change of state and manner components (e.g. 'chop').

Levin categorizes verbs based on their syntax, i.e. based on which alternations (argument structure) a verb can take [7]. This categorization scheme resulted in classes of verbs with semantic similarities. VerbNet is a digital resource based on Levin's verb classes [20]. For each verb, it provides the possible argument structures and a logical representation of their semantics. This has been a useful resource for NLP researchers in the past. But, in the case of result and manner verbs that denote physically observable events, it does not provide enough detail to enable a robot to ground the verb in its perceptions. For example, the VerbNet semantic representation of 'cut' specifies only that there some 'degradation of the material integrity' of the object as a result of the action.

Chapter 4

A Pilot Study and Ontology of Change of State

4.1 Crowdsource Setup

In order to build an ontology of the types of changes of state that verbs in the cooking domain may denote, we conducted a pilot study. Verb-object pairs were presented to turkers via Amazon Mechanical Turk (AMT) and turkers were asked to describe the changes of state that occur to the object as a result of the verb. Then the turker's open-ended descriptions were analyzed and categorized.

Verbs and objects from the TACoS corpus were chosen for this crowdsourcing study. The TACoS corpus [17] is a collection of natural language descriptions of the actions that occur in a set of cooking videos. I.e., it contains 18227 sentences collected via AMT that describe various cooking events (preparing a cucumber, scrambling eggs, etc.). This is an ideal corpus to explore the types of changes of state and manners that verbs denote since it contains mainly descriptions of concrete actions. Moreover, possibly because most actions in the cooking domain are goal-directed, a majority of the verbs in the descriptions denote results of action (changes of state).

The ten verbs (shown in Table 4.1) were chosen based on the criteria that they take an agent argument and that they occur relatively frequently in the corpus and with a variety of different direct objects. Furthermore, they must be concrete, meaning that they denote

Verb	Object 1	Object 2	Object 3
clean	cutting board	dishes	counter
rinse	cutting board	dishes	ginger
wipe	counter	knife	hands
cut	cucumber	beans	leek
chop	cucumber	beans	leek
mix	eggs	leeks, salt, and pepper	ingredients
stir	eggs	leeks, salt, and pepper	ingredients
add	water	eggs	leeks
open	bread packaging	drawer	pomegranate
shake	spices	bowl	broccoli

Table 4.1: Verbs and objects used for data collection (pilot)

some observable event in the world. Verbs of this type are the most relevant for a kitchen robot. Lastly, five of the verbs were chosen because they only denote a change of state, and the other five were chosen because they denote some manner of action (possibly in addition to change of state).

To examine how CoS depends on the context, we paired verbs with different objects (3 objects per verb, shown in Table 4.1) and presented the verbs to turkers with and without a video of the action described by the verb (+/-scene). Objects were chosen based on the criteria that they are dissimilar to each other, since we hypothesize that the change of state indicated by the verb will differ depending on the object's features. For example, *broccoli* and *bowl* were chosen as objects for the verb *shake* because one is a vegetable and the other a kitchen utensil, having very different features. Thus, there are $10 \times 3 \times 2 = 60$ conditions. For each condition we collected 30 turker responses.

In addition to turkers responses about what changes of state the verbs indicated, we also collected responses about the manner of the action.

4.2 Ontology Design and Annotation

Based on the types of descriptions the turkers provided we developed an ontology of change of state, which is show in Table 4.2. This section will explain this ontology in detail.

18 attributes of state changes were identified from the change of state descriptions

provided by the turkers. Because we are interested in lower level state changes, which can easily sensed with a camera and computer vision algorithm, we did not include higher level attributes in the categorization such as *Cleanliness*, which may be identified in terms of lower level attributes but are more difficult to sense automatically. Although, note that some of these attributes are at different levels in terms of visual perception, e.g. *Shape* vs. *Wetness*. Wetness must be visually identified in terms of some lower level features such as color, whereas shape is directly perceivable. Some examples of descriptions containing these attribute are shown below. It is also worth knowing that some descriptions actually describe multiple changes of state as shown in Figure 4.1.

Given that both adjectives and CoS verbs have their semantics defined in terms of a scale structure (for gradable verbs and adjectives), some of the above attributes are motivated by the semantic types of adjectives from Dixon and Aikhenvald's categorization [3]. These adjective categories include Dimension, Color, Physical Property, Quantification, and Position.

Although the instructions given to the turkers specifically requested a description of the change of state undergone by the *object*, several responses contained descriptions of changes of other objects. Often, a part of the direct object was described, rather than the whole object. And, sometimes some completely different object, that was still associated in the action, was described. Thus, CoS descriptions can be categorized as describing a change to the DirectObject, PartOfObject, or AssociatedObject. Some examples are shown below.

DirectObject

Cut-cucumber: "The size of **the cucumber** changes"

PartOfObject

Wipe-knife: "The knife gets cleaner. More **metal** is showing"

AssociatedObject

Clean-dishes: "**Debris and residue** fall away from the dishes"

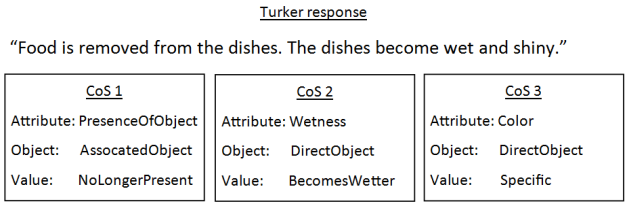


Figure 4.1: Example of CoS frame applied to a description of change of state

In addition to attribute of change and the object undergoing change, the turkers descriptions often contained a third important aspect of a change of state: the result value. I.e. the result value of the attribute after it changes. These values can be categorized in several different ways depending on the attribute, but generally there are two polar values. For example, the SizeLengthVolumeThickness, Wetness, NumberOfPieces, etc. attribute may *Increase* or *Ddecrease* in value. On the other hand, not all result values can be categorized in this way. For example, the Shape attribute is usually described simply as having changed in some vague way, or to have undergone a specific change. Thus, *Specific* and *Change* are two more general result values.

These three aspects of a change of state, the *Attribute*, *Object*, and result *Value*, make up the *CoS frame* which can be used to label a verb-object pair, as shown in Figure 4.1 for a sentence which describes three changes of state. Thus, the CoS ontology presented here consists of a CoS frame and the options used to fill the frame slots.

4.3 Analysis

4.3.1 Types of CoS Descriptions

The data shows that generally turkers described changes of state in one of three ways. (1) They describe the attribute directly, e.g. cut-cucumber: "The *size* of the cucumber changes". (2) They describe the change of state with a resultative phrase, e.g. cut-cucumber: "The cucumber is cut *into small pieces*." Here, the CoS is indicated by the semantics of *small pieces*. And, (3) the CoS can be described with another verb that denotes the same or similar CoS as the verb presented to the turker, e.g. stir-ingredients: "The ingredients are

Type of CoS	Attribute	Attribute Result Value
Dimension	Size, length, volume, thickness	Changes, increases, decreases, specific
	Shape	Changes, specific (cylindrical, flat, etc.)
Color/Texture	Color	Appear, disappear, changes, mix, separate, specific (becomes green, red, etc.)
	Texture	Changes, specific (slippery, frothy, bubbly , soft, etc.)
Physical Property	Weight	Increase, decrease
	Flavor, smell	Changes, intensifies, specific
	Solidity	Liquefies, solidifies, specific (paste, soggy, etc.)
	Wetness	Becomes wet(ter), dry(er)
	Visibility	Appears, disappears
	Temperature	Increases, decreases
	Containment	Becomes filled, emptied, hollow
	Surface Integrity	A hole or opening appears
Quantification	Number of pieces	Increases, one becomes many, decreases, many becomes one
Position	Location	Changes, enter/exit container, specific
	Occlusion	Becomes covered, uncovered
	Attachment	Becomes detached
	Presence	No longer present, becomes present

Table 4.2: Attributes and result values for change of state

mixed together".

