2008

University of North Carolina Wilmington

Master of Science in

Computer Science and Information Systems

Proceedings

https://csbapp.uncw.edu/mscsis

DISAMBIGUATING HUMAN SPOKEN DIARY ENTRIES USING CONTEXT INFORMATION

Daniel James Rayburn-Reeves

A Thesis Submitted to the
University of North Carolina Wilmington in Partial Fulfillment
of the Requirements for the Degree of

Master of Science

Department of Computer Science
Department of Information Systems and Operations Management

University of North Carolina Wilmington

2008

Approved by

Advisory Committee

Dr. Thomas Janicki Dr. Gene Tagliarini

Dr. Curry Guinn (Chair)

Accepted by

Dean, Graduate School

TABLE OF CONTENTS

ABSTRACT

The EPA has commissioned studies to gather fine-grained time / activity / location / exposure data from a diverse cross-section of the population. The information is recorded into digital voice diaries and transcribed by a human for classification into a standard representational system, the Consolidated Human Activity Database. Analysis of the diary entries is a long and tedious process for a human encoder. Automating the process and providing useful information can greatly assist a human encoder in correctly classifying the diary entries.

This paper will discuss utilizing Natural Language Processing (NLP) techniques to analyze spoken diary entries and classify the locations and activities into semantic categories. There will be three main foci that form the hypotheses of the study: improving diary classification accuracy using context information, using thresholds to balance precision and recall tradeoffs, and utilizing the CHAD database structure to improve accuracy by generalizing the semantic ontologies.

The word and context based system shows the relevance of using context information to improve CHAD code classification by using the surrounding diary entry context to augment the word analysis of the diary entries. The threshold-based system shows relative difference levels between top scoring CHAD codes can be utilized to balance tradeoffs between precision and recall. The semantic ontology system shows that generalizing semantic ontologies by employing the CHAD database structure can improve classification accuracy by reducing granularity.

ACKNOWLEDGEMENTS

My thanks go to Dr. Gene Tagliarini for requiring me to work for any assertion I make and keeping me honest in the process. My thanks also go out to Dr. Thomas Janicki for being a mentor and introducing me to new ways of seeing the world. I would like to thank Emma Kay Thornton for pretty much everything; I don't know what we would do without her. I would like to thank Dr. Douglas Kline for always giving me something to get fired up about, and laugh about afterward. I would like to thank Dr. Ron Vetter for tirelessly working on behalf of the graduate students and the program and always having an ear for a curious student. I would like to thank Ference Altrichter for leading me down the strict path of logic and, by extension, language and motivating me to continually experience the boundless joy that is the logical /academic process. I thank Allen Randall and Jessie Berrick for offering time and equipment for the timely completion of my thesis. I would also like to thank Eddie Dunn for helping me retrieve my thesis from the depth of a destroyed laptop at the last minute, thank you.

I would especially like to thank Dr. Curry Guinn for his patience, his understanding, his assistance, his guidance, and most of all his enthusiasm for furthering academic knowledge. He is a mentor and a motivator who has given me a profound respect for the academic process and the possibilities that come when good ideas are explored scientifically.

I would like to thank the departments of Computer Science and Information Systems for providing moral and financial support through my studies, and also providing the facilities necessary to complete this work. I would also like to extend my thanks to RTI out

of Research Triangle Park for providing the diary information and funding my research and time.

I would like to thank my family for the support, moral and financial, and the help through out this long process. I give some of my deepest thanks to my girlfriend Cheryl for being my support through the thesis process. She gave me strength when I was week, advice when I was lost, and the perspective to put the work down when I was too engrossed.

I would like to extend my greatest thanks to my entire committee, for the time, patience, and work they have put into my research. Each person contributed greatly to the learning process of this academic study as well as my academic growth.

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

1. Introduction

1.1 Overview

This thesis investigates the effectiveness of using statistical Natural Language Processing (NLP) to perform text abstraction and classification of human spoken diary entries. The Environmental Protection Agency (EPA) desires fine-grained time/activity/location/exposure (TALE) data to analyze human exposure to various chemicals. The EPA commissioned a study to collect a variety of actual TALE data from a cross section of the population. Subjects in the study were instructed to keep a diary of their daily activities and locations. Each subject used a digital voice recorder to input his or her diary information throughout the day. The voice diaries were then transcribed into a text format for further analysis and record keeping. The transcribed diary entries were analyzed by a human researcher to classify the activities and locations using a uniform method for representation, the Consolidated Human Activity Database (CHAD). The CHAD database was created by the EPA to be used as a unified representation for both location and activity information. The process for classifying diary entries into appropriate CHAD activity and location codes is cumbersome, time-consuming, and error prone for human encoders. The following example illustrates how a typical diary entry might be encoded in the CHAD database:

Example Diary Entry: “In the kitchen about to make some eggs”

Location CHAD code: **30121 - Kitchen**

Activity CHAD code: **11100 - Prepare and clean up food**

1.2 Previous Work

This thesis builds upon previous research in automatically classifying diary entries into semantic categories using statistical NLP (Guinn *et al.*, 2006). In this previous study, referred to as the word-only system, diary entries were categorized using a statistical technique requiring the construction of n-grams of the data. N-grams are a common statistical NLP method for performing text abstraction (Jurafsky and Martin, 2000). Creating n-grams consists of breaking diary entries into sequences of fixed length. Those substrings are then analyzed according to their frequency in categories within the training data. Each category that contains any of the n-grams in the diary entry is given a score based on the statistics from the training data. The category with the highest score is chosen to be the most likely category for the diary entry.

The word-only system always chose the highest scoring semantic category as the correct classification for the diary entry in question. The system performed well when tested against the training set as a benchmark (high 90% correct classifications), but when tested against new diary entries it performed far worse (below 70% correct classifications). The word-only system used the n-gram model to analyze the words present in a single diary entry taken out of context. The original data contained more information than just words within the diary entries; it also included the time and dates for particular diary entries and the sequence of diary entries.

1.3 Limitations of the Word-only System

The word-only system's method, though very successful when applied to the training set, was much less successful when applied to the test set. Out of 949 diary entries tested in

the training set, only 17 locations were misclassified (98.21% accuracy) and 23 activities were misclassified (97.57% accuracy). When run against the 272 diary entries in the testing set, 86 locations were misclassified (68.38% accuracy), and 115 activities were misclassified (57.73% accuracy).

1.4 Problem Definition

The goal of the research in this thesis is to develop a system to assist a human encoder in classifying human spoken diary entries. The automated system should classify diary entries with their proper Location and Activity CHAD codes with high accuracy. However, if the system is not able to classify a diary entry with high accuracy, it should alert the human coder of that fact.

1.5 Hypotheses

Hypothesis 1: Performing statistical NLP text abstraction using multi-diary entry contextual information will improve the disambiguation of human speech diary entries over the word-only n-gram model applied to single diary entries.

Hypothesis 2: Thresholds can be found to balance trade-offs between precision and recall.

Hypothesis 3: The hierarchical structure of the CHAD database can be exploited for a more general model of human activity/location with higher classification accuracy.

1.6 Basic Approach to Each Hypothesis

Hypothesis 1:

The original data contained information that was disregarded in the word-only system which encoded the location or activity of a diary entry based solely on the words. For example, the activity of diary entry "Walking down the street going to the bus stop to wait for the bus" was correctly encoded as the CHAD code **18000 - Travel, general**. However, the location coding was incorrectly selected as the CHAD code **30120 -Your residence, indoor**. It is apparent that location and activity encodings for this utterance are incompatible with each other. The approach for hypothesis one will be to combine the calculations from the word-only statistical analysis with calculations derived from a statistical analysis of the diary entry's context.

Hypothesis 2:

The word-only system always made classified a diary entry, regardless of what the relationship of the scores were to one another. For example, the diary entry "going to lay [sic] in bed for 20 to 30 minutes" was classified as the location **30122 - Living room/family room** with a score of 0.6448 for its first choice and **30125 - Bedroom** with a score of 0.6296 for its second choice. These scores are relatively close (less than difference 3%) and indicate an ambiguous situation. The approach for the second hypothesis examines the usefulness of setting thresholds to determine the system's confidence in making a database encoding. Then the resultant tradeoff between precision and recall will be analyzed.

Hypothesis 3:

Another issue has to do with the granularity provided by the CHAD code database. The CHAD database was built inherently hierarchical, where the codes represent more granularity as the digits move to the right. The word-only system did not utilize this granularity in its calculations which could also be utilized to improve accuracy over the previous system. For example, the diary entry "going to make Kool-Aid in the kitchen" was encoded as the activity **11220 – Clean house** for the first encoding, **11100 - Prepare food** for the second encoding, and **11000 - General household activities** for the third encoding. The first two encodings are subcategories of the third encoding. If less granularity is required by a study, the general category **11000 – General household activities** could be used to encode this diary entry and it would be a correct, though less specific, encoding.

2. Review of Literature Review and Analysis

2.1 Capturing Environmental Exposure Data

The definition of environmental exposure provided by Ott (Ott, 1982) and later adapted by others (Lioy, 1990; NAS 1991; USEPA, 1992; Zartarian *et al.*, 1997) is "an event that occurs when there is direct contact at a boundary between a human and the environment with a contaminant of a specific concentration for an interval of time." This definition implies that four variables must be measured to characterize exposure accurately: location (L), time (T), activity (A), and concentration (C).

With renewed EPA interest in understanding the relationship between activity and exposure, the early activity data collected as part of exposure field studies and activity pattern surveys (*e.g.*, (Johnson, 1987)) needed to be unified in a single representation. The EPA developed a system to store and analyze the available data systematically, the Consolidated Human Activity Database (CHAD) (McCurdy *et al.*, 2000).

Relatively recent surveys to capture time activity data have been undertaken both on a state level (Robinson *et al.*, 1989) and the national level (Klepeis *et al.*, 1998). These surveys utilized the 24-hour recall method, which was a snapshot in time of an individual's locations and activities recorded throughout the diary.

Paper-based diaries, electronic diaries, voice-recorded diaries, and observational techniques have all been used to collect data about temporal activities, spatial locations, product use, and dietary consumption of research participants (Johnson *et al.*, 2001). Diary methods relying on recall are not highly reliable and have a relatively high respondent burden, which negatively impacts participant compliance. Observational techniques are costly and burdensome. Post-study processing of diary entries is labor intensive unless

simplified reporting protocols are employed and automatic processing systems are developed.

In an earlier study of carbon monoxide exposures, the investigators recognized the need to automate the data collection and pilot tested a hand-held data entry device to be used by the participant to enter the location/activity code throughout the monitoring period (Akland *et al.*, 1985). This methodology utilized a programmable HP 41C calculator, where the keys were programmed to represent specific locations or activities (Zelon, 1989). This basic system was modified by Freeman (Freeman *et al.*, 1991) although a different micro-processor system was used and an internal clock was included in the package. More sophisticated electronics available today enable the participant to record a broader range of locations and activities by use of hand-held devices, including the Palm Pilot or Personal Digital Assistant (*e.g.*, (Braeur, *et al.*, 1999)). The hardware and protocol used in the data gathered in the RTI experiment involved having subjects record their diary entries using a digital voice recorder (RTI, 2001; Guinn *et al.*, 2006; Guinn and Rayburn-Reeves, 2008).

2.2 Text Abstraction

The primary task of the word-only system was to take spoken language diary entries like "I am on the bus on my way to South Square Mall" and select the appropriate CHAD location code (**31140 - Travel by bus**) and activity code (**18400 - Travel for goods and services**). In the computational linguistic community, this task has been called text abstraction. Text abstraction has a long history in computational linguistics with a variety of techniques employed to tackle the task of deriving relevant information from human spoken text. For the problem domain tackled by the word-only system, two techniques have

previously been implemented: one involving semantic grammars and minimum distance parsing; and one using statistical language processing (Guinn *et al.*, 2006).

2.3 Semantic Grammars

Semantic grammars have long been employed in practical natural language systems (Burton, 1976). Semantic grammars are efficient representations of limited domains. The drawback is that these grammars are typically hand-crafted. For this voice diary domain, a hand-crafted semantic grammar was constructed with approximately 60 man-hours of work. The number of grammar rules was 752 (Guinn *et al.*, 2006).

2.4 Minimum Distance Parsing

The stream of words that the parser has to process in a spoken language system often is "ungrammatical" for numerous reasons: 1) speakers often violate prescriptive grammar rules, 2) speakers frequently use sentence fragments, 3) normal speech is filled with pauses, restarts and other disfluencies. In fact, even the best speech recognition systems rarely perform better than 90% on word-for-word transcription (and it's often much worse). To combat these issues, the semantic grammar system relies on a minimum distance parsing technique that finds the "best" match in the grammar based on the fewest additions/deletions/substitutions needed to make the input fit a grammar rule (Hipp, 1992).

2.5 Statistical Natural Language Processing

With hand-built semantic grammar, engineers must predict what words and grammatical structures are likely for the domain. Minimum distance parser offers some

flexibility, but the further the speakers deviate from the predicted structures, the lower the precision.

An alternative approach is to use statistical NLP techniques that analyze a training corpus to build probabilities that can be used to choose the most likely semantic categories for a diary entry. In this study, a selection of the total corpus was set aside as training data. From this training set, a human coder selected CHAD activity and location codes for each diary entry.

2.6 Naïve Bayes

Using Bayesian statistics, unigram, bigram, and trigram probabilities were generated for each set of words for each activity and location code. (An overview of Bayesian statistics may be found in *Speech and Language Processing* (Jurafsky and Martin, 2000)). The naïve Bayes rule is the statistical technique used in the first study of this domain (Guinn *et al.*, 2006). The goal is to improve upon this baseline by augmenting the technique with contextual information.

There are other statistical techniques and methodologies that may be applicable to this domain including latent semantic indexing (LSI) (Deerwester *et al.*, 1990). This technique is used to match documents which contain large amounts of text (*e.g.*, a web page). In the spoken human diary domain, the textual data tends to be very small (on average 9.5 words). Neural networks are another statistical technique that has been successfully employed in text categorization (Ruiz and Srinivasan, 2004). An area of future research in this area would be to contrast the performance of a neural network with the word-only system's Bayesian approach. However, the word+context system's goal is to improve upon the baseline

approach of the word-only system by adding context information. Thus, it is desirable to use the same underlying statistical techniques to have a more direct comparison.

2.7 Using Context to Resolve Ambiguities

A primary goal of this research project is to use pragmatic, situational context to resolve ambiguities at the diary entry level. Theories that use text coherence (Grosz *et al.*, 1995) or rhetorical structure theory (Mann and Thompson, 1988) have some relevance, but the domains in which these theories have previously been applied are in coherent, tightly connected texts. The diary entries of the subjects do not constitute a coherent dialog.