4.3.2 Multiple CoS Labels per Verb

A turker's change of state description does not necessarily only contain a single change of state. In fact, all the descriptions described between 0 and 3 changes of state, as seen in Figure 4.2. Most descriptions (43%) contained only a single change of state. Also, a large percentage (36%) contained no change of state. In actuality, some of the descriptions that were annotated as containing no change of state, described changes of state with high level attributes which do not fit into our categories (e.g. Cleanliness). Others contained verbal descriptions of CoS (e.g. Stir-ingredients: "The ingredients are mixed together."). We did not annotate descriptions which contained these circular definitions.

For each verb, we calculated the distribution of CoS annotations over each attribute. Figure 4.3 shows the CoS attribute distributions for two verbs, *clean* and *rinse*. The CoS

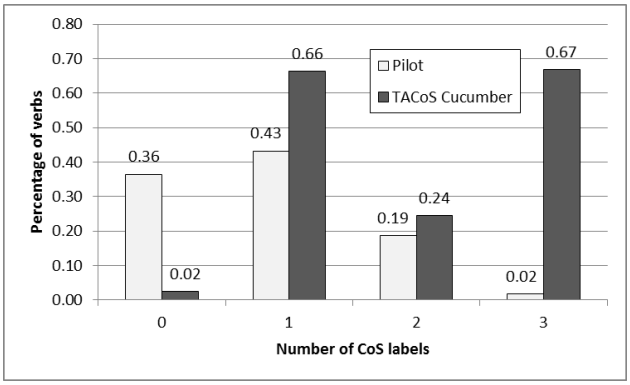


Figure 4.2: Percentage of samples describing 0 to 3 changes of state (pilot and cucumber datasets)

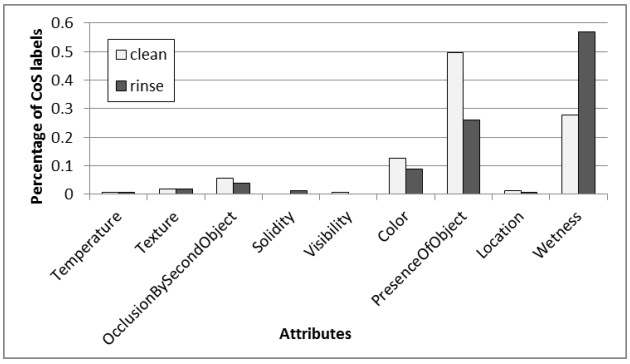


Figure 4.3: CoS distributions over attributes for *clean* and *rinse* (pilot dataset)

distributions are closely related to the semantics of the verbs they label. For example, CoS labels with the attributes Wetness and PresenceOfObject (referring to dirt that is removed) are more frequent for the verbs *clean* and *rinse* than CoS labels with other attributes. This is because the semantics of these verbs indicate some object is cleaned away, possibly with water. Notice that *clean*, the result verb, has a much lower frequency of the Wetness attribute (which is related more to the manner of cleaning) and a higher frequency of PresenceOfObject (which is related to the intended result). On the other hand, the manner verb *rinse* has these distributions the other way around.

Another observation regarding the CoS distributions in the pilot dataset is that not all the descriptions describe the same attributes. For example, for *clean*, most of the descriptions describe Wetness and PresenceOfObject, but there is also some distribution over the attributes Texture, OcclusionBySecondObject, Color, etc. One reason this may happen is

because when a verb-object pair is presented to a turker without an accompanying scene, the turker may rely more on their imagination when describing the change of state, whereas when the scene is shown, they can see the change directly. For example, if the verb object pair is *shake broccoli* but the turker does not see that the broccoli is covered with water, they will not describe the water droplets that come off as it is shaken. Moreover, even in conditions when the scene is shown, the turkers may describe the same change of state in different ways resulting in different CoS annotations. For example, *clean dishes*: "Food is removed from the dishes" describes the PresenceOfObject attribute of the AssociatedObject, while *clean dishes*: "Dish surface is cleared of debris and/or muck." describes the OcclusionBySecondObject attribute of the DirectObject.

4.3.3 The Role of Visual Context

To determine how the presence of a scene or how the object of a verb affects the types of CoS turkers describe in their responses, we defined a Jensen-Shannon divergence based metric *variability*. The JSD of two distributions P and Q is given by the formula below.

$$JSD(P||Q) = \frac{1}{2}D(P||M) + \frac{1}{2}D(Q||M) \tag{4.1}$$

$$\text{where } M = \frac{1}{2}(P + Q) \tag{4.2}$$

D is the Kullback-Leibler divergence, a non-symmetric measure of the difference between two distributions.

$$D(P||Q) = \sum_i P(i) \ln \frac{P(i)}{Q(i)} \tag{4.3}$$

The advantage of using JSD is that it is a symmetric measure. It shows similarity between two distributions, equaling 0 when the distributions are the same and approaching 1 as they become more different.

Verb	+/-Scene	3 Objects	Objects +scene	Objects -scene
clean	0.03	0.04	0.08	0.05
rinse	0.01	0.05	0.09	0.07
wipe	0.02	0.14	0.15	0.19
cut	0.01	0.02	0.01	0.05
chop	0.02	0.03	0.03	0.04
mix	0.05	0.13	0.15	0.17
stir	0.09	0.21	0.24	0.25
add	0.12	0.22	0.19	0.33
open	0.09	0.32	0.34	0.41
shake	0.18	0.42	0.42	0.43

Table 4.3: Variability between CoS frequencies per object and scene conditions (pilot dataset)

The variability describes how the CoS distribution of a verb differ depending on a certain variable (+/-scene or object). We compute the variability by averaging the sum of JSD for each pair of CoS distributions of the verb, where the distributions of each pair are taken over different values of the variable. For example, the variability of the verb *shake* over the scene conditions is found by dividing the JSD of the CoS distributions in the +scene and -scene conditions by 1 (the number of pairs of conditions). Moreover, the variability over the object conditions is found by summing the JSD of the CoS distributions for each pair of object conditions (three unique pairs), and dividing by 3. The general variability formula is shown in Equation 4.4.

$$variability = \frac{\sum_{distr\ pairs\ (i,j)} JSD(d_i, d_j)}{num\ pairs} \tag{4.4}$$

The variabilities between various conditions for each verb in the pilot dataset are shown in Table 4.3. The variability metric shows that there is indeed some difference between the CoS distributions in the +scene and -scene conditions. Moreover, the variability is much higher for some verbs (shake 0.18, add 0.12) than others (cut 0.01, rinse 0.01). This may be because for verbs like shake and add, without the accompanying scene it is not clear how the state of the object will change.

Verb	Label card.	+Scene	-Scene
add	0.64	0.78	0.52
chop	1.46	1.44	1.48
clean	0.88	0.99	0.78
cut	1.42	1.41	1.43
mix	0.71	0.68	0.76
open	0.82	0.91	0.73
rinse	0.88	0.89	0.87
shake	0.52	0.53	0.53
stir	0.52	0.61	0.47
wipe	0.65	0.78	0.53

Table 4.4: Label cardinality per verb (pilot dataset)

Table 4.4 shows the average number of CoS labels for each of the turkers' descriptions. For most verbs (6 of 10), more changes of state are described when the scene is shown, indicating that the scene presents more information about CoS to the turker for these verbs. Taken together this data shows that visual context is important to determine the change of state denoted by a verb.

4.3.4 Effects of Direct Object on CoS

Table 4.3 also shows the variability for each verb in the pilot dataset computed over the three object conditions. The variability among objects is different for each verb, showing that for some verbs in this kitchen domain the CoS depends more on the object of the verb than for others. The verb with the highest variability is again shake (0.42). The resulting state change from shaking a (wet) piece of broccoli is very different than shaking a container of spices over food, or a bowl filled with eggs. This shows that even though the verb sense is the same for in all these descriptions, the CoS indicated by the verb may depend on the object of the verb.

4.3.5 Verb Semantic Similarity based on CoS

To compare the CoS distributions between each pair of verbs in the pilot dataset we computed the Jensen-Shannon divergence of each pair. Table 4.5 shows that the distributions for verbs from the pilot data are very similar for verbs with similar se-

Verb pair	JSD	Verb pair	JSD	Verb pair	JSD
cut-chop	0.01	mix-shake	0.36	rinse-stir	0.45
mix-stir	0.03	chop-stir	0.37	rinse-mix	0.45
rinse-wipe	0.04	cut-stir	0.39	chop-add	0.47
clean-rinse	0.05	wipe-add	0.39	cut-add	0.48
clean-wipe	0.06	clean-shake	0.39	wipe-open	0.51
add-shake	0.11	chop-open	0.40	clean-open	0.53
stir-add	0.20	chop-mix	0.42	rinse-open	0.57
add-open	0.23	rinse-add	0.42	chop-shake	0.59
mix-add	0.27	wipe-stir	0.42	cut-shake	0.59
open-shake	0.30	cut-open	0.43	wipe-chop	0.67
wipe-shake	0.31	cut-mix	0.43	clean-chop	0.67
stir-shake	0.32	clean-mix	0.43	wipe-cut	0.68
stir-open	0.32	clean-stir	0.43	clean-cut	0.68
rinse-shake	0.33	wipe-mix	0.44	rinse-cut	0.68
mix-open	0.34	clean-add	0.45	rinse-chop	0.68

Table 4.5: Jensen-Shannon divergence between CoS distributions for verb pairs (pilot dataset)

mantics (e.g. JSD(cut,chop)=0.01 and JSD(mix,stir)=0.03 vs. JSD(cut,shake)=0.59 and JSD(rinse,chop)=0.68). This shows that the CoS frame is capturing relevant semantic information.

Chapter 5

Automated Identification of CoS

5.1 The Cucumber Dataset

To discover which features are important for predicting CoS labels we collected a second dataset by automatically identifying 553 verbs from the TACoS corpus manually annotated them with CoS frame labels. In particular, we collected all the verbs within sentences describing the cucumber preparation activity (slicing a cucumber and placing it on a plate). A student with no previous knowledge of the project was recruited and trained to label each verb with up to three CoS frames. They were shown the verb, its patient (both automatically identified), and the original sentence (e.g.,"The person **chops** the **cucumber** into slices on the cutting board"). Then they were tasked with annotating the change of state that occurred to the patient as a result of the verb by choosing options from the CoS ontology to fill up to three CoS frames. If the change of state was not clear from this information, the student could view the videos.

Figure 4.2 shows the percentage of the TACoS sentences that received 0 to 3 CoS labels. Compared to the crowdsourced data almost all of these verbs received at least one CoS annotation. This is because in the pilot study some of the turkers' responses did not describe changes of state and some of their descriptions did not fit into our CoS categories.

Table 5.1 shows the number of tokens of the top ten most frequent verbs from the TACoS cucumber video descriptions, their average number of CoS labels, and the number

Verb	Count	Label card.	Num obj.
get	106	1.44	17
take	88	1.51	15
wash	58	1.10	9
cut	56	1.09	14
rinse	50	1.36	9
slice	30	1.07	17
place	30	1.07	16
peel	24	2.67	10
put	21	1.00	8
remove	16	2.31	10

Table 5.1: Verb counts and label cardinality of top ten most frequent verbs (cucumber dataset)

of different objects (e.g. 'cucumber', 'bowl', etc.) they occurred with. The average number of object types that each verb takes is 12.46.

5.2 Multilabel Classification

To explore the effect of different feature sets on CoS prediction, we formulated the problem as a multilabel classification problem. The goal of a typical classification problem (i.e. supervised learning with discrete labels) is to learn a classifier that can predict the correct label for a sample given a vector of feature values that represent a sample. A set of N samples $(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)$, where x_i is the vector of feature values for sample i and y_i is sample i's label, is used to train and test a classification model. In a multilabel classification problem, y_i is a *set* of labels rather than a single label [25, 15].

In this particular problem a sample i consists of a single verb-object pair and its CoS frame annotations. The feature vector x_i (e.g., linguistic features of the sentence containing the verb or visual features of the video it describes) is extracted from the sample. The label set y_i consists of the Attributes of the annotations (for example, NumberOfPieces, Wetness, etc.). Note that although we collected Object and result Value information for each verb, we did not include it as part of the prediction problem.

We explored two methods of multilabel classification, Binary Relevance and Label Com-

bination. Binary relevance works by training 18 separate classifiers (one for each attribute) and then applying them independently to predict the labels for each sample. Thus, correlations between attributes are not taken into account with this method. In contrast, the label combination method takes into account correlations between the labels by treating each unique label set as a single label/class and applying traditional classification techniques.

Because multilabel classification is a significantly different problem from traditional classification, special metrics are used for evaluation. Three common metrics for multilabel classification are *Exact Match* [26, 15], *Hamming Score* [5, 26, 15], and *Jaccard Index* [26, 15, 16]. Exact match is an example-based metric of accuracy computed by the number of label sets that are correct divided by the number of samples. This is a strict metric because "a single false positive or false negative label makes the [sample] incorrect" [15]. Furthermore, Hamming score is a label-based metric of accuracy. It is computed by averaging the accuracy for each label calculated independently. This is considered a lax metric because it "tends to be overly lenient due to the typical sparsity of labels in multi-label data," rewarding highly conservative prediction [15]. Lastly, Jaccard index provides an accuracy metric between Exact match and Hamming score in terms of strictness. It is a "ratio of the size of the union and intersection of the predicted and actual labels sets, for each example and averaged over the number of examples" [15].

5.3 Results

5.3.1 Linguistic and Visual Features

The first linguistic feature set, bag of words (BoW), is extracted from the sentence containing the verb-object pair. In the case of data from the pilot study BoW features were extracted from the open-ended descriptions of CoS. This feature set contains the lemmas of each word in the sentence concatenated to its part-of-speech (e.g., 'cut-verb', 'cucumber-noun', etc.). Thus some syntactic information is contained in the BoW feature set. The set is then converted into a vector of 1s and 0s, representing whether or not each PoS-lemma

combination is present in the sentence. This procedure yielded 264 features for samples from the cucumber dataset and 1159 features for samples from the pilot dataset.

The verb+object feature set (VO) is the second linguistic feature set. The verb+object feature set consists of a binary representation of the lemmatized verb and object of each sample's verb-object pair.