One area of research that is similar in its goal to the word+context system is the attempt to predict future user behavior from past user behavior (such as: interacting with computer software, making web queries). The word+context research could utilize either of two primary data structures: Markov models (Mayrhofer *et al.*, 2003) or Bayesian networks (Oliver and Horvitz, 2005). The latter work uses statistical techniques similar to the techniques utilized in this study. A Markov model approach might be applicable to this study's domain. An area of future work would be to do an explicit comparison of Markov models contrasted with the naïve Bayes approach.

3. Methodology

3.1 The Word-only System

3.1.1 Data Organization and Definition of Terms

There are a variety of terms used in this section to describe various elements of the word-only system:

Diary Entry: a test subject's recorded spoken diary entry for a particular activity and location.

Each diary entry is transcribed by a human into a textual format. Then each diary entry is hand-coded by a human with an appropriate CHAD code. The human coder is instructed to use all available information (*e.g.*, surrounding utterances, time of day, day of the week) when encoding each diary entry. This information included: the words in the entry, the previous and following utterances in the diary, and the context of the utterance itself (for example using the subject's activity to discern their location). For example, examine this diary entry along with its corresponding CHAD code:

Diary Entry: "in bathroom going to do my daily routine"

Location CHAD code: **30124 - Bathroom**

Activity CHAD code: **14600 - Other Personal Needs**

Database: the collection of all diary entries for all subjects over all days of recording.

The diary database consists of a text file with all diary entries with the corresponding Activity and Location CHAD codes.

Shown in Table 1 is a sample of the database entries for the word-only system:

Table 1: Sample of the data base from the word-only study

Utterance	Location CHAD	Activity CHAD
brushing my teeth	30124 - Bathroom	14600 - Dress, Groom
I'm disconnecting the electrodes in order to take a shower	30124 - Bathroom	14700 - Other personal needs
alright heading back from the kitchen to the bedroom so I can go back and lay down	30125 - Bedroom	14500 - Sleep or nap
from the kitchen to the bedroom	30125 - Bedroom	11000 - General household activities
I'm in the car driving now	31110 – Travel by car	18200 -
alright I'm back inside the car now	31110 – Travel by car	18200 -
I've finished my back exercises and am now going back downstairs	30100 – Your residence, Indoor	11000 - General household activities
going downstairs	30100 – Your residence, Indoor	11000 - General household activities
I've finished my back exercises and I'm now going downstairs	30100 – Your residence, Indoor	17130 - Exercise
going out to the garage to start up the truck	35200 – Public garage/parking lot	18200 – Travel to/from work
completed my exercise and I'm now walking down a hallway and the eh stairs so I can get ready to take a shower	30100 – Your residence, Indoor	14600 - Dress, Groom
went down the hallway to my office	30126 – Study / Office	10000 - Work and other income producing activities, general

Testing Set: Testing set is comprised of a sampling from the entire data set.

Training Set: The training set is comprised of all of the diary entries that are not in the testing set.

3.1.2 Word-only Classification

The primary task of the word-only system (Guinn *et al.*, 2006) is to examine spoken language utterances such as, "I am on the bus on my way to South Square Mall", and automatically select the appropriate Activity and Location CHAD codes that most correspond. The task of this research, mentioned in section 2.2, is commonly referred to as "text abstraction" and is performed using statistical NLP techniques.

The statistical NLP technique consists of segmenting the original data into a training corpus and a testing corpus. The training corpus is utilized to perform statistical analysis and build probabilities that can be used to choose the most likely semantic categories for new utterances. The testing corpus is utilized to provide new utterances that the system is not trained on to test it's classification ability.

The training set is then analyzed to get probabilities for a naïve Bayes classifier to classify the utterances of the testing set. For any particular diary entry in the training set, statistics were generated based on the unigram (single word), bigram (word pair), and trigram (word triple) probabilities and combined together to classify an utterance into a CHAD activity and location code. In this paper, when referring in general to unigrams, bigrams and trigrams, the term "n-grams" will be used in their place.

3.1.3 Bayes Classifier for N-grams

N-gram probabilities are determined for each utterance in the training set. Examine the unigram "kitchen", to determine the probability that the occurrence of the word "kitchen" in a diary entry corresponds to the semantic category **30121 - Kitchen**, the Bayes' rule is employed:

$$P(A \mid B) = P(B \mid A) * P(A) / P(B)$$

or

$$P(Kitchen \mid "kitchen") = \frac{P("kitchen" \mid Kitchen) * P(Kitchen)}{P("kitchen")}$$

The formula for P("kitchen"| **30121 - Kitchen**) was computed by the percentage of times the word "kitchen" appears in utterances that have been classified as the category **30121 - Kitchen**. P(**30121 - Kitchen**) is the probability that an utterance is of the semantic category **30121 - Kitchen**, and P("kitchen") is the probability that "kitchen" appears in any utterance.

3.1.4 N-grams and Basic Scores

Given a string of words I = {$word_1$, …, $word_n$}, the unigram formula for the semantic category Kitchen is:

$$Unigram(Kitchen) = 1 - \prod_{i=1}^{n} (1 - P(Kitchen \mid word_i))$$

This formula assumes that P(**30121 - Kitchen** |$word_i$) is independent of P(**30121 - Kitchen** |$word_j$). These values are not strictly independent; however the formula is a good approximation when employed with a large enough corpus (Guinn *et al*., 2006). More

generally for any semantic category, S, and any input string I = {$word_1$,…, $word_n$}, the unigram estimate is:

$$Unigram(S,I) = 1 - \prod_{i=1}^{n} \left(1 - P\left(S \middle| word_i\right)\right)$$

(Guinn et al., 2006). Similar probabilities are computed for bigrams and trigrams:

$$Bigram(S,I) = 1 - \prod_{i=1}^{n-1} \left(1 - P\left(S \middle| word_i, word_{i+1}\right)\right)$$

$$Trigram(S,I) = 1 - \prod_{i=1}^{n-2} \left(1 - P\left(S \middle| word_i, word_{i+1}, word_{i+2}\right)\right)$$

3.1.5 N-gram Score Combination and Diary Entry Classification

Trigrams should be given higher weight than bigrams and bigrams higher than unigrams due to the relevance of three word sequences to particular CHAD encodings. In the case of unigrams, the likelihood that a unigram is in multiple CHAD categories can be high (*e.g.*, the words: to, for, a). Bigrams have a lower likelihood that there are multiple possible categories, and for trigrams the likelihood is even less . For example, examine the trigram "riding to work". An activity coding from the general **Travel to/from…** categories deserve much higher precedence based on the trigram than the presence of the unigrams "riding", "to", and "work" somewhere else in the input string. Since trigrams are more relevant in encoding diary entries than bigrams, and bigrams are more relevant than unigrams, weights will be applied according to relevance.

In the word-only system (Guinn *et al.*, 2006), a simple linear combination is employed where, for each input and possible semantic category, the formula is:

$$Score(S,I) = 1 * Unigram(S,I) + 1.5 * Bigram(S,I) + 2.5 * Trigram(S,I)$$

To determine which semantic category is most likely for each input I (I=utterance), the formula Score(S,I) is calculated for all possible values of S (S=CHAD code for either location or activity), and the maximum score is chosen as the correct CHAD code classification. Essentially, the CHAD code with the highest overall score is the system's choice as the correct semantic category for the diary entry.

3.2 Re-testing the Word-only System

3.2.1 Data Re-organization

There are two main concerns with the word-only system and the nature of its training and testing sets: the inconsistencies in CHAD encoding, and the requirement of chronological data for the context contribution.

The first issue with the data from the word-only system is inconsistencies in CHAD encodings. This study analyzes the words in the diary entries and relies on a naïve Bayes classifier to classify for each diary entry. If the classifications for similar diary entries are inconsistent, there will be inconsistencies in the classification of similar diary entries. The data set from the word-only system contains many diary entries that are similar in their semantic category, yet are classified into different CHAD codes. When the human encoder classified the diary entries, he/she classified similar diary entries into different CHAD codes. For example, consider the diary entries:

> Diary Entry: "on my way walking to the store ran through the shortcut in the woods"
> Human encoded Activity CHAD code: **18000 - Travel, general**
> Human encoded Location CHAD code: **31210 - Travel by walk**
>
> Diary Entry: "walking to the store to get stuff for gathering tonight"
> Human encoded Activity CHAD code: **13210 - Shop for food**
> Human encoded Location CHAD code: **31210 - Travel by walk**

One problem is that the activities of the two diary entries are similar, yet the CHAD encodings differ greatly. In both diary entries, the subject is walking to the store to go shopping. The CHAD Location code is encoded the same for both diary entries, which is correct, but the CHAD Activity codes are encoded differently, although the activities the subject is engaging in are the same. It makes little sense to encode the activities as being different given the words of the diary entry.

The other problem is the CHAD Activity codes in which both entries are classified into do not accurately describe the activity referenced in the diary entries. The activity CHAD codes for the diary entries in the example were, **18000 - Travel, general** and **13210 - Shop for food.** A more suitable and more often used CHAD code, **18400 - Travel for goods and services**, would accurately describe the activity with sufficient detail and be more consistent.

The inconsistencies in the human encoded diary entries are due in part to the ambiguity that is inherent in the CHAD database. Within the CHAD database there are multiple codes that could be considered in the same semantic categories. For example, consider the activity and location codes below:

Diary Entry: “at the bank”

Valid Activity CHAD encoding: **13200 - Shop / Run Errands**
Valid Activity CHAD encoding: **13230 - Run Errands**
Valid Activity CHAD encoding: **13500 - Obtain government / financial services**
Valid Activity CHAD encoding: **13800 - Other services**
Valid Activity CHAD encoding: **18400 - Travel for Goods and Services**

Valid Location CHAD encoding: **31300 - Waiting**
Valid Location CHAD encoding: **32100 - Office building / bank / post office**
Valid Location CHAD encoding: **35900 - Pool, river, lake**

The diary entry “at the bank” can be encoded as multiple encodings within the CHAD database. A human encoder could reasonably encode the diary entry into one of the five activity categories listed and any one of the three location categories listed above.

Another encoding issue was the use of “u” to encode diary entries with unintelligible encodings and “x” to encode diary entries missing either location or activity information. Though descriptive to the human encoder as to the nature of the voice data, they provide no information for the word-only system or the word+context system to utilize in classification. In fact, “u” and “x” become possible valid encodings when testing a diary entry as they are considered CHAD code classifications in the training set statistics. It was decided that listings of this type must be removed for the sake of consistency in the database of diary entries.

Though it may be difficult for a human to distinguish which CHAD code to encode a particular diary entry, the code that is chosen for the semantic category should be consistently applied throughout the entire data set. This includes CHAD codes that make little sense and need to be removed entirely. Thus, the full data set has to be re-analyzed to ensure that similar diary entries are encoded with the same CHAD code and that superfluous CHAD encodings are removed.

The second issue is that the training and testing sets are selected to portray a sampling of the entire data set. Testing and training sets chosen to represent the diversity of words in the entire data set properly show the word-only system’s ability to classify diary entries accurately. Though this format is convenient for the purposes of the word-only system, it is not indicative of actual data from a study similar to the EPA study (i.e., a series of chronological diary entries). The addition of context information requires that the

chronological structure of the original diary entries be preserved, thus it is a necessity to re-order the word-only system's data set to reflect the original format more closely.

These two issues required resolution before the word+context system could be combined with the word-only system. The database used in the word-only system had to be reconfigured. First the data was grouped by subject. Then, within each subject, the diary entries were put in chronological order. Then divisions were noted where each day of data started and ended. Thus the resulting database consisted of eight subjects' data and a total of 42 days of data.

The reordering process changed the data so drastically that the word-only system needed to be re-tested. This was done both to see what changes occurred with a better organized data set, and so the word+context results could be compared to the word-only results.

3.2.2 Terms for the New Data Organization

Following are definitions that will be used for the context and word system:

Day of recordings: a collection of diary entries for a particular subject for a particular day.

Each subject is instructed to record his/her activities and locations as they changed through the day (noted in Section 1.1). These collections of entries are preserved by date of recording to keep the subject's original sequence of utterances. So a day of recordings consists of chronological diary entries for one subject.

Database: the collection of all diary entries for all subjects over all days of recording.

The diary database consists of a text file with all diary entries with the corresponding Activity and Location CHAD codes. The diary entries are organized by subject and in chronological order. Table 2 shows a sample of the database entries:

Table 2: Sample of the database used in the re-organization for the word+context system

Time	Recorded Utterance	CHAD Location	CHAD Activity
8:57 AM	in the bedroom starting housework	30125 - Bedroom	11200 - Indoor chores
8:59 AM	carrying clothes to the laundry room	30128 - Utility room / Laundry room	11410 - Wash clothes
9:00 AM	the bedroom getting more clothes	30125 - Bedroom	11410 - Wash clothes
9:05 AM	loading the washing machine in the laundry room	30128 - Utility room / Laundry room	11410 - Wash clothes
9:06 AM	sitting down going to watch twenty minutes of Regis	30122 - Living room / family room	17223 - Watch TV
9:23 AM	I'm going to be brushing the dog in the family room	30122 - Living room / family room	11800 - Care for pets/animals
9:29 AM	the laundry room moving the clothes from the washer to the dryer	30128 - Utility room / Laundry room	11410 - Wash clothes
9:34 AM	taking the dog for another walk in the rain	30210 - Your residence, outdoor	11800 - Care for pets/animals
9:45 AM	kitchen doing dishes	30121 - Kitchen	11210 - Clean-up food
10:00 AM	in my office checking email	30126 - Study / Office	17160 - Use of computers
10:10 AM	in the hallway playing ball with Fetzer	30120 - Your residence, indoor	11800 - Care for pets/animals
10:18 AM	back to the laundry room to load and unload clothes	30128 - Utility room / Laundry room	11410- Wash clothes
10:54 AM	pulled the sheets from the dryer I'm going to making the bed and I'm going to be switching stuff from the washer to the dryer and reloading the washing machine reloading washing machine	30128 - Utility room / Laundry room	11410- Wash clothes
11:13 AM	at the computer	30126 - Study / Office	17160 - Use of computers
11:34 AM	on the phone dealing with clients	30120 - Study / Office	10000 - Work and other income producing activities

Testing Set: The testing set consists of a collection of chronological diary entries from each subject in the study, a total of 272 diary entries.

The total diary size is 1220 entries from 42 days of subject data. The two criteria for creating the testing set are: 1) the entries must be in chronological order by subject, and 2) the testing set should contain data from each of the eight subjects in the study.

The need for chronology has been expressed in section 3.2.1 as a requirement of the context system. The addition of context information requires information from the surrounding diary entries. The data is organized by subject for easier analysis and to assist in generating the testing and training sets.

The testing set needs to represent the variety of data present in the full diary, including data from all of the subjects. Each subject used different vocabulary, described activities and locations in different levels of detail, and described different locations and activities.

In light of the two constraints and the limited size of the total diary (only 42 days) the testing set is created by removing one day of data from each subject. The resulting test set is made up of eight days of subject data, one day from each subject. This is done to preserve the sequence of diary entries while sampling from all subjects in the study.