The third feature set is visual rather than linguistic. Before we extract the visual features, we make the assumption that the ground truth correspondence of the verb and object in the video is known. The visual features are extracted from the video clip described by the verb and include the following.

1. Difference in area of object at the beginning and end of video clip
2. Distance between start and end location of the object
3. Difference in color (euclidean distance) of the object between the start and end of the video clip
4. Difference in texture (euclidean distance between HoG features) of the object at the start and end of the video clip
5. Difference in the object's moments of inertia at the start and end of the video clip – this may indicate change of orientation
6. Whether or not the object was occluded at the beginning and end of the video clip
7. Whether or not the object was present in the scene at the beginning and end of the video clip

5.3.2 Comparison of Features on Cucumber Dataset

Three different feature sets and their combinations were tried. The goal is to evaluate how important different linguistic and visual features are to the identification of change of state.

	Jaccard index	Hamming score	Exact match
DT+BR,BoW	0.601 +/- 0.121	0.951 +/- 0.019	0.363 +/- 0.259
DT+BR,BoW+VO	0.612 +/- 0.128	0.952 +/- 0.019	0.372 +/- 0.271
DT+LC,BoW	0.752 +/- 0.054	0.968 +/- 0.009	0.696 +/- 0.066
DT+LC,BoW+VO	0.855 +/- 0.048	0.983 +/- 0.007	0.790 +/- 0.067
LR+BR,BoW	0.698 +/- 0.052	0.965 +/- 0.006	0.602 +/- 0.045
LR+BR,BoW+VO	0.778 +/- 0.045	0.976 +/- 0.007	0.694 +/- 0.048
LR+LC,BoW	0.775 +/- 0.038	0.971 +/- 0.005	0.720 +/- 0.039
LR+LC,BoW+VO	0.854 +/- 0.021	0.983 +/- 0.003	0.801 +/- 0.028
BL	0.868	0.988	0.790

Table 5.2: CoS attribute classification accuracy (cucumber dataset)

Initially only the linguistic features were used in combination with two types of classifiers, decision tree (DT) and logistic regression (LR), and the two methods of multilabel classification BR and LC, described in Section 5.2. The results for the cucumber dataset are shown in Table 5.2. The data was randomly split into 80% for training and 20% for testing. Five replicates of the experiment were done and their accuracy measurements averaged. The baseline (BL) predicts each verb to have the majority label set for that verb, i.e. it predicts label sets based on the verb.

Table 5.3 shows the prediction results for the cucumber dataset using all combinations of the two linguistic and one visual feature sets. The data was randomly split into 60% for training and 40% for testing. Logistic regression with l1 regularization was used for classification. The scores show the average of five replicates and the standard deviations.

The results show that only the BoW combined with the VO feature set, as well as all three features sets combined, performed better than the baseline (exact match 0.8486 and 0.8495 vs. 0.790). The visual features alone do not perform nearly as well as the baseline (0.3667 vs. 0.790); however, they do perform better than random assignment of label sets (0.3667 vs. 0.0532), showing that they do contain some useful information related to the attributes undergoing change.

Table 5.4 shows per verb classification for the three most frequent verbs. Predictions were made using logistic regression with l1 regularization with 60% of the data used for train-

	Exact match	Jaccard index	Hamming score
BL	0.790	0.868	0.984
Random assignment of label sets	0.0532 +/- 0.0125	0.1319 +/- 0.0093	0.8648 +/- 0.0029
BoW	0.7369 +/- 0.0190	0.7833 +/- 0.0164	0.9736 +/- 0.0015
VO	0.7595 +/- 0.0259	0.8393 +/- 0.0222	0.9819 +/- 0.0024
BoW+VO	0.8486 +/- 0.0225	0.8844 +/- 0.0202	0.9870 +/- 0.0022
Visual	0.3667 +/- 0.0301	0.4480 +/- 0.031	0.9306 +/- 0.0042
BoW+Visual	0.7405 +/- 0.0244	0.7868 +/- 0.0201	0.9742 +/- 0.002
VO+Visual	0.7450 +/- 0.0234	0.8328 +/- 0.0202	0.9813 +/- 0.0021
BoW+VO+Visual	0.8495 +/- 0.0218	0.8847 +/- 0.0195	0.9871 +/- 0.0021

Table 5.3: CoS attribute classification accuracy using various feature sets (cucumber dataset)

ing and 40% for testing. The label combination method was used for multilabel classification. The scores show the average of five replicates and the standard deviations.

The results show that the features sets that give the best performance depend on the verb. The verb *get* has the best exact match score with the visual and BoW features (0.819 vs. baseline 0.642). Adding the VO features offers no further improvement. *Take* has the best performance with only BoW features (exact match 0.872 vs. baseline 0.439). Adding the VO features, visual features, or both to the BoW offers no further improvement. Lastly, *cut* has the best performance with the BoW and VO features combined (exact match 0.791 vs. baseline 0.687). Moreover, when the visual features are used in addition to this feature set the performance for this verb actually decreases down to exact match 0.783. Overall, these results show that the most important features for predicting the CoS of a verb may depend on the verb, rather than there being a single best feature set for all.

5.3.3 Predicting the Attribute Described by NL Descriptions of CoS

We classified the CoS descriptions from the pilot data in order to determine which features are important in predicting the change of state. Note that this dataset provides a distinct problem from the other three datasets. In the case of the pilot data, the turker has provided a direct description of a change of state, from which the features for classification are extracted. On the other hand, for the cucumber dataset the features are extracted from

		get	take	cut
Num examples		106	88	56
Num unique label sets		3	5	5
BL	EM	0.642 +/- 0.062	0.439 +/- 0.069	0.687 +/- 0.084
	JI	0.809 +/- 0.004	0.700 +/- 0.003	0.748 +/- 0.007
	HS	0.979 +/- 0.004	0.965 +/- 0.003	0.974 +/- 0.007
BoW	EM	0.814 +/- 0.015	0.872 +/- 0.065	0.774 +/- 0.058
	JI	0.900 +/- 0.001	0.926 +/- 0.004	0.830 +/- 0.007
	HS	0.989 +/- 0.001	0.991 +/- 0.004	0.983 +/- 0.007
VO	EM	0.619 +/- 0.035	0.628 +/- 0.084	0.757 +/- 0.052
	JI	0.798 +/- 0.003	0.792 +/- 0.006	0.817 +/- 0.004
	HS	0.978 +/- 0.003	0.975 +/- 0.006	0.982 +/- 0.004
BoW+VO	EM	0.800 +/- 0.019	0.872 +/- 0.065	0.791 +/- 0.051
	JI	0.893 +/- 0.001	0.926 +/- 0.004	0.848 +/- 0.006
	HS	0.988 +/- 0.001	0.991 +/- 0.004	0.985 +/- 0.006
Visual	EM	0.600 +/- 0.043	0.472 +/- 0.039	0.757 +/- 0.09
	JI	0.788 +/- 0.003	0.714 +/- 0.002	0.817 +/- 0.007
	HS	0.976 +/- 0.003	0.967 +/- 0.002	0.982 +/- 0.007
BoW+Visual	EM	0.819 +/- 0.023	0.872 +/- 0.065	0.774 +/- 0.064
	JI	0.902 +/- 0.002	0.926 +/- 0.004	0.830 +/- 0.007
	HS	0.989 +/- 0.002	0.991 +/- 0.004	0.983 +/- 0.007
VO+Visual	EM	0.605 +/- 0.044	0.617 +/- 0.121	0.765 +/- 0.065
	JI	0.793 +/- 0.003	0.789 +/- 0.007	0.826 +/- 0.005
	HS	0.977 +/- 0.003	0.975 +/- 0.007	0.983 +/- 0.005
BoW+VO+Visual	EM	0.819 +/- 0.023	0.872 +/- 0.065	0.783 +/- 0.073
	JI	0.902 +/- 0.002	0.926 +/- 0.004	0.839 +/- 0.007
	HS	0.989 +/- 0.002	0.991 +/- 0.004	0.984 +/- 0.007

Table 5.4: Per verb CoS attribute classification accuracy using various feature sets (cucumber dataset)

sentences which describe the action (rather than change of state). If a robot is not able to understand the CoS from a human's narration of an action in the kitchen, then it should be able to ask what CoS is indicated and subsequently extract the CoS from the human's description.

Table 5.5 shows the results for classification of the pilot data using all combinations of the linguistic features sets with two types of classifiers, decision tree (DT) and logistic regression (LR), and the two methods of multilabel classification BR and LC, described in Section 5.2. The data was randomly split into 80% for training and 20% for testing. Five

	Jaccard index	Hamming score	Exact match
DT+BR,BoW	0.426 +/- 0.197	0.949 +/- 0.019	0.271 +/- 0.256
DT+BR,BoW+VO	0.429 +/- 0.200	0.949 +/- 0.019	0.273 +/- 0.259
DT+LC,BoW	0.726 +/- 0.021	0.976 +/- 0.002	0.666 +/- 0.026
DT+LC,BoW+VO	0.732 +/- 0.022	0.977 +/- 0.002	0.674 +/- 0.022
LR+BR,BoW	0.691 +/- 0.027	0.970 +/- 0.002	0.624 +/- 0.025
LR+BR,BoW+VO	0.610 +/- 0.037	0.961 +/- 0.003	0.540 +/- 0.037
LR+LC,BoW	0.738	0.983	0.684
LR+LC,BoW+VO	0.743	0.983	0.689
BL	0.486	0.973	0.429

Table 5.5: CoS attribute classification accuracy for open ended change of state descriptions (pilot dataset)

replicates of the experiment were done and their accuracy measurements averaged. The baseline (BL) predicts each verb to have the majority label set for that verb, i.e. it predicts label sets based on the verb.

The results show that the simple feature set bag of words (BoW) was enough for performance above the baseline (BL); although, when the verb and object (VO) were also provided as features accuracy increased. Furthermore, the results show two interesting things. (1) LC, the multilabel classification method that takes into account correlations between labels, performs better than BR, which does not take correlations into consideration. This is in line with our observation that some attributes are correlated. And (2) in comparison to the results for the cucumber dataset the classification accuracy for the pilot dataset is not as high. We might expect the oposite since the pilot sentences are direct descriptions of the CoS. This may be because of the large number of sentences from the pilot dataset are annotated as describing no CoS resulting in less examples to learn from. Moreover, there is much more variation in the CoS descriptions (pilot data) because some turkers gave thorough, full sentence responses while other only responded with one or two words.

Chapter 6

Complexity of Verb Semantics based on CoS

6.1 Multilevel Dataset and Crowdsource Study

The TACoS Multilevel corpus consists of descriptions of cooking videos (the same videos as from the original TACoS corpus) at three levels of detail, including single sentence, short (about five sentences), and detailed descriptions (no more than 15 sentences) [18, 21]. The corpus contains 20 triples of descriptions for each video and there are five videos per activity. An example of descriptions for the cucumber preparation activity is shown below. (Note that these sentences were processed to have the same subject, etc.)

Single sentence

"The person entered the kitchen and sliced a cucumber"

Short

"The person walked into the kitchen. The person got a cutting board, knife, and cucumber. The person washed the cucumber. The person put the cucumber on the cutting board. The person sliced the cucumber. The person put the cucumber on the plate."

Detailed

"The person walked into the kitchen. The person removed a cutting board and knife

from the drawer. The person put the plate on the counter. The person washed the cucumber at the sink. The person placed the cucumber and the plate next to the cutting board..."

We used this corpus to examine several questions about how the level of detail of an activity affects the CoS denoted by verbs in the description.

6.1.1 Bread and Cucumber Multilevel Dataset

To examine the effects of visual context we collected CoS annotations for the sentences describing the bread and cucumber activities of the TACoS multilevel corpus. Note that these sentences are distinct from the cucumber dataset from Section 5.1. CoS annotations were collected by presenting a sentence from the corpus containing a verb-object pair and sometimes accompanied by the video clip described by the sentence. The difference in responses to sentences with and without the clip will elucidate the effects of visual context on CoS as in the pilot study. The annotations of verbs at different levels of detail allow for analyses about the relations between the levels in terms of CoS.

Turkers were tasked with filling out the CoS frame for up to three changes of state by selecting the frame slot options from a drop down menu. They were also instructed to check 'Current change of state frame is not applicable' (CoS-NA) if none of the options satisfactorily described the change of state denoted by the verb. Furthermore, they could check 'No change of state' (No-CoS) if the verb did not denote a change of state. In latter two cases turkers were prompted to provide a reason for their response.

Annotations for five of the tuples of descriptions (out of 20) were collected with a +/- scene condition for each of the ten bread and cucumber activity videos. We collected three turker responses for each verb to ensure inter-annotator agreement. Thus, 10 videos $\times$ 21 sentences $\times$ 5 tuples $\times$ 2 video conditions $\times$ 3 replicates $\rightarrow$ we can expect 6300 responses. However, in reality we collected 5100 responses for the cucumber and bread videos. This is because not all the short and detailed descriptions contained five or fifteen sentences, respectively. And, not all sentences contained only a single verb, or a verb that takes a

patient (the object that undergoes a change of state).

From the three turker responses collected for each verb-object-scene condition we marked unanimous labels whenever at least two of the attribute, CoS-NA, or No-CoS labels match (e.g., a unanimous No-Cos label is marked if at least two of three of the responses' No-Cos labels match). Because we are only examining the attributes of the CoS frames, we do not consider the object and value portions of the frame for now.

6.1.2 24 Activities Multilevel Dataset

Lastly, a fourth dataset was collected in order to validate the CoS ontology by showing that the CoS frame and frame slot options apply to a wide range of kitchen activities. We chose one video of each of the remaining 24 activities from the TACoS multilevel corpus and turkers were tasked with annotating descriptions of the accompanying sentences. These activities include dicing an onion, frying eggs, etc. This broad scope of kitchen activities ensures that a variety of verbs will appear in the descriptions. Annotations for five (out of 20) of the tuples for each video were collected as in Section 6.1.1. Collecting annotations for this diverse set of activities should tell us how well the CoS ontology covers the kitchen domain.

With 24 videos $\times$ 21 sentences $\times$ 5 tuples $\times$ 3 replicates $\rightarrow$ we expected to collect 7560 turker responses. However, in total we collected 8256 turker responses for the same reasons as stated in section 6.1.1. Unanimous labels were also computed as in the section above.