Training Set: The training set is comprised of the remaining subject data that is not put into a testing set, a total of 948 diary entries.

3.2.3 Word-only Re-testing

The reorganized data had to be re-tested to see how the results differed from the word-only system. The training and testing process is similar to the word-only study as detailed in Section 3.1, with only a few differences. The first difference is the composition of the testing and training set, an example of which is shown in Table 1. The other difference is the weights applied to the word-only system, see the following formula.

$$Score(S,I) = 0.2 * Unigram(S,I) + 0.3 * Bigram(S,I) + 0.5 * Trigram(S,I)$$

The weights are chosen so they added up to one; thus, the final score represented a combination of elements that are parts of a whole (Jelinek, 1990).

3.3 Hypothesis 1: Using Context

Hypothesis 1:

"Performing statistical NLP text abstraction using multi-diary entry contextual information will improve the disambiguation of human speech diary entries over the word-only n-gram model applied to single diary entries"

3.3.1 Using Context to Resolve Ambiguities

This study's primary goal is to improve the baseline accuracy of the word-only system by utilizing contextual information present within the data to augment the classification of diary entries.

The data collected from each subject in the voice recorders is initially in chronological order. When the diary entries are transcribed and encoded, the chronological order is preserved. Thus the sequence of Activity and Location CHAD codes is preserved.

This sequence is exploited to generate probabilities for each diary entry's CHAD code classification relationships (past and current classifications for location and activity) that are used to augment the word-only system's classification scores.

3.3.2 Primary Reasoning

There are a few general relationships concerning a subject's locations and activities that serve as the basis for using context. These relationships are used to derive context information from the training set to assist in the classification of diary entries. In the word-only system, the relationship between a word and the words around it is utilized for classification. In the case of context, a relationship exists between activities and locations, essentially the connection between where a subject is/was and what the subject is/was doing. This relationship is utilized by the word+context system to augment the word-only system score.

One of the context relationships is that current activities and locations are related. For example, if a subject is cooking food that subject is typically in the kitchen. It is also highly unlikely that the subject is in the bathroom or on the sidewalk outside when their activity is cooking. Another of these relationships is the connection between a subject's current location and the subject's previous location. Some typical examples of these relationships are: where a subject currently is and where that subject previously was are related, what a subject is doing and what a subject was just doing are related as well. In all there are six relationships that formed the base of the context calculations:

- 3 for activity
    - Current Activity given Current Location (CAct given CLoc)
    - Current Activity given Previous Activity (CAct given PAct)
    - Current Activity given Previous Location (CAct given PLoc)

- 3 for location
    - Current Location given Current Activity (CLoc given CAct)
    - Current Location given Previous Location (CLoc given PLoc)
    - Current Location given Previous Activity (CLoc given PAct)

An example of a difficult location for the word-only system to encode is a home office. An entry such as "sitting at my desk" out of context has ambiguities at whether the location is a home office or an office at a place of business (but is distinguished in the CHAD database). Since offices at places of business are more frequent in the corpus, the word-only system would likely select **32100 - Office building / bank / post office** as opposed to **30126 - Study / Office**. However, if the previous location indicated the subject was at home, this information should be factored into the analysis.

3.3.3 Context Formulas

For each contextual relationship, a formula serves as the basis of the calculations to augment the word-only system. The generic formula for each of the six calculations is a naïve Bayes probability of the current CHAD code given the relationships described in section 3.3.2.

CLoc given CAct = $P(CLoc \mid CAct)$
CLoc given PAct = $P(CLoc \mid PAct)$
CLoc given PLoc = $P(CLoc \mid PLoc)$

A parallel structure exists for Activity CHAD codes.

3.3.4 Training and Creating Context Probabilities

In the word-only system, the training begins by breaking each utterance in the training corpus down into word n-grams. Then probabilities are assessed based on the frequency that a particular n-gram was classified into a specific CHAD code for both locations and activities. In general, the word-only system reads each diary entry, calculates probabilities based on the word n-grams, and stores those probabilities for use in the testing process.

When the word+context system tests new diary entries a similar process is employed. Each diary entry is read in and broken down into n-grams (unigrams, bigrams and trigrams). Then each n-gram is used as a basis to reference the corresponding probabilities found within the training set. The CHAD codes that are found with the n-grams in the training set are kept as possible categories in which to classify a new diary entry. A score is calculated for each n-gram using the data from the training set and weighted accordingly: low weight for unigrams, higher weight for bigrams, and the highest weight given to trigrams. Scores are tabulated for each possible CHAD code and the CHAD code with the highest overall score would be considered the correct utterance encoding for both activities and locations.

In the word+context system, a similar process is followed for context data that is used in the word-only system. The context training begins with recording the CHAD classification of each utterance in the training set preserving the chronological of the original diary entries. Then probabilities are assessed for each of the six context relationships for all utterances in the training set. These six probabilities are preserved in order to test new diary entries from the testing set.

When the word+context system is employed on new diary entries, a process similar to the training process is utilized. The chronological sequence of CHAD codes from the testing set is preserved and used to access probabilities from the training set. The six context relationships from the testing data reference the corresponding training probabilities. These six formulas are calculated for each diary entry in the testing set and combined with the word-only system's score to generate an activity score and location score.

3.3.5 Score Combination

When calculating the word-only system, there is only one probability of importance, the word n-grams. In the case of the word+context system, there are four factors to combine for both location and activity. Assuming a linear combination of weights, optimal weights are determined using a brute force, or (an almost) exhaustive search, method. Optimal weights for both the word+context system as well as the word-only system refer to the set of weights that produces the highest diary entry classification accuracy for a singe data set.

The weights are assigned to each calculation (for Activity the relevant scores are N-gram score, Current Location, Previous Activity, and Previous Location; for Location, N-gram score, Current Activity, Previous Location, and Previous Activity) and then combined to get an overall score for a particular CHAD code classification:

$$ActivitySc\ ore = w_1 * PerWord + w_2 * PAct + w_3 * PLoc + w_4 * CLoc$$

$$Location\ Score = w_5 * PerWord + w_6 * PLoc + w_7 * PAct + w_8 * CAct$$

And the weights conform to the requirement:

$$\sum_{i=1}^{4} w_i = \sum_{i=5}^{8} w_i = 1$$

The weights for location all add up to one, and the weights for activity all add up to one. They are determined by brute force by trying all combinations between 0.0 and .95 with a step size of 0.05.

3.4 Hypothesis 2: Thresholds

"Thresholds can be found to balance trade-offs between precision and recall."

3.4.1 Score Thresholds

The purpose of the word+context system is to assist a human encoder in the process of classifying diary entries of the type in this study. Classifying all of the utterances correctly with a computer would be the ultimate goal of a system such as this, but achieving 100% accuracy on unseen diary entries would be extremely difficult to do.

The word+context system uses only the highest scoring CHAD code as the correct semantic classification. The CHAD code with the highest probability may be the most likely based on the training data, but other CHAD codes may be very close. This method disregards the relationship between the top scores for any particular diary entry. Finding thresholds within the scores returned by the system would allow for it to either encode with confidence or alert the human encoder to the ambiguous encoding. The threshold system analyzes the top scores for a diary entry and flags ambiguous classifications.

3.4.2 Primary Reasoning

The scores returned from the word+context system are calculated from the probability for each CHAD code based on the words in the diary entry and the six context formulas. Thus the score represents the system's calculations on how likely a particular CHAD code is the correct encoding for a particular utterance. It also follows that the relative closeness of the CHAD code scores represents the closeness of their probabilities. If two CHAD code scores are relatively close, then they are both likely encodings for a particular diary entry.

In the case where there are multiple likely encodings for a diary entry, the information from the training set suggests that the highest scores are similarly probable as the correct encoding. This case is an ambiguous situation, where there is no clear "most probable" CHAD encoding for the diary entry. Ambiguous situations can then be flagged for later analysis by a human encoder and the system can move on to the next utterance.

For example, the diary entry "going to lay in bed for 20 to 30 minutes" was classified for location as **30122 - Living room/family room** with a score of 0.6448. The system's second location choice was **30125 - Bedroom** with a score of 0.6296. These scores are relatively close (less than 3% difference between them) and this closeness indicates an example of an ambiguous situation.

3.4.3 Determining Score Thresholds

The threshold system in this study will analyze the relationship between the two top scores and measure the relative difference between them. The formula for the comparison is as follows:

$$Threshold(highest\_score, second\_highest\_score) = \frac{(highest\_score - second\_highest\_score)}{highest\_score}$$

The threshold system uses the relative difference between the top two scoring CHAD codes as the relationship for flagging ambiguous situations. If the difference is low, then the scores are close to one another; if it is large they are further apart. The threshold is a measure of difference between the top CHAD code scores. If the difference is high enough, the diary entry will be flagged for follow up analysis; if it is not, the system will encode the top score as the correct CHAD code for the diary entry.

It is difficult to know by simply looking at the data what the threshold levels for the relative difference should be. Thus, the threshold levels will be determined experimentally and could later be utilized to balance precision and recall data for all utterances.

3.4.4 Guesses for the threshold system

The threshold system is, in part, based on the concept of guessing; whether the system classifies a diary entry or not. In the word-only and the word+context systems, each diary entry is classified with a CHAD code regardless of the relationship of the scores for each. The threshold system relies on a different method altogether, either it classifies a diary entry as particular CHAD location and activity code or is does not. The threshold system guesses based on the relationship between the top two scoring CHAD codes, if the relative difference exceeds the threshold level, it classifies the diary entry. A diary classification, correct or not, is considered a guess by the threshold system.

3.4.5 Measuring Score Thresholds: Precision and Recall

Precision and recall are the basis for measuring threshold levels. Precision refers to the fraction of correct word+context system guesses relative to all guesses: for all selections made, how many are actually made (Jurafsky and Martin, 2000). Recall is the fraction of guesses out of the entire data set: for all possible selections, how many are accurately selected (Jurafsky and Martin, 2000). Usually there is a trade-off between precision and recall; the higher the precision the lower the recall.

For example, assume a student is given an examination with 10 questions. Suppose the student answers only seven of the questions, but he/she answers them all correctly. The student's precision is 100% (7 answers right out of 7 attempts) and their recall is 70% (7 correctly answered questions out of 10 total questions).

Threshold levels will be assessed experimentally to attempt to find the trade-off between precision and recall. Finding these levels would allow for desired precision levels to be set, and the corresponding recall levels will be known. Essentially one can set a desired accuracy level for the system to achieve and know how many of the diary entries will be correctly classified.

3.5 Hypothesis 3: Exploiting Semantic Ontologies

"The hierarchical structure of the CHAD database can be exploited for a more general model of human activity/location with higher classification accuracy."

3.5.1 The CHAD Database Structure

The CHAD database consists of semantic categories of locations in which humans are commonly found in and activities humans commonly participate in. These categories (referred to as semantic categories or semantic ontologies) are organized in a structure that is hierarchical in nature. Its representation structure is a five-digit code and the semantic category to which it is referring.

30000 - Residence, general
30010 - Your residence
30020 - Other's residence
30100 - Residence, indoor
30120 - Your residence, indoor
30121 - Kitchen
30122 - Living room / family room
30123 - Dining room
30124 - Bathroom
30125 - Bedroom
30126 - Study / office

This example of the CHAD database's shows it's hierarchical structure well. The first database entry is **3000 - Residence, General** this category encompasses all residence locations from a subjects residence, another's residence, and areas associated with a person's residence. Then the CHAD code is **30100 - Residence, indoor** and the subsequent entries are places within a residence

The general structure for the CHAD code database is a five-digit code. The numbering system is set up in a way that general categories are on shallower level and more specific categories are on a deeper one. The code representation reflects this hierarchical structure with the leftmost digit being the most general and the rightmost being the most specific. The structure goes as follows:

- The first digit: denotes whether the CHAD code is an activity or location code. The number three for locations and the number one for activities.

- The second digit: denotes a general category for location and activity.
    - Example:
        - 10000 - Work and other income producing activities, general
        - 11000 - General household activities
        - 12000 - Child care, general
        - 13000 - Obtain goods and services, general
        - 14000 - Personal needs and care, general

- The third digit: denotes subcategories in each category for location and activity
    - Example:
        - 10000 - Work and other income producing activities, general
            - 10100 - Work, general
            - 10200 - Unemployment
            - 10300 - Breaks
        - 11000 - General household activities
            - 11100 - Prepare food
            - 11200 - Indoor chores
            - 11300 - Outdoor chores
        - 12000 – Child care, general
            - 12100 – Care of baby
            - 12200 – Care of child
            - 12300 – Help / teach

- The fourth and fifth digits: denote granularities in the subcategories.
    - Example:
        - 17000 – Leisure, general
        - 17100 – Participate in sports and active leisure
        - 17110 – Participate in sorts
        - 17111 – Hunting, fishing, hiking
        - 17112 – Golf
        - 17113 – Bowling / pool / ping pong / pinball

Some categories utilize all five digits; some only three. As the examples show, the more digits that are used in the CHAD code, the more granular the category the code represents will be. The exploitation of the CHAD code structure is the basis behind the third hypothesis.

3.5.2 Primary Reasoning

The CHAD database's hierarchical structure is based on similar structures that exist in the real world. In the case of location, there are hierarchical relationships inherent to a person's location that can be delineated; the same is true for a subject's activities. Take the example of a person's residence. At the highest level of the residence, there are the sublevels of being outside or inside of the residence. For inside the residence there are semantic ontologies as well, upstairs *versus* downstairs and then all of the various rooms. If a person's location is the bedroom, their location is also the indoors of the residence and in the general area of the residence itself.

3.5.3 Exploiting the Semantic Ontology Structure

The CHAD database's structure allows us to easily decrease granularity with the hope of improving accuracy. This can be accomplished easily by harnessing the numbering system discussed in Section 3.5.1. By disregarding the rightmost digits, the number of which can be decided by the granularity required by the study, the word+context system can be modified to test less specific locations or activities. The system disregards digits in the testing process starting with the rightmost digit. The semantic ontology system will consist of three main ontologies: testing with a four-digit CHAD code, testing with a three-digit CHAD code and testing with a two-digit CHAD code. Testing less than a two-digit code would be meaningless as the left digit stands simply for location or activity. When the list of possible CHAD codes is generated, the semantic ontology system disregards digits it is not using, depending on the number of digits being used.

For example, if the system is testing the diary entry “I'm walking down the hallway gonna use the bathroom”, the location would likely be encoded as **30124 - Bathroom**. But a human could reasonably encode the location as **30129 - Other indoor** because it is not technically the bathroom or **30120 - Your residence, indoor** because it is actually inside of the house. If the test is run with a four-digit semantic ontology, the rightmost digit is disregarded and the system only tests the first four digits. Then all three of these reasonable categories would be classified as correct. The correct classification would be **3012x** defaulting to **30120 - Your residence, indoor** which is makes sense the hallway and the bathroom both are inside of the residence. By disregarding the rightmost digits in the testing process, a diary entry with multiple possible encodings could be accurately classified by sacrificing the granularity of the categories.