6.2 Results of Human Studies

6.2.1 Coverage of CoS Ontology

Table 6.1 and Table 6.2 show how well the CoS frame options cover the verbs in the the bread and cucumber and the 24 activities datasets, respectively. These statistics were computed by counting the number of turker responses that contained 'Current change of state frame is not applicable' and 'No change of state' labels. Unanimous (UN) labels exist whenever two or more turkers agree on a label for a given verb. 0.5 percent of the bread and

	All	Single	Short	Detailed	+Scene	-Scene
Num. verb types	94	16	43	79	94	94
Num. verb tokens	1676	140	456	1080	838	838
Num. CoS-NA (UN)	8	0	5	3	5	3
Perc. CoS-NA (UN)	0.0048	0.0000	0.0110	0.0028	0.0060	0.0036
Num. No-CoS (UN)	4	0	2	2	1	3
Perc. No-Cos (UN)	0.0024	0.0000	0.0044	0.0019	0.0012	0.0036
Num. turker responses	5100	433	1384	3283	2548	2552
Num. CoS-NA	57	3	22	32	22	35
Perc. CoS-NA	0.0112	0.0069	0.0159	0.0097	0.0086	0.0137
Num. No-CoS	49	4	11	34	20	29
Perc. No-CoS	0.0096	0.0092	0.0079	0.0104	0.0078	0.0114

Table 6.1: Coverage of the CoS frame options (bread and cucumber multilevel dataset)

	All	Single	Short	Detailed
Num. verb types	222	53	112	201
Num. verb tokens	2715	200	658	1857
Num. CoS-NA (UN)	12	0	4	8
Perc. CoS-NA (UN)	0.0044	0.0000	0.0061	0.0043
Num. No-CoS (UN)	15	0	2	13
Perc. No-Cos (UN)	0.0055	0.0000	0.0030	0.0070
Num. turker responses	8256	611	2005	5640
Num. CoS-NA	115	9	27	79
Perc. CoS-NA	0.0139	0.0147	0.0135	0.0140
Num. No-CoS	133	4	29	100
Perc. No-CoS	0.0161	0.0065	0.0145	0.0177

Table 6.2: Coverage of the CoS frame options (24 activities multilevel dataset)

cucumber verb instances and 0.4 percent of the instances from the other 24 activities have unanimous CoS-NA labels. The low number of CoS-NA labels suggests that the coverage of the CoS ontology is quite thorough.

Based on the feedback provided by the turkers, there are several reasons that they selected CoS-NA. The most frequent reason is a mistake by the automatic part of speech tagger (e.g., labeling a noun modifier as a verb (e.g. *cutting board*) or the labeled verb was actually a deverbal adjective describing a non-changing state of the verb (e.g. *the crumbs that remained*, *sealed package*)). Also, sometimes there were mistakes made by the automatic semantic role labeler (e.g., mislabeling agents as patients). In some cases the verb presented

to the turker was actually not a concrete action verb (e.g. *start*, *finish*, *continue*, need). Lastly, the CoS frame does not apply to sentences from the original TACoS corpus that did not make sense (e.g. *The person leaked the folk.*). None of the above reasons are result from the design of the CoS ontology but other factors.

Some interesting verbs that the CoS ontology does not cover include 'taste' and 'test' (e.g., "The person tested the temperature with his finger"), where the CoS occurs in the knowledge of the agent, and 'took' along with other verbs where the CoS is a change of possession. Change of state of knowledge and change of possession however are abstract concepts, not observable in terms of concrete visual features, so they were intentionally left out of the ontology. There were also vague verbs for which the CoS was not clear from only the linguistic context, for example 'prepare' and 'make'. Also, the verb 'use', which actually seems to indicate that its patient is an instrument in another action. Furthermore, in cases where the verb was used to state the purpose of the action (e.g., "The person took out an orange to make orange juice") or an attempt that may or may not be successful (e.g., "The person tried to remove the lid") the verb does not clearly describe a CoS that actually occurs. These instances were labeled with No-CoS. Lastly, the only concrete change of state that should have been included in the CoS ontology was *turn on/off*. For example, this change of state is visible by the flow of water from a sink faucet.

6.2.2 Effects of Visual Context

Table 6.3 compares the label sets between the +/- scene conditions of the top 20 most frequent verbs from the multilevel bread and cucumber dataset. The metrics show the variability between corresponding verb-object instances' label sets in the two scene conditions. The label cardinality indicates the average number of labels each verb was annotated with.

Overall, the data shows that the label sets do not vary greatly between the +scene and -scene conditions for all verbs (variability of 0.005). I.e. the visual context does not play a large effect on humans interpretations of these verbs. Moreover, the average number of labels does is the same for both scene conditions (0.86 labels). However, the data shows that

Verb	Count	Variability	+S label card.	-S label card.
all verbs	838	0.005	0.86	0.86
slice	96	0.015	0.80	0.73
cut	81	0.018	0.74	0.80
put	67	0.018	0.96	1.00
get	61	0.017	0.93	0.97
place	58	0.006	1.02	1.02
take	57	0.016	0.96	0.93
take out	42	0.010	0.95	0.88
wash	32	0	1.00	1.00
open	31	0.086	0.48	0.55
remove	26	0.047	0.88	0.88

Table 6.3: Comparison between +/-scene conditions (bread and cucumber multilevel dataset)

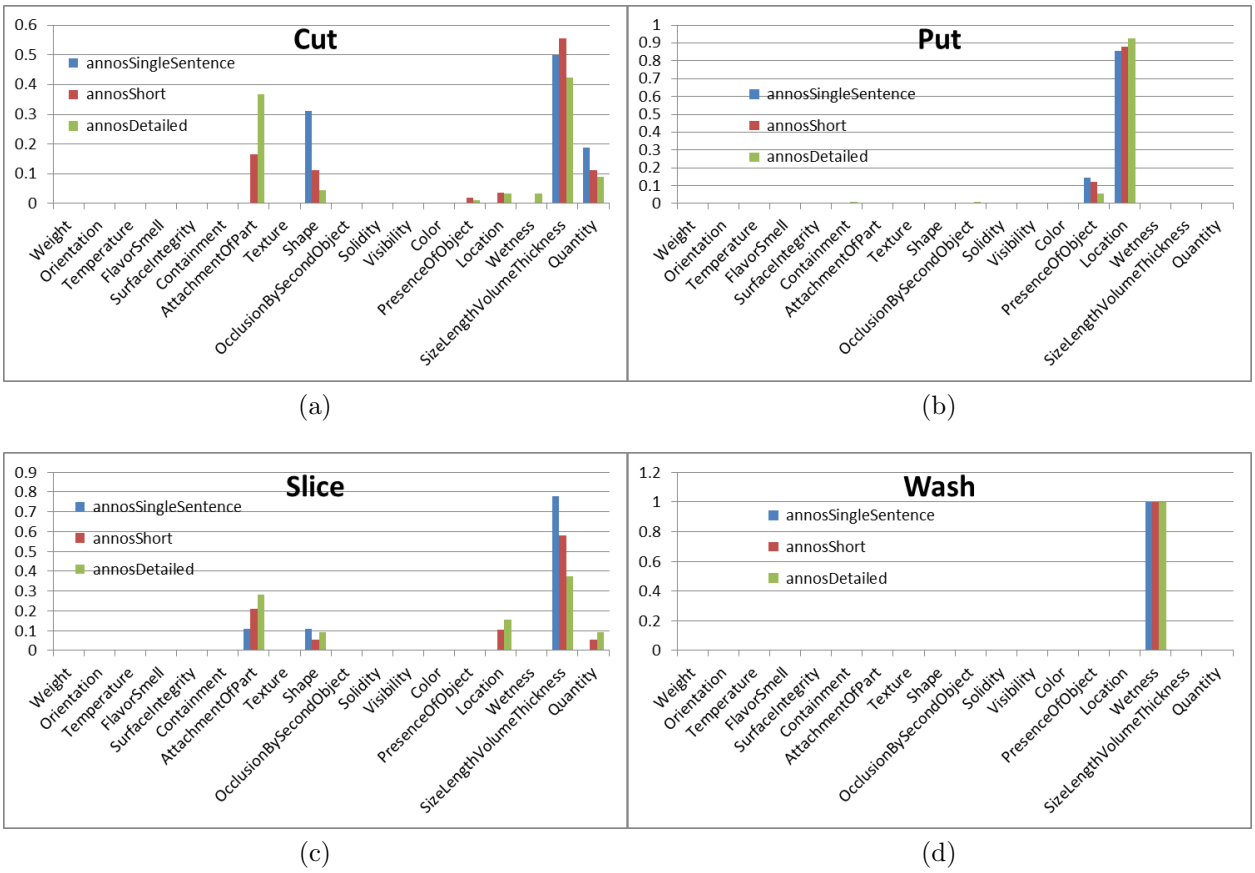
some verbs CoS interpretation depends on the visual context such as *open* with variability 0.086 and label cardinality 0.48 in +scene, and 0.55 in -scene. But the visual context does not affect the interpretation of other verbs as much, such as *wash* with variability 0 and label cardinality 1.00 in both scene conditions.