4 Experimental Results

4.1 Single Training and Testing Set

The word-only system achieved a location classification accuracy rate of 69.1% and an activity classification accuracy rate of 58.4%. The addition of context with the word+context system improved for location classification accuracy to 75.0% with a p-value of 0.0001, and activity classification accuracy to 66.1% with a p-value of 0.0007. The results are given in Table 3.

Table 3: Average word-only versus word+context CHAD location code classification accuracy for the single data set.

	Word-only	Word+context	Improvement	P-value
Locations	69.1%	75.0%	8.5%	0.0001
Activities	58.4%	66.1%	13.4%	0.0007

The average optimal weights found for this test are given in Table 4.

Table 4: Weight values that yields highest accuracy for both activity and location in the single data set

Activity				**Location**		
Per-word	w_1	0.35		**Per-word**	w_5	0.65
Previous Activity	w_2	0.3		**Previous Location**	w_6	0.15
Previous Location	w_3	0.1		**Previous Activity**	w_7	0.05
Current Location	w_4	0.25		**Current Activity**	w_8	0.15

4.2 Leave-one-out Testing

Classifying diary entries into semantic categories is a difficult process regardless of diary size or utterance quality. In this study, the data set is smaller than what would be desired for a system based on probability relationships within a data set to perform diary entry classification. The larger the diary, the larger the training set that can be made from

which the word+context system generates probabilities. Larger training sets allow for more numerous diary entries from with to analyze word and context relationships. Leave-one-out testing can be utilized to allow for more data sets and greater coverage of the full data set

Leave-one-out testing commonly consists of creating multiple testing and training sets. Each testing set consists of one data point and the training sets consist of the remaining data points. Typically, there are as many testing sets as there are data points in the data set, allowing for the testing of all of the data in the set.

The data set for this study consisted of 42 days of subject diary entries, 1220 diary entries in all. Since the chronological structure of the database needs to be retained in the leave-one-out data testing and training sets, a different segmentation of the data is used than was used in the single data set. As described in Section 3.2.2, the segmentation of the first set is comprised of one day of data from each subject in the study. For the leave-one-out data sets, the data segmentation is based on individual days and testing sets were generated from a day of subject data each leaving 42 testing and training sets in all.

The word-only system was tested on the leave-one-out data sets to determine its accuracy on multiple files, and then the word+context system is tested to see if there was improvement.

4.2.1 Word-only Testing

When the word-only system is applied to the leave-one-out data sets, the activity results are described in Figure 1, the location results are presented in Figure 2, and the combined weighted averages across the data sets is given in Table 5.

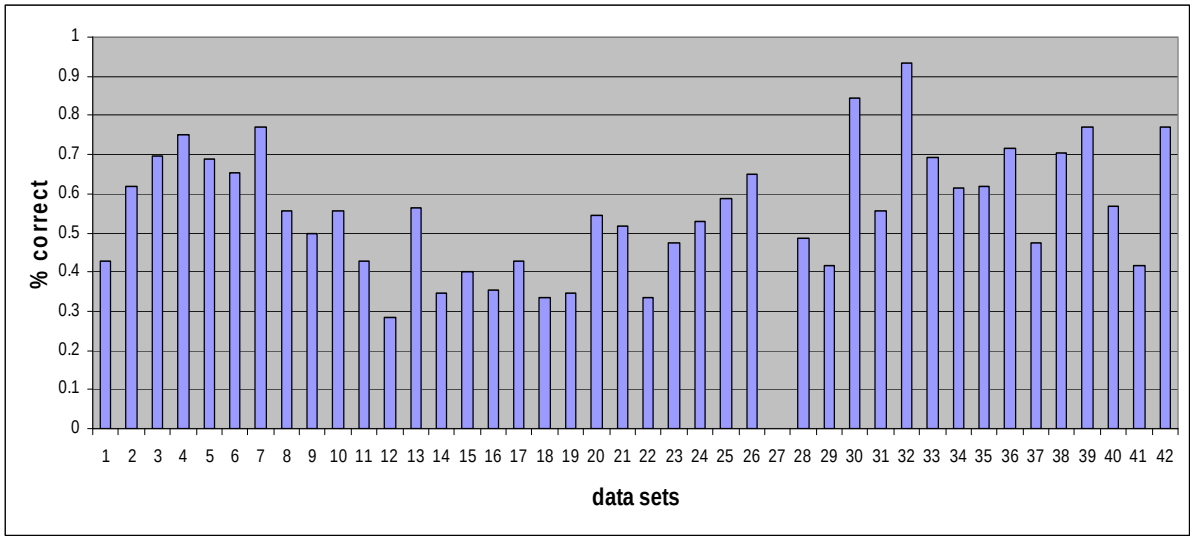

Figure 1: Word-only CHAD activity code classification accuracy for leave-one-out testing

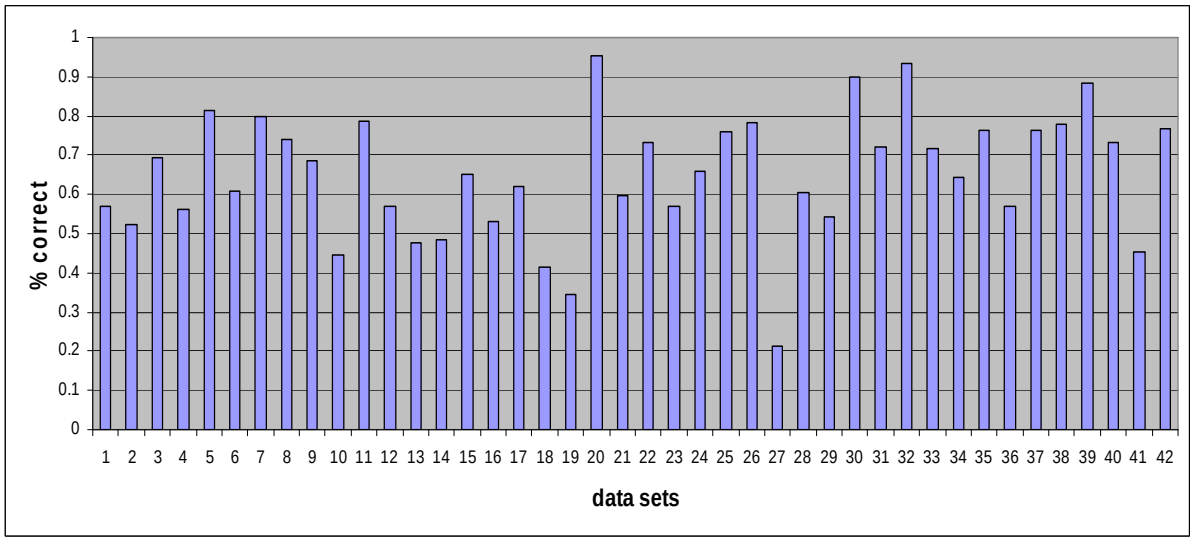

Figure 2: Word-only CHAD location code classification accuracy for leave-one-out testing

Table 5: Average word-only CHAD location code classification accuracy for leave-one-out testing

	Average	Max	Min
Location	65.5%	95.5%	21.4%
Activity	55.3%	93.3%	0%

4.2.2 With Contextual Information

When contextual information is combined with the word-only calculations, the activity results show improvement as illustrated in Figure 3. The difference between the word-only versus the word+context is graphically illustrated in Figure 4. Similarly, results for location are given in Figure 5 and Figure 6.

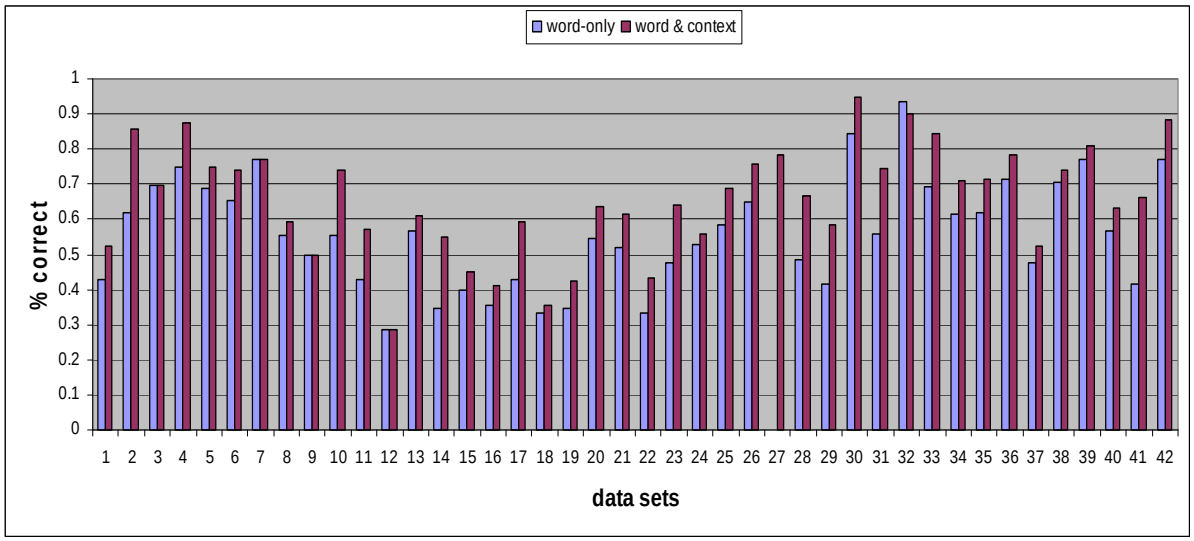

Figure 3: Word-only versus word+context CHAD activity code classification accuracy for leave-one-out testing

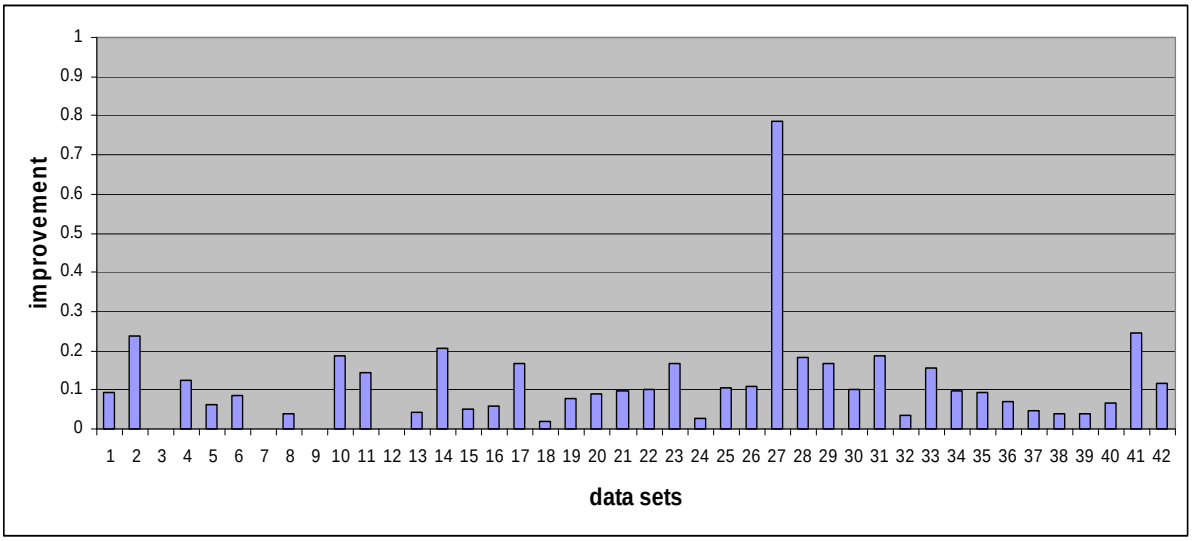

Figure 4: Word+context system improvements in activity classification accuracy over the word-only system for leave-one-out data sets

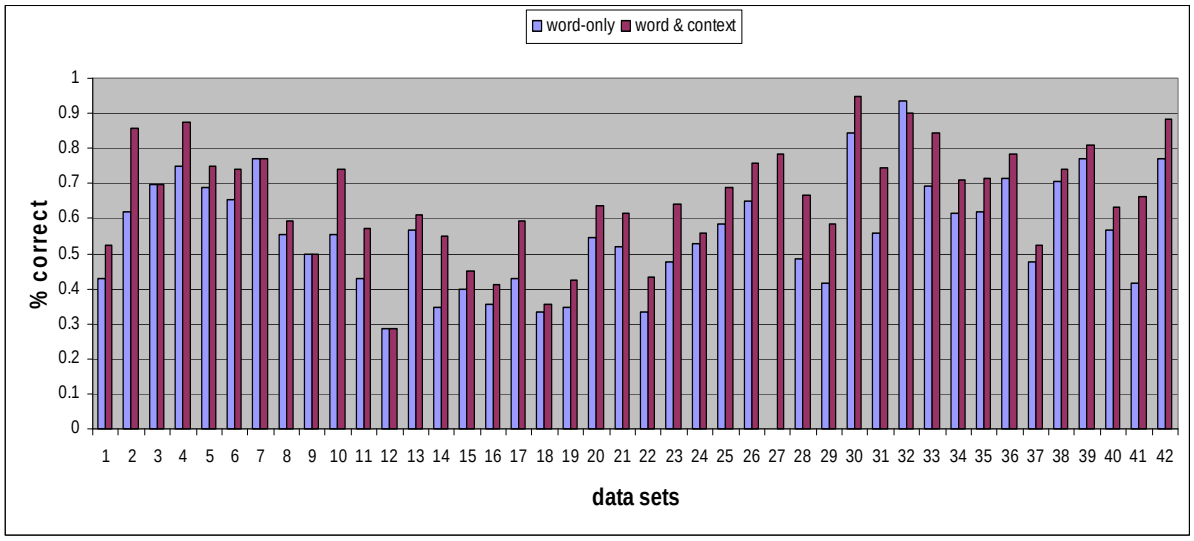

Figure 5: Word-only versus word+context CHAD location code classification accuracy for leave-one-out testing

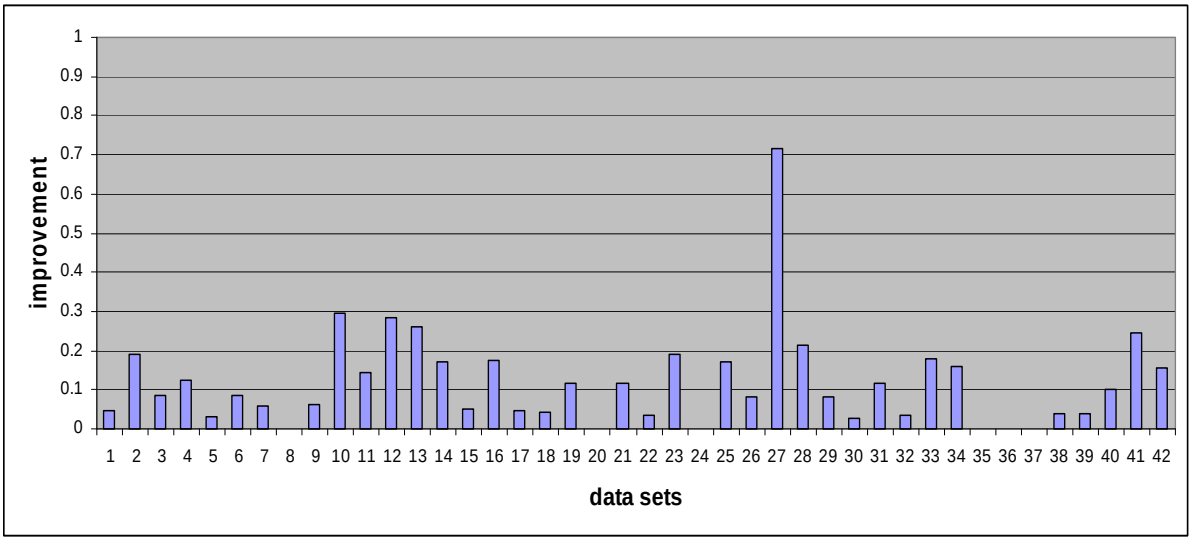

Figure 6: Word+context system improvements in location classification accuracy over the word-only system for leave-one-out data sets

The average accuracy results for activity and location are summarized in Table 6. The average weights obtained for the six context relationships and the words is given in Table 7.

Table 6: Average word+context it semantic ontology location classification accuracy improvement over the word-only system including % improvement

	Word-only	Word+context	Improvement
Locations	65.5%	76.0%	16.0%
Activities	55.3%	66.1%	19.5%

Table 7: Activity and location weights for score combination in the word+context system testing

Activity				Location		
Per-word	w_1	0.353571		**Per-word**	w_5	0.294048
Previous Activity	w_2	0.177381		**Previous Location**	w_6	0.146429
Previous Location	w_3	0.20119		**Previous Activity**	w_7	0.27381
Current Location	w_4	0.267857		**Current Activity**	w_8	0.285714

4.2.3 With Semantic Ontologies

4.2.3.1 Four Digit CHAD Code

When the semantic ontology system is combined with the context-word system using a CHAD code four digits in length, the activity results are summarized in Figure 7 and the location results are given in Figure 8. The marked improvement in location results can be more clearly seen in Figure 9 which graphs the difference between the five-digit results and the four-digit results for location classification.

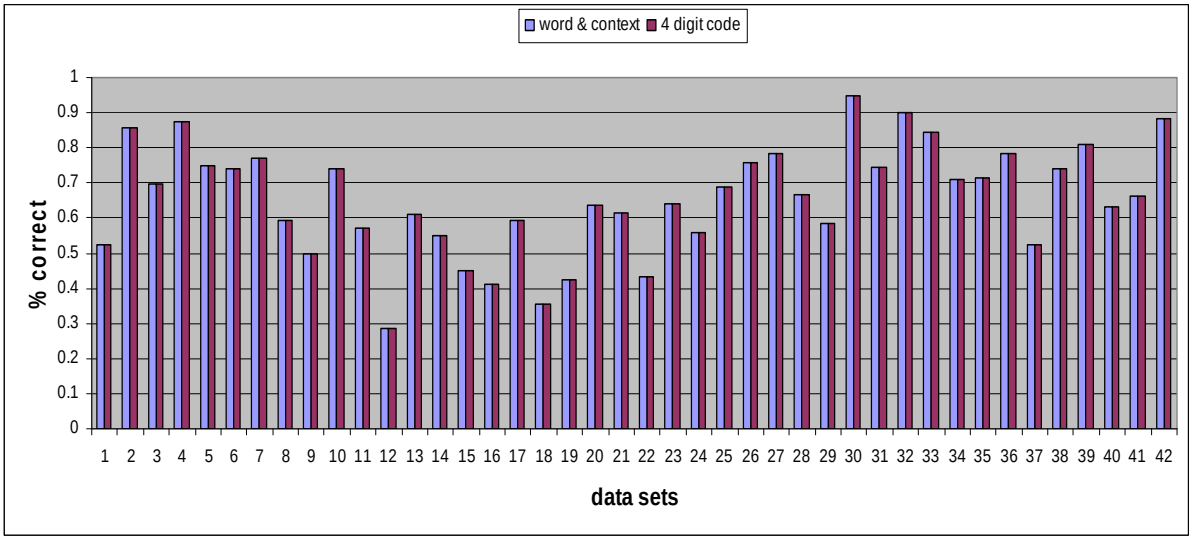

Figure 7: Word+context versus four-digit semantic ontology CHAD activity code classification accuracy for leave-one-out testing

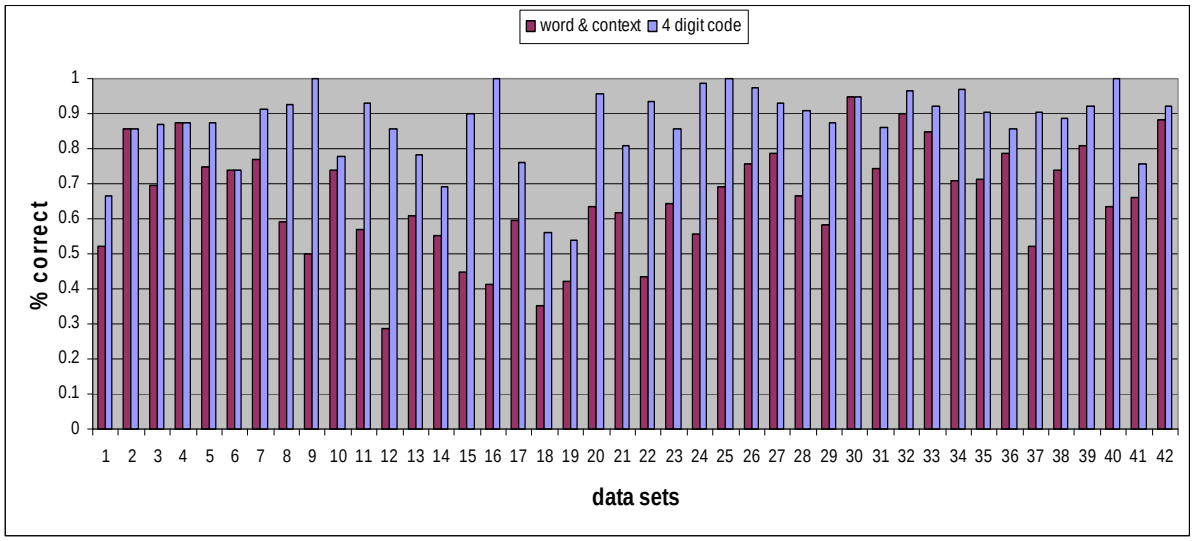

Figure 8: Word+context versus four-digit semantic ontology CHAD location code classification accuracy for leave-one-out testing.

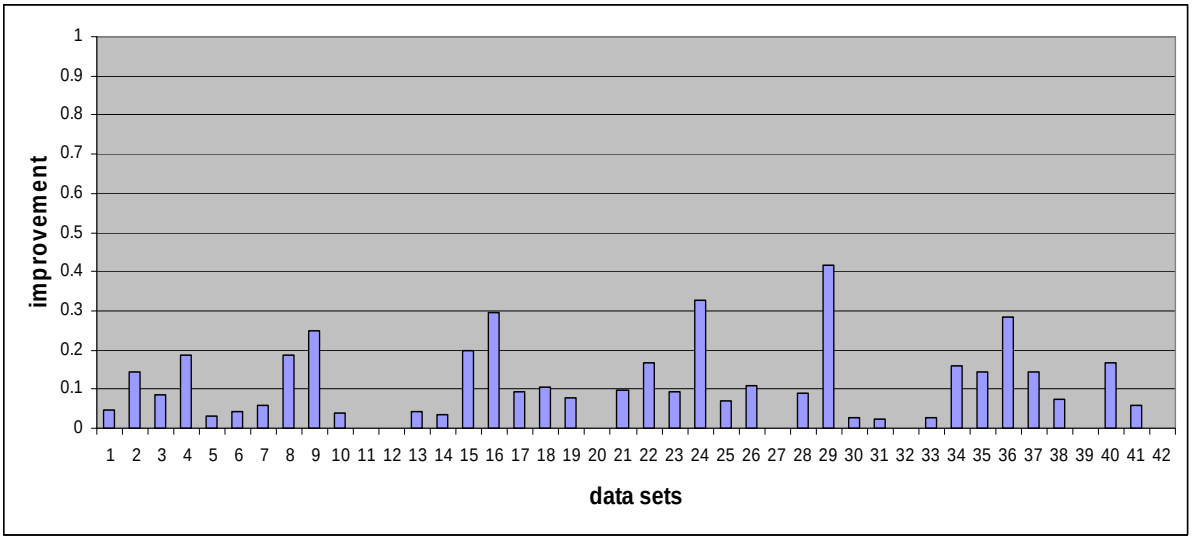

Figure 9: Four-digit semantic ontology system improvements in location classification accuracy over the word+context system for leave-one-out data sets

The average accuracy results for activity and location are summarized in Table 8.

Table 8: Average four-digit semantic ontology location classification accuracy improvement over the word+context system including % improvement

	Word+context	4 Digit code	Improvement
Locations	76.0%	86.6%	13.9%
Activities	66.1%	66.1%	0.0%

4.2.3.2 Three Digit CHAD Code

When the semantic ontology system is combined with the word+context system using a CHAD code three digits in length, the activity results are summarized in Figure 10 with the difference between the five-digit results and the three-digit results graphed in Figure 11.

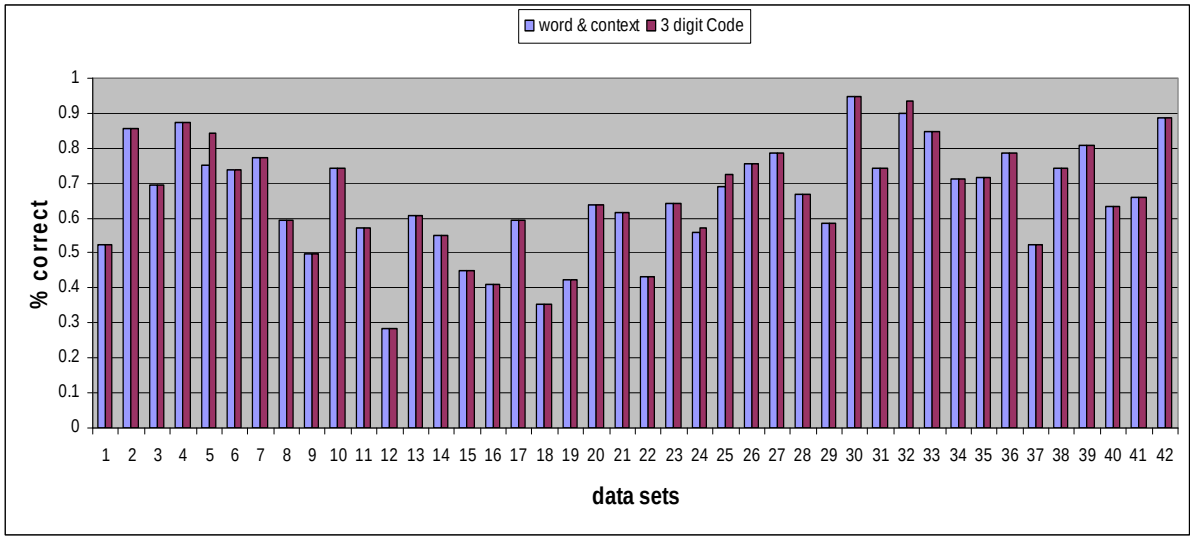

Figure 10: Word+context versus three-digit semantic ontology CHAD activity code classification accuracy for leave-one-out testing.

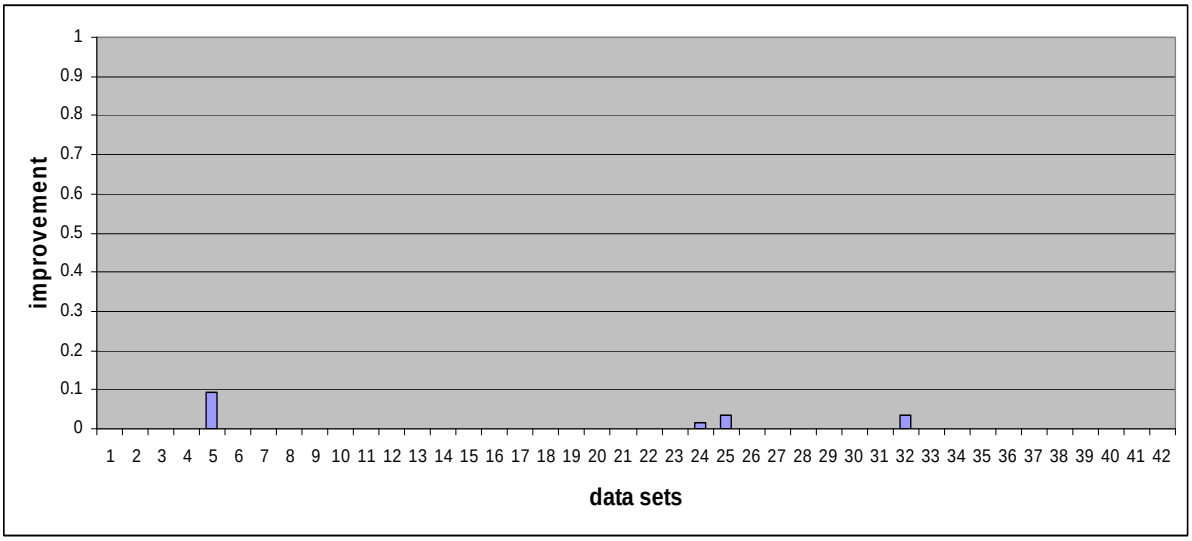

Figure 11: Three-digit semantic ontology system improvements in activity classification accuracy over the word+context system for leave-one-out data sets

The combined location three-digit semantic ontology system and five-digit word+context system results are shown in Figure 12 and Figure 13.