6.2.3 Level of Detail's Effects on CoS

Does the level of detail in which a verb appears affect the change of state that it denotes? Figure 6.1 shows how the labels are distributed over the attributes for four verbs that occur frequently in all three levels of the 24 activity dataset. We can see that for some of the verbs (*put* and *wash*) the distributions are the same for all three levels. However, for the verbs *cut* and *slice* the distributions differ depending on the level of detail in which the verb occurs. This suggests that there is some ambiguity in determining which changes of state the verbs denotes. Furthermore, the change of state may be determined based on the context of use.

6.2.4 Level of Detail's Effects on Verb Frequency and Distribution

How does the level of detail affect which words appear most frequently? Figure 6.2 shows the frequencies of the top twenty most common verbs in each of the levels from the 24 activity dataset. As observed previously in the TACoS multilevel corpus there are differences between the verbs used at different levels of detail [18, 21]. For example, the single



(a) (b)

(c) (d)

Figure 6.1: CoS distributions over attributes for verbs at three levels of detail (24 activities multilevel dataset)

sentence descriptions contain vague words like *cook*, *prepare*, and *make*, as well as verbs such as *demonstrate* which describes the activity at a higher level of abstraction. These verbs appear less frequently or not at all at more detailed levels of description.

How are the tokens of each verb distributed across the three levels? Table 6.4 shows the distributions of each verb in the 24 activity dataset over three levels of description. Additionally, the entropy of each distribution shows how likely the verb is to appear equally in all three levels (high entropy, max. of 1.0986...) or only to appear in one level (low entropy, min. 0). The entropy of a verb's distribution among the three levels is given by Equation 6.1, where $\delta(v_l)$ is the distribution of verb v in level l.

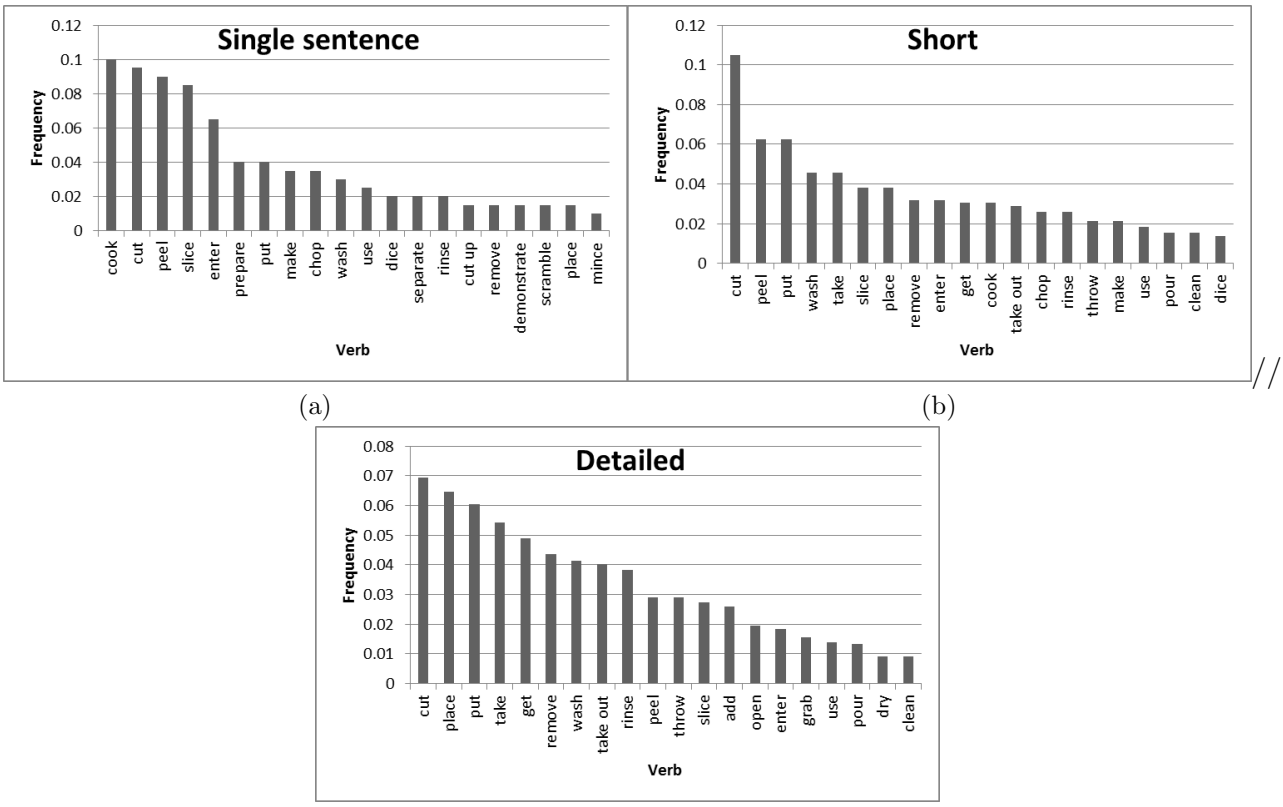


Figure 6.2: Frequencies of occurrence of the top twenty verbs (24 activities multilevel dataset)

$$H(V) = -\sum_{l} \delta(v_l) log_2(\delta(v_l)) \tag{6.1}$$

The data shows that indeed some abstract verbs like *prepare*, *cook*, and *make* are more highly distributed at levels of less detail (with 95%, 72%, and 56% chance of appearing in a single sentence description, respectively). These verbs describe the main activity in the video but there is no reason to use them at greater levels of detail where the sequence of actions is described. Consequentially, these verbs have relatively low entropy (0.22, 0.74, and 0.93, respectively). On the other hand, the verbs *cut*, *wash*, and *rinse* are equally likely to appear in any level (with high entropy 1.08, 1.08, and 1.06, respectively). When used in low levels of detail they describe the most salient part of the video – i.e., the goal of the