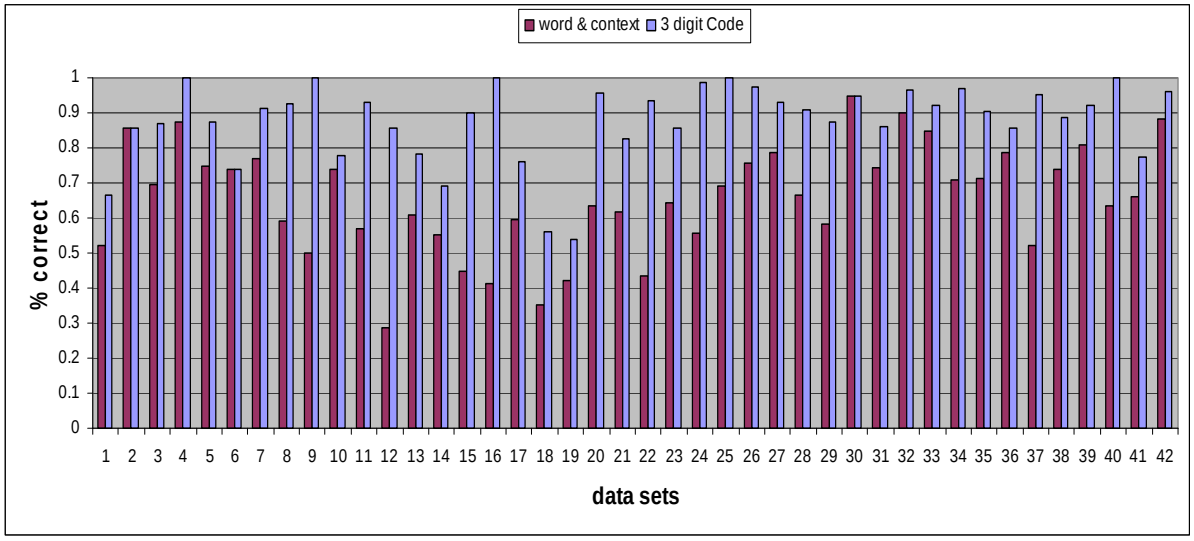

Figure 12: Word+context versus three-digit semantic ontology CHAD location code classification accuracy for leave-one-out testing.

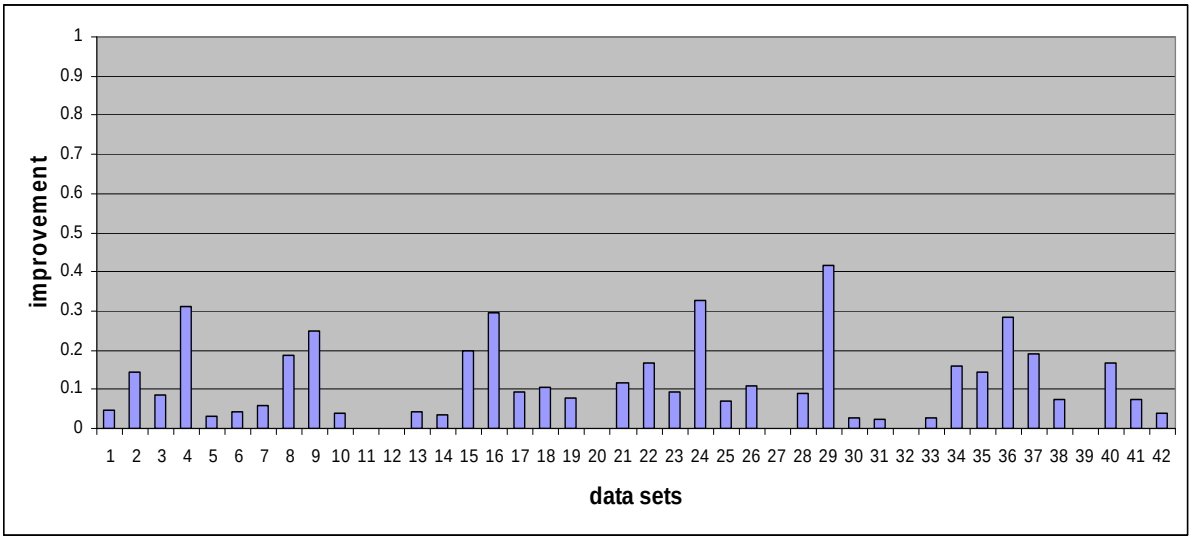

Figure 13: Three-digit semantic ontology system improvements in location classification accuracy over the word+context system for leave-one-out data sets

The average accuracy results for activity and location are for the three-digit code and its improvement over the five-digit code is summarized in Table 9.

Table 9: Average three-digit semantic ontology location classification accuracy improvement over the word-only system including the % improvement

	Word+context	3 Digit code	Improvement
Locations	76.0%	87.0%	14.5%
Activities	66.1%	66.5%	0.6%

4.2.3.3 Two-digit CHAD code

When the semantic ontology system is combined with the context-word system using a CHAD code two digit in length, the activity results are summarized in Figure 14 and Figure 15; and the location results summarized in Figure 16 and Figure 17.

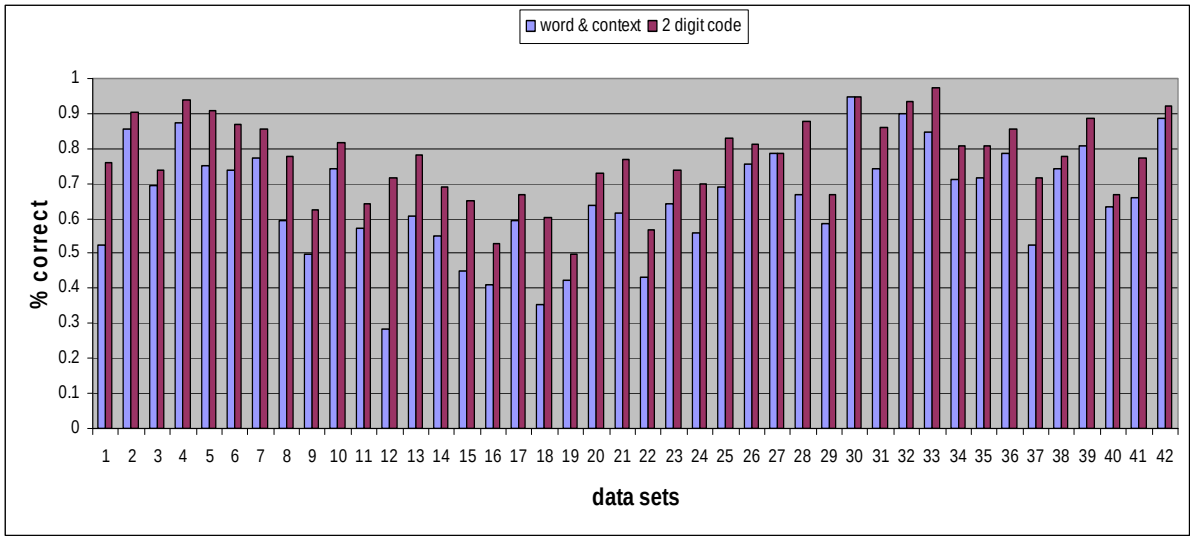

Figure 14: Word+context versus two-digit semantic ontology CHAD activity code classification accuracy for leave-one-out testing.

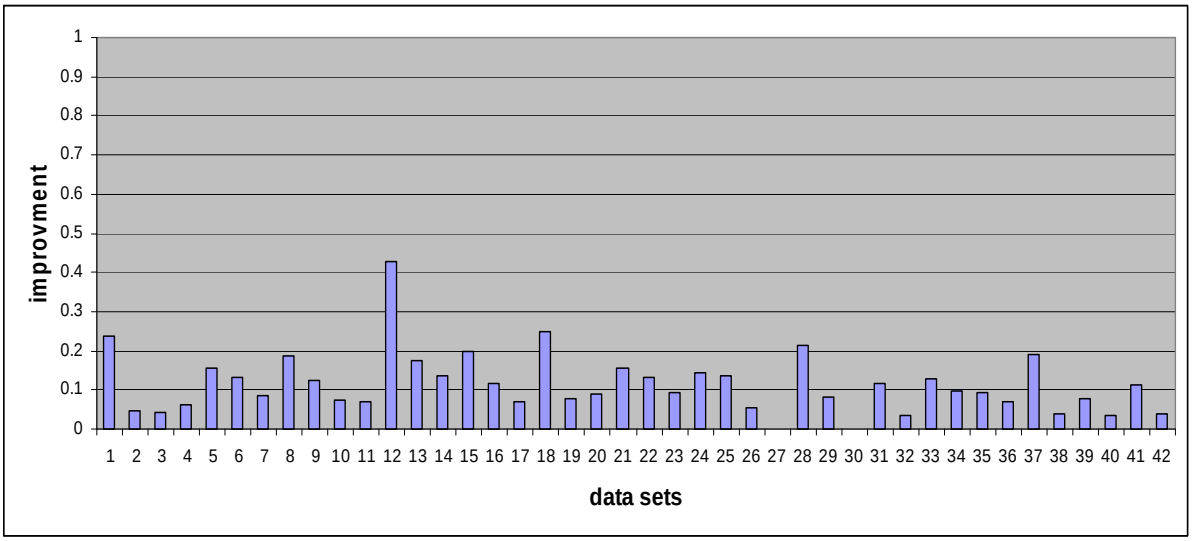

Figure 15: Two-digit semantic ontology system improvements in activity classification accuracy over the word+context system for leave-one-out data sets

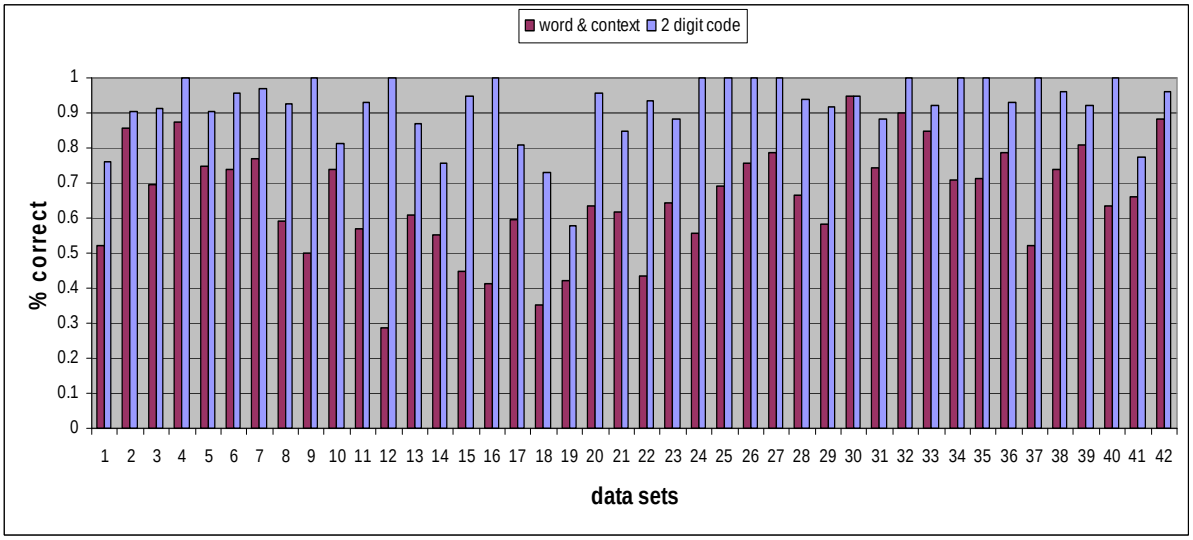

Figure 16: Word+context versus two-digit semantic ontology CHAD location code classification accuracy for leave one out testing.

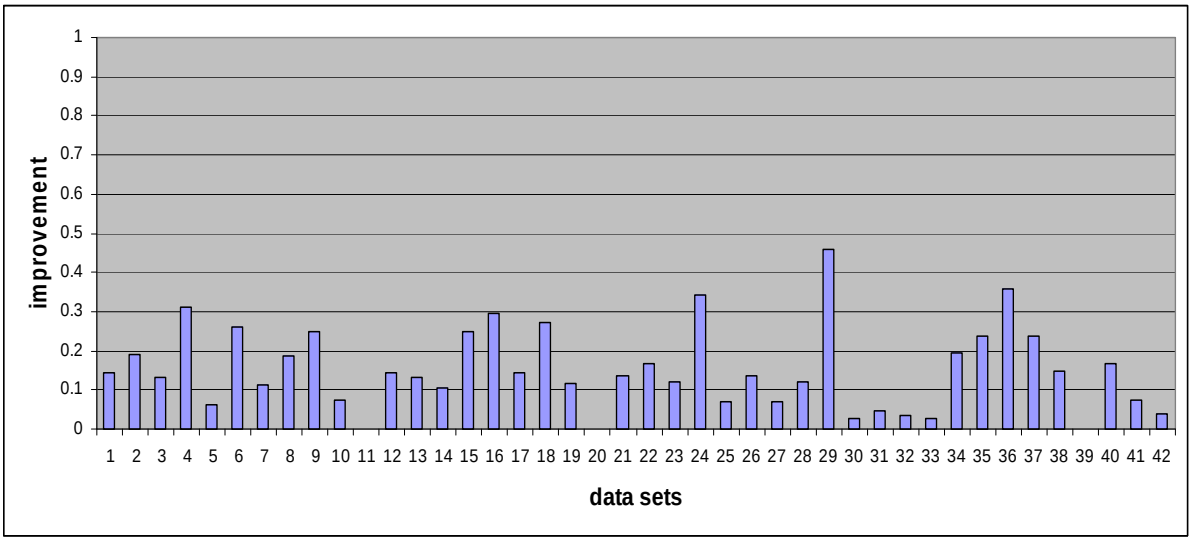

Figure 17: Two-digit semantic ontology system improvements in location classification accuracy over the word+context system for leave-one-out data sets

The average accuracy results for using the two-digit encoding for activity and location are summarized in Table 10.

Table 10: Average two-digit semantic ontology location classification accuracy improvement over the word-only system including the % improvement

	Word+context	2 Digit code	Improvement
Locations	76.0%	90.8%	19.5%
Activities	66.1%	77.4%	17.1%

4.2.4 With Thresholds

Precision and recall results thresholds in increments of 0.1 between the two top CHAD Activity scores are graphed in Figure 18 and CHAD Location scores are graphed in Figure 19.

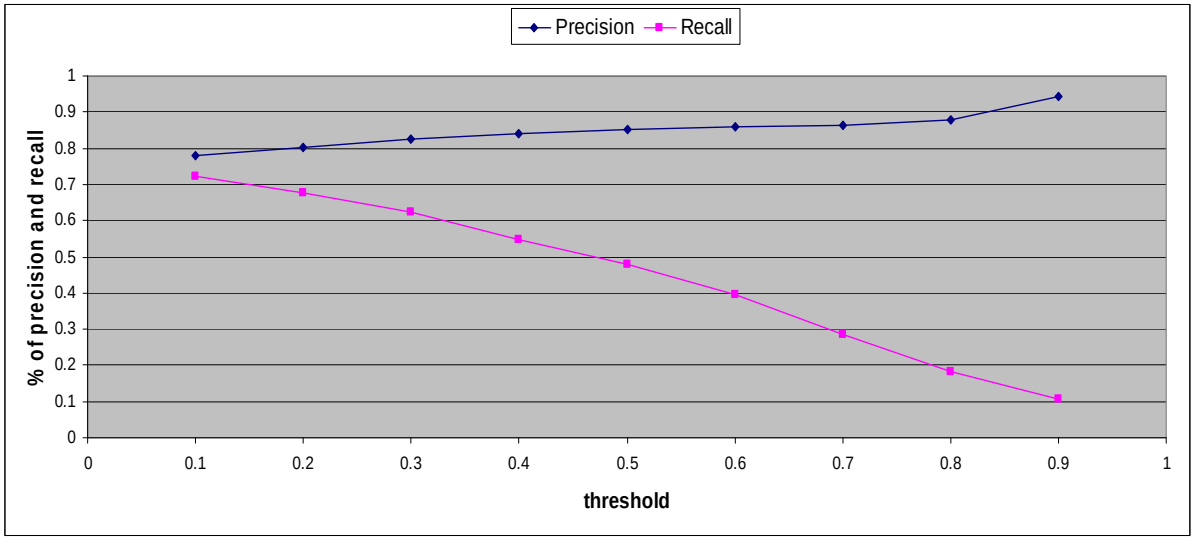

Figure 18: Precision and recall results for difference threshold levels ranging from 0.1 to 0.9 for activity

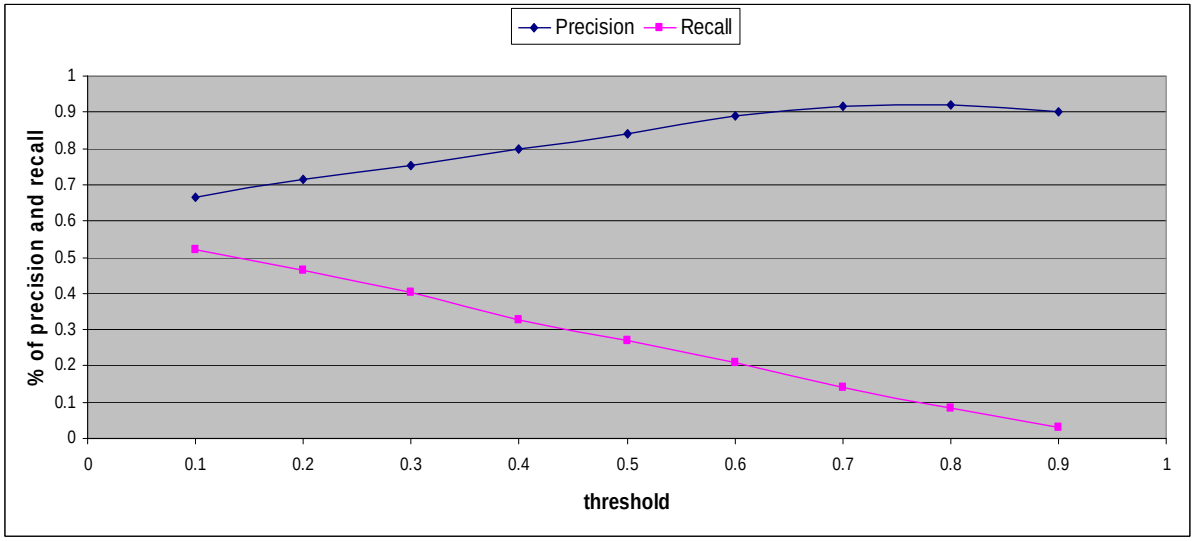

Figure 19: Precision and recall results for difference threshold levels ranging from 0.1 to 0.9 for location

5. Discussion

5.1 Word-only System

5.1.1 Results

The word-only system performs extremely well when tested the training data. Using the word-only n-grams produces 97.5% Location CHAD code classification accuracy and 96.8% Activity CHAD code classification accuracy. This result makes sense considering the fact that the word-only system is trained on the data it is being tested upon. The high results from the testing of the training sets show the word-only system's ability to accurately classify diary entries.

When testing with new diary entries from the test set (which was disjoint from the training set), however, the accuracy of the word-only system drops to 69.1% for Location and 58.4% for Activity CHAD codes. The testing results are lower than what was expected for the testing corpus although the training set was perceived to be large and diverse. The training set was three times larger than the testing set, and the data was organized so that all participants were included in the testing set. Clearly, the word-only system performs extremely well when applied to the training set, but the performance declines precipitously when applied to the disjoint testing set

5.2 Word+context System

5.2.1 Results

As detailed in Section 4.2.1, the word-only system yields accuracy rates below 70% for location and below 60% correct for activity. The word+context system improved classification accuracy for both location and activity. The location classification accuracy

improved to 75%, an increase of 8.5%, and the activity classification accuracy improved to 66.1%, an increase of 13.4%. Similar to the word-only study, the location classification accuracy was higher that that of the activity, but the gap between location and activity accuracies fell. The gap between the accuracy for locations and activity decreased from 18.3% difference to 13.3% difference. Context had a greater impact on the classification accuracy of activities than it did locations. How much of an impact can only be known by examining the weights used in the scoring of the diary entries.

5.2.2 Weights

A working hypothesis was that the words in the utterance would be of primary importance, and the context would serve only to augment the score. This assumption held true with Location CHAD code classification but not for Activity CHAD code classification. The optimal weights for location and activity are given in Table 4 and Table 11.

Table 11: Table 4 repeated for clarity

Activity				**Location**		
Per-word	w_1	0.35		**Per-word**	w_5	0.65
Previous Activity	w_2	0.3		**Previous Location**	w_6	0.15
Previous Location	w_3	0.1		**Previous Activity**	w_7	0.05
Current Location	w_4	0.25		**Current Activity**	w_8	0.15

In the location classifications, 65% of the overall score was derived from the words in the diary entry, and the three context calculations only contributed 35%. The previous location and the current activity received the same weight of 15% each and the previous activity was the least important with only 5% contribution to the overall score. The

proportion between the context weights makes sense under the intuition that, of the three available context calculations, the most likely indicators of a subject's current location are: where their previous location was and the activity in which the subject is participating.

On the other hand, the activity weights were quite different. The word-only calculations contributed less to the activity classification accuracy, roughly half as important comprising only 35% of the total score. The previous activity was nearly as relevant to the classification accuracy at 30% of the overall score. The third most relevant score was the current location with 25%, with the previous location as the least relevant with only 10% of the overall score. The low relevance of the words in the dairy entries to activity classification accuracy was a surprising result.

Based on the test results of the word+context system it was expected that the words were going to play a diminished role. What was not expected is that word relevance would be less than half of the overall score for CHAD code classification. In fact, compared to the results of the location weights, the proportions of words to context was reversed with only 35% score relevance derived from words and 65% derived from context.

A similar weight proportion was examined for both location and activity in the weights for the six context formulas. For location the proportions of the three context scores was 3:3:1 (three for current activity; three for previous location; one for previous activity). Similarly, for activity, the proportions for the three context scores were 3:3:1 (three for current location; three for previous activity; and one for previous location). The context formulas for location and activity are mirrors of one another, where activity calculations use: current location, previous location and previous activity, and the location calculations use; current activity, previous activity, and previous location. Essentially they are similar

contextual constructs, the word-only type, activity or location being tested (referred to as the *test type*), and the context being added from surrounding diary entries (referred to as the *other type*). The three general categories are: the test type and the current other type (weights two and six), the test type and the previous test type (weights three and seven), the test type and the previous other type (weights four and eight). The ratio for both location and activity weights showed that weights two and six and weights three and seven were three times as relevant to the overall score as were weights four and eight.

The results seem to suggest that, for location, the current activity and the previous location are relevant, while the previous activity is negligible. The same is true for activity, the current location and previous activity are relevant, but the previous location is negligible.

5.2.3 Examples

Example 1: Current diary entry

Diary Entry: "back from my walk"

- Correct Activity: **11000 - General household activities**
- Current Location: **30120 - Your Residence indoor**

Table 12: Comparison between top word-only scores and word+context scores for example diary entry "back from my walk"

Top 3 Location word-only choices:	Top 3 Location word+context choices:
1. 0.652 - 10000 - Work and other income producing activities, general	1. 0.259 - 11000 - General household activities
2. 0.214 - 11000 - General household activities	2. 0.243 - 10000 - Work and other income producing activities, general
3. 0.120 - 18000 - Travel, general	3. 0.119 - 11800 - Care for pets/animals

In this example, the word-only system chooses the CHAD code for **10000 - Work and other income producing activities** as the subject's current activity. The score is three times larger than the next score which was the correct one. The subject's current location is their residence, not their place of business. Based on the probabilities within the training set,

the system overwhelmingly chooses an incorrect score. With the context information of the current location added to the probability of the overall score, the word+context system's highest scoring CHAD code is the correct one. The scores are close, but the classification is correct. The diary entry itself is a difficult one to classify in general without some surrounding context, it's not difficult to see why the word-only system has difficulty with this diary entry.

Example 2: Previous diary entry

Diary Entry: "I'm sitting down now eating pizza"

- Correct Location: **30121 - Kitchen**
- Previous Activity: **11110 - Prepare and clean-up food**
- Previous Location: **30121 - Kitchen**

Table 13: Comparison between top word-only scores and word+context scores for example diary entry "I'm sitting down now eating pizza"

Top 3 Activity Word-only Choices:	Top 3 Activity Context + word Choices:
1. 0.285 - 30125 - Bedroom	1. 0.320 - 30121 - Kitchen
2. 0.127 - 32100 - Office building/bank/post office	2. 0.285 - 30122 - Living room / family room
3. 0.031 - 35200 - Public garage / parking	3. 0.254 - 30125 - Bedroom

In this example, the word-only system classifies the CHAD code **30125 - Bedroom** as the subject's current location. In fact, the correct CHAD code is not present in the top three scores. The subject's previous location was **30121 - Kitchen** and his previous activity was **11110 - Prepare and clean up food**. The fact that the correct CHAD code receives such a poor score in the word-only system means that context played a more important role in this diary entries classification.

Example 3: Previous diary entry

Diary Entry: “in the office at the computer”

- Correct Location: **30126 - Study or Home Office**
- Previous Location: **30122 - Living room / family room**

Table 14: Comparison between top word-only scores and word+context scores for example diary entry “in the office at the computer”

Top 3 Location Word-only Choices:	Top 3 Location Context + word Choices:
1. 0.904 - 32100 - Office building/bank/post office	1. 0.502 - 30126 - Study / Office
2. 0.217 - 32900 - Public building/library/museum /theater	2. 0.349 - 32100 - Office building/bank/post office
3. 0.053 - 35200 - Public garage / parking	3. 0.212 - 32900 - Public building/library/museum /theater

This is one of the better examples of the word+context system’s ability to disambiguate a diary entry where the word-only system fails to. Here the subject is using the computer in his/her home office. The human encoder used context when encoding this entry thus the encoder knows that the subject is in his/her home office and not in an office building. The word-only system does not have enough information to encode this diary entry correctly, and indeed, if a human attempts to encode this utterance, without context, the human encoder may well improperly encode it too. Looking at only the words in the diary entry it, seems obvious that the person is at his/her place of business and he/she is working at the computer. Only with the addition of context information can diary entries such as these be correctly encoded

5.3 Leave-one-out Testing Results

5.3.1 Word-only

The word-only system was tested on the leave-one-out data sets to contrast with the word+context system’s classification accuracy. The average results are similar to the single data set’s results with an average accuracy rate of 65.4% for locations and 55.3% for

activities. These accuracies are slightly lower than that of the single data set, 7.2% lower for location accuracy and 5.6% lower for activity accuracy.

5.3.2 Word+context

The word-only system, as noted above in Section 4.2.1, achieves an average accuracy of 65.5% for locations and 55.3% for activities. When the context data is combined with the word calculations, there is a marked improvement. Applying the context system to the test sets, the location accuracy improves to an average of 76.0%, an improvement of 16%, and the activity accuracy improves to an average of 66.2%, an improvement of 19.5%. This improvement is greater than that of the single data set for location at 1.3% higher accuracy, but the activity accuracy is unchanged.

The leave-one-out data sets perform worse in the word-only testing, but they performed better than the word+context system than the single data set. When the entire data set is considered, context has a greater impact on improving the CHAD code classification accuracy than when a single data set is constructed. To examine the impact of context on classification accuracy, the weights from the leave-one-out tests will need to be analyzed.

5.3.3 Weights

With the leave-one out testing weights, there is no single weight configuration that provides optimal results for all of the data sets. Further research will have to be done to see if there is a pattern, range or some other method for generalizing weights for all leave-one-out data sets. The average weights that performed well are detailed in Table 7.

Table 15: Table 7 repeated for clarity

Activity				Location		
Per-word	w_1	0.353571		**Per-word**	w_5	0.294048
Previous Activity	w_2	0.177381		**Previous Location**	w_6	0.146429
Previous Location	w_3	0.20119		**Previous Activity**	w_7	0.27381
Current Location	w_4	0.267857		**Current Activity**	w_8	0.285714

These results are surprising due to the low relevance words play in correctly classifying the diary entries. The results returned from the system showed that context was weighted at 65% of the score for activity and 70% of the score for location. The word-only study produces relatively high accuracy without the use of any context information; consequently, the assumption that the words alone would be the most important factor with the context factors simply augmenting the word score. This assumption holds true in the single data set test results for location but not activity. The observation that context constituted approximately 65% of the total score shows the assumption was incorrect, and that context holds more importance for classifying diary entries of this type.

The proportions between context calculations that was observed in the single data set were different for the leave-one-out testing results. The largest difference was the relevance of weights four and eight (the test type and the previous other type). In the single data set tests, weights 4 and 8 were the least relevant to the score calculation for both activity and location. In the leave-one-out testing weights 4 and 8 were the second most relevant calculations to the overall score, comprising 27% of the location score and 20% of the activity score. The discrepancy between the single data set weights and the leave-one-out weights seems to imply that more research is needed to determine weight relevance.

5.4 Threshold Results

Threshold levels were tested in increments of 0.1 to see the system's ability to give a confidence valuation to the top scoring CHAD codes. As predicted, the higher the precision of the system achieved, the lower the recall fell.

The precision and recall numbers for location started higher than expected for the 10% difference threshold. The precision was close to 80% and the recall was close to 75%. This means that the word+context system can classify, with 80% accuracy, 75% of the total testing set correctly. At the other end of the spectrum, when there is a 90% difference between the top 2 scores the precision is nearly 95% while the recall is near 10% of the testing set.

For activity classification the precision and recall numbers were lower. For a 10% difference threshold, the precision was approximately 65% and the recall was just below 55%. With a threshold difference level of 90% the precision was approximately 90% but the recall was approximately 4%. The precision and recall numbers for activity were lower than those for location. This indicates that the relationship of the top two scores is less relevant for activity classification than it is for location classification. At the same time, results achieved with the 10% difference threshold are quite promising, the high precision and recall numbers could prove immensely useful in the application of the word+context system.

The weights chosen for the threshold testing are the optimal weights found in the experimental testing. It was decided that the best way to test the threshold system is to use the weights that returned the highest classification accuracy. Thus, the threshold results would be representative of the best known precision and recall values achievable.

5.5 Semantic Ontology Results

The results from the semantic ontology results were: a higher average accuracy was attained by applying the semantic ontology system for both location and activity, but the overall numbers ended up different than expected. There were three general applications of the semantic ontology system: a four-digit CHAD code, a three-digit CHAD code and a two-digit CHAD code. Each application returned different results when tested on the leave-one-out data, with improvement for both location and activity.

5.5.1 Four Digit Semantic Ontology Results

For location, the four-digit semantic ontology improved the accuracy from 76% correct classification to 86.6% correct classification, in total a 14% improvement over the word+context system. This improvement included bringing the classification accuracy of four of the leave-one-out data sets up to 100%. The improvements in this ontology level were greater for Location CHAD classifications than they were for Activity CHAD classifications.

Primarily, the improvements in location classification accuracy were attained from the subject's residence, and the subcategories within a subject's residence. The code structure for a subject's residence is as follows:

30100 - Residence, indoor
30120 - Your residence, indoor
30121 - Kitchen
30122 - Living room / family room
30123 - Dining room

Utilizing the four-digit ontology system classifies all subcategories within a subject's residence as the same code. This homogenizes all of the residence codes into a single code, removing any confusion between areas within the home.

For activity however, the four-digit semantic ontology had no affect on the classification accuracy. As can be seen in Figure 7 for each leave-one-out results and in Table 8 for the total average, the scores were unchanged. This was surprising as it was assumed that there would be improvements for each of the semantic ontology systems, even if the improvement was small. The fact that there was no improvement in the activity four-digit ontology accuracy when the accuracy for location improved as well as it did raises questions about the structure of the Activity section of the CHAD database.

5.5.2 Three Digit Semantic Ontology Results

This level of the semantic ontology structure within the CHAD data base consists of removing the codes that deal with granularities within the subcategories, and simply focus on the subcategories themselves. In the case of residence, when the last two digits are disregarded all granular categories for residence are classified into the CHAD code **30100 - Residence indoor** together. The **30100 - Residence indoor** CHAD code includes the subject's residence and all of its rooms, other people's residence and all of their rooms, and the other indoor categories.

The three-digit semantic ontology improved in accuracy from 76% correct classification to 87% correct classification for location. This improvement was 14.5% more accurate than the word+context system, but only 0.5% accuracy improvement over the four-

digit semantic ontology system. The three-digit semantic ontology system's improvement included five leave-one-out test sets with 100% classification accuracy.

One reason there was little improvement for location accuracy at this semantic ontology level is the makeup of the database. Many of the categories for location are in the last digit of the CHAD code system. Most of the discrepancies for location in the deeper semantic ontology levels occurred at the five-digit level not the four-digit level (for example a subject's residence). Disregarding the fourth digit made a much smaller impact than did disregarding the fifth digit.

There was a small amount of improvement using the three-digit semantic ontology for activity. The activity classification accuracy improved from 66.1% to 66.5%, an improvement of 0.6%. This result does show that the semantic ontology system can have some impact on improving activity classification accuracy. The improvement was small in comparison to the accuracy improvements of the word+context system over the word-only system, and the location accuracy improvements with the four-digit semantic ontology system. The low improvement could be due to the structure of the CHAD database and the difference between the Activity and Location CHAD codes.

An interesting observation about both the location and activity results for the three-digit semantic ontology is the similarity of their improvement. Both activity and location accuracy improved about ½ percent over the three-digit semantic ontology. The similarity between location and activity accuracy improvement is surprising in the lack of considerable improvement. Consolidating all granularities with the subcategories and utilizing only three digits of the CHAD code system did not have much of an improvement to the overall accuracy for either location or activity.

5.5.3 Two Digit Semantic Ontology Results

This level of the semantic ontology structure within the CHAD data base consists of removing codes for subcategories and granularities within them and focusing on the general categories themselves. Consider the residence example again; disregarding the three last digits places all subcategories of **30000 - Residence general together**. This includes indoors and outside of the residence, other people's residences, and all areas in the same category. The same is true for other categories as well, such as:

- Travel - all codes consolidated into **31000 - Travel general** CHAD code
- General indoors - all consolidated into codes **32000 - Other indoor** CHAD code

In the case of Location CHAD code classification, the two-digit semantic ontology raised the accuracy from 76% correct classification to an accuracy of 90.8% correct classification. This is an improvement of 19.5% over the word+context system, an improvement of 4.8% over the four-digit semantic ontology accuracy and an improvement of 4.4% over the three-digit semantic ontology accuracy. Along with high accuracy improvement, 13 datasets achieved 100% classification accuracy, adding eight data sets to the total number of correctly classified data sets for location.

The primary reason for the improved accuracy was the limited number of categories within location when only two digits are present. When the three last digits are ignored, only seven categories remain in the location section of the CHAD database. Thus, the system only had seven choices with which to classify a diary entry's location.

For activity, the two-digit semantic ontology system drastically improved accuracy over the word+context system. The activity accuracy improved from 66.1% to 77.4%, an improvement of 17.1% over the word+context and the four-digit semantic ontology system, and an improvement of 16.4% over the three-digit semantic ontology system. This is a substantial increase over the word+context score, nearly as much of an increase in accuracy as the two-digit semantic ontology system did for location.

5.5.4 Distinctions Between the Location and Activity CHAD Code Semantic Ontologies

It is clear from the semantic ontology results that location and activity improvements, though similar in improvement percentage, differed in the source of the improvement. Location accuracy gained its greatest increase in accuracy from the four-digit semantic ontology system. The last digit was the most important in improving the accuracy and its importance may be due to the nature of the diary entries in this study.

The subject's residence was one of the primary locations from the EPA study. There are many diary entries taken from within the home that were either long or contained multiple locations and activities. Diary entries such as these made the classification process difficult the human encoder, and even more difficult for the word-only system to correctly classify. Take the example diary entry "have been and will continue to be going back and forth between the kitchen and living room preparing lunch and doing laundry". It is difficult for a human to accurately classify this diary entry into a single Activity and Location CHAD code. It seems that the person recorded two different locations at one time. The words present in the diary entry play an important part in the calculation of the correct CHAD code score. With a diary entry of this nature, the word+context system will have difficult

differentiating which location category this diary entry should be classified into. The four-digit semantic ontology system is able to classify this diary entry correctly because each of the individual locations mentioned in the utterance (**30121 - Kitchen** and **30122 - Living room/family room**) are both subcategories of the general category **30120 - Your residence indoor**. Due to the high frequency of diary entries occurring in the subject's residence, location classification accuracy was improved the greatest with the four-digit semantic ontology system.

The semantic ontology results were quite different for activity. There was no improvement to the activity accuracy for the four semantic ontology system, and little more for three-digit semantic ontology system. The accuracy was only greatly affected by the two-digit semantic ontology system, improving by 17.5% over the word+context system. Where the improvement of the location score was primarily in the four-digit system, areas of the most granularity, the activity score was primarily affected by the areas of least granularity.

There a two possible reasons for the lack of improvement in the four-digit semantic ontology system for activity accuracy when there was a high improvement in location accuracy. The granularity of the activity section of the CHAD database did not adequately utilize a semantic ontology structure in the fourth and fifth digits, or the diary entries from the EPA consisted of activities that did not utilize the fourth and fifth digits like the locations did.

5.5.5 Semantic Ontology Weights

The weights chosen for the semantic ontology testing were the optimal weights found in the exhaustive testing of the word+context system. It was decided that the best way to test

the semantic ontology system was to use the best weight combinations found for word+context calculations. Thus, the semantic ontology results would be representative of the highest achievable accuracy the system could provide, combining the words, the surrounding context and semantic ontologies.

6. Conclusions

6.1 Hypothesis 1: Adding Context to Improve Precision

> "Performing statistical NLP text abstraction using multi-diary entry contextual information will improve the disambiguation of human spoken diary entries over the word-only n-gram model applied to single diary entries."

When contextual information was combined with the word-only system, classification accuracy improved for both location and activity. In the single data set testing, the activity classification accuracy improved by 13.4% and the location classification accuracy improved by 8.5%. Whereas when the 42 leave-one-out data sets were tested, the average activity classification accuracy improved by 19.5% and the average location classification accuracy improved by 16.0%. In both experiments, single data set and leave-one-out data set, the diary entry classification accuracy improved for location and activity. The consistent classification accuracy improvement shows that adding context information can improve a word-only n-gram system.

A surprising result was the weighing of the word-only score relative to the weights applied to the context information. Initially, it was thought that words within a diary entry would be the most important factor in high classification accuracy. The weights from the single data set experiment showed the relevance of context information was different for location and activity. 65% of the total score for location was from derived the words and only 35% of the total score was derived from the words for activity. The weights from the leave-one-out tests were a somewhat different. The word relevance for activity was still low at 29% but for location word relevance was also low dropping to 35%. In both experiments,

the weight given to the context information was higher than the weight given to the word-only score.

6.2 Hypothesis 2: Using Thresholds to Balance Precision and Recall

"Thresholds can be found to balance trade-offs between precision and recall."

When the threshold system classified new utterances, it used the optimal word+context weights, and threshold levels were generated and tested. The threshold levels were set at intervals of 0.1 between zero and one, and the precision and recall trade-offs emerged at higher levels than expected.

An assumption was held about the relationship between the threshold level and the precision of the system; the lower the difference between the highest scores, the lower the precision will be. This assumption, when combined with the general relationship between precision and recall stated in Section 3.4.5 (that the higher the precision is the lower the recall will be), meant that for low threshold levels, the precision would be low as well. The results show that the assumption proved to be true; higher setting of threshold lead to higher precision and lower recall.

The results for the threshold system test show that threshold levels can be set and corresponding precision and recall values can be supplied for location and activity. For example, if a situation required an accuracy rate of 85%, the breadth of the data set that could be classified at that rate would be known. The results from this study show the recall rate would be 48% for location and 21% for activity.

6.3 Hypothesis 3: Using Semantic Ontologies to Improve Precision

"The hierarchical structure of the CHAD database can be exploited for a more general model of human activity/location with higher classification accuracy."

The semantic ontology system results showed both location and activity accuracy can be improved upon by utilizing the hierarchical structure of the CHAD database. Location classification accuracy improved with each semantic ontology level, eventually improving by 19.5%. Activity, on the other hand, had no improvement from the two-digit; negligible improvement from the three-digit; and massive improvement from the four-digit semantic ontology level, improving by 17.1%. Thus in the case of both activity and location, utilizing the hierarchical structure improved diary entry classification accuracy.

The location results showed most of its improvement in the four-digit semantic ontology system (71% of total improvement) due primarily to the generalization the subject's residence. Study data of this type will have many diary entries from within the home and it can improve accuracy by generalizing all areas of the home into one category. The four-digit ontology level improved location classification accuracy the most, and proves that the semantic ontology system can garner higher results if less granularity is required.

The activity results were nearly the opposite in the case of the best performing semantic ontology level. The two-digit semantic ontology system most improved classification accuracy (97% of total improvement), in fact provided the bulk of the increase. Human activities are more difficult to arrange hierarchically as there is often no single obvious "parent" for an activity. There are redundancies in the CHAD activity database that make any encoding ambiguous, whether done by human or computer. Locations, generally,

are much more inherently hierarchical whether the location is a country, state, county, city, etc. or the location is a neighborhood, house, floor, room. Activities are not as hierarchical in nature, and the range of possible activities is more diverse than locations. Thus activity classification accuracy only improved when the categories became so few that the hierarchical structure was gone, and there were only 9 very general categories.

6.4 Limitations

6.4.1 Optimal Classifier

In order to directly compare this study's results with the results from the previous study, a naïve Bayes classifier is used to calculate n-gram probabilities for the semantic category classification. Bayes is an optimal classifier if the data satisfies the constraints that it is multivariate, normally distributed. If it is not, there exist a number of classifiers that would perform better. It is not obvious by looking at the data to tell if it is multivariate, normally distributed and more analysis of the data is necessary before it is know if Bayes is an optimal classifier for this data set.

6.4.2 More Context Information

While the word+context system takes advantage of some contextual information (the sequence of diary entries) present in the data, there is more information available in the data that is not utilized for this study. The word+context system does not incorporate information such as: time between diary entries, the time of day, or the day of the week. The time that has elapsed between diary entries is relevant to the usefulness of the context information as it is time sensitive. If the previous diary entry is four hours previous to the current one, the

context information should be less pertinent than context information from five minutes before the current dairy entry. With the breadth of context information available in the original study data, a number of methods could be created for utilizing it and improving diary entry classification accuracy. Future refinements could utilize this additional information to provide a higher degree of disambiguation and greater classification accuracy.

6.4.3 Data Set Size

Possibly the largest deficiency of this study was the small size of the total data set. To generate statistically significant probabilities for the variety of diary entries generated in a study of this type requires a training set of large size. A small data set means that either the diary entries will be too diverse and all CHAD codes would be essentially equally likely, or that there are too few codes for the word+context system to analyze enough diverse data. In a system reliant heavily on the probabilities it gleans from training sets, larger training sets with diverse data should yield high CHAD code classification accuracy.

It is unclear what is a sufficient database size needed for a study similar to this one. The database size in this study was only 1220 diary entries. Further experiments need to be conducted to determine the gain in precision and recall as the size of the training set increases. To accurately test the systems from this study, (the word+context system, the threshold system, and semantic ontology system) a larger corpus of diary entries is required. The data from this study was diverse and it was representative of common activities and locations subjects would typically be associated with, but forty two days worth of diary data is probably not sufficient to build reliable context data.

6.4.4 Reliance on the Word-only System

Another limitation for this system may be its reliance on the word-only system to generate lists of possible CHAD codes and their scores to augment. The word+context system builds off of the work done in the word-only system. The word+context system uses the word-only system as a first pass when it creates a list of initial possible CHAD codes. The context information is used only to augment the word-only score, not to choose possible CHAD codes for testing. The word-only system was improved upon with context information yet context information was not utilized in build lists of possible CHAD codes for each diary entry.

6.5 Future Work

6.5.1 Optimal Classifier

A naïve Bayes classifier may not have been optimal for the study. Exploring alternative classification algorithms is a natural extension to this line of research. An example classifier to contrast with the naïve Bayes classifier is a feed-forward neural network trained with the back-propagation algorithm. If these classifiers were to converge, it would show that the data exhibits characteristics of being multivariate, normally distributed.

Another common methodology for predicting sequences of actions uses Markov models. Further research could involve contrasting Markov modeling against the results of this study

6.5.2 Utilize More Contextual Data

Experimenting with a limited amount of context information improved diary entry classification over a word-only model significantly. Data such as time, date, time of day, and day of the week could improve diary entry classification accuracy even more. Future studies could analyze the database and extract more context data to incorporate into the word+context system.

6.5.3 Larger Corpus

Future experiments could benefit