Verb	Entropy	Single distr.	Short distr.	Det. distr.	Single count	Short count	Det. count	All count
cut	1.08	0.35	0.39	0.26	19	69	129	217
wash	1.08	0.26	0.39	0.35	6	30	77	113
put	1.08	0.25	0.38	0.37	8	41	112	161
use	1.07	0.44	0.32	0.24	5	12	26	43
rinse	1.06	0.24	0.31	0.45	4	17	71	92
remove	1.02	0.17	0.35	0.48	3	21	81	105
throw	1.02	0.17	0.35	0.48	2	14	54	70
peel	1.01	0.50	0.34	0.16	18	41	54	113
dice	0.98	0.51	0.35	0.14	4	9	10	23
slice	0.98	0.56	0.25	0.18	17	25	51	93
chop	0.97	0.50	0.37	0.12	7	17	16	40
enter	0.97	0.56	0.28	0.16	13	21	34	68
place	0.96	0.13	0.32	0.55	3	25	120	148
make	0.93	0.56	0.34	0.10	7	14	12	33
cook	0.74	0.72	0.22	0.06	20	20	16	56
take	0.69	0.00	0.46	0.54	0	30	101	131
separate	0.68	0.78	0.12	0.10	4	2	5	11
take out	0.68	0.00	0.42	0.58	0	19	75	94
get	0.67	0.00	0.38	0.62	0	20	91	111
add	0.63	0.00	0.32	0.68	0	8	48	56
prepare	0.22	0.95	0.04	0.01	8	1	1	10

Table 6.4: The entropy of the distributions of each verb over three levels (24 activities multilevel dataset)

action (e.g., "The person cut the onion"). Conversely, they can also be used in high levels of detail where they describe a specific action in a sequence of finer actions (e.g., "The person set the onion on the cutting board. He cut the onion into small pieces. He put the pieces in a bowl...").

Chapter 7

Discussion and Conclusion

In conclusion, we have designed an ontology of changes of state as denoted by verbs based on representations of verbal semantics. Furthermore, three datasets containing change of state frame annotations of verbs in the kitchen domain were collected by building on top of the preexisting TACoS and TACoS multilevel corpora. Several analyses showed how visual context, the object of the verb, and level of description affect the way in which a person understand a verb to indicate change of state. It was also demonstrated that the CoS ontology can be used to annotate the changes of state denoted by a wide variety of verbs in the kitchen domain. And lastly, we showed that the CoS indicated by a verb can be predicted to some degree automatically based on linguistic and visual features.

In the future it would be interesting to create a set of classifiers to predict all the slots of the CoS frame (attribute, object, and value) rather than only the attribute. Furthermore, more in depth study is needed to show how specifically the features of the object, the surrounding linguistic context of the verb, and the visual context of the scene described by the verb determine which CoS a verb denotes. This work may draw more inspiration from Generative Lexicon which specifies how the properties of noun arguments may affect the meanings of verbs [12].

The current work may be relevant to future work on action learning by instruction (in combination with demonstration) for robots as it provides a set of changes of state synonymous with the goal of the action. Furthermore, we demonstrated with the prediction

results that even if the verb is unknown (as is the case when hearing a novel verb), that the number of CoS hypotheses can be narrowed down based on the surrounding linguistic and visual context.

Also, after the robot has learned a new action, the human may provide further feedback by describing the CoS of the verb (i.e., goal of the action) in more detail. For example imagine that we are teaching the robot a novel word that means *slice thinly*. If the instructor sees that the robots slices are too thick, an extra level of detail can be added to the robot's CoS representation for the verb, which states that the resulting slices should be greater than a certain thickness.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] John Beavers and Andrew Koontz-Garboden. Manner and Result in the Roots of Verbal Meaning. *Linguistic Inquiry*, 43(3):331–369, 2012.

[2] Yu-Wei Chao, Zhan Wang, Rada Mihalcea, and Jia Deng. Mining semantic affordances of visual object categories. *Journal of Computer Vision*, 88(2):303–338, 2010.

[3] R.M.W. Dixon and A.Y. Aikhenvald. *Adjective Classes: A Cross-linguistic Typology*. Explorations in Language and Space C. OUP Oxford, 2006.

[4] Jane Gillette, Henry Gleitman, Lila Gleitman, and Anne Lederer. Human simulations of vocabulary learning. *Cognition*, 73(2):135–176, 1999.

[5] Shantanu Godbole and Sunita Sarawagi. Discriminative methods for multi-labeled classification. In *Advances in Knowledge Discovery and Data*, volume LNCS3056, pages 22–30. Springer, 2004.

[6] Christopher Kennedy and Louise McNally. Scale structure and the semantic typology of gradable predicates. *Language*, 81(2):345–381, 2005.

[7] Beth Levin. *English verb classes and alternations: A preliminary investigation*. University of Chicago press, 1993.

[8] Beth Levin and Malka Rappaport Hovav. Lexicalized scales and verbs of scalar change. Presented at 46th Annual Meeting of the Chicago Linguistics Society, 2010.

[9] J M Mandler. How to build a baby: II. Conceptual primitives. *Psychological review*, 99(4):587–604, 1992.

[10] George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, November 1995.

[11] Sanmit Narvekar and Kristen Grauman. Relative Attributes. In *IEEE International Conference on Computer Vision*, pages 503–510, 2011.

[12] James Pustejovsky. The generative lexicon. *Computational linguistics*, 17(4):409–441, 1991.

[13] Malka Rappaport Hovav and Beth Levin. Reflections on manner/result complementarity. *Lecture notes*, 2008.

[14] Malka Rappaport Hovav and Beth Levin. Reflections on manner / result complementarity. In *Lexical Semantics, Syntax, and Event Structure*, pages 21–38. Oxford University Press, 2010.

[15] Jesse Read. *Scalable Multi-label Classification*. PhD thesis, University of Waikato, 2010.

[16] Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. Classifier chains for multi-label classification. *Machine learning*, 85(3):333–359, 2011.

[17] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics (TACL)*, 1:25–36, 2013.

[18] Anna Rohrbach, Marcus Rohrbach, Wei Qiu, Annemarie Friedrich, Manfred Pinkal, and Bernt Schiele. Coherent multi-sentence video description with variable level of detail. In *GCPR*, 2014.

[19] Marcus Rohrbach, Wei Qiu, Ivan Titov, Stefan Thater, Manfred Pinkal, and Bernt Schiele. Translating video content to natural language descriptions. In *ICCV*, 2013.

[20] Karin Kipper Schuler. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. PhD thesis, University of Pennsylvania, 2005.

[21] Anna Senina, Marcus Rohrbach, Wei Qiu, Annemarie Friedrich, Sinkandar Amin, Mykhaylo Andrilika, Manfred Pinkal, and Bernt Schiele. Coherent multi-sentence video description with variable level of detail. *arXiv:1403.6173*, 2014.

[22] Aashish Sheshadri, Ian Endres, Derek Hoiem, and David Forsyth. Describing Objects by their Attributes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[23] J Siskind. Grounding Lexical Semantics of Verbs in Visual Perception Using Force Dynamics and Even Logic. *Journal of AI Research*, 15:31–90, 2001.

[24] Jeffrey Mark Siskind. Grounding Language in Perception. *Artificial Intelligence Review*, 8:371–391, 1995.

[25] Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification. *International Journal of Data Warehousing and Mining*, 3(3):1–13, 2007.

[26] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. Mining multi-label data. In *Data mining and knowledge discovery handbook*, pages 667–685. Springer, 2010.

[27] Josiah Wang, Katja Markert, and Mark Everingham. Learning Models for Object Recognition from Natural Language Descriptions. *Procedings of the British Machine Vision Conference 2009*, pages 2.1–2.11, 2009.

[28] Ran Xu, Caiming Xiong, Wei Chen, and Jason J Corso. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In *AAAI*, 2015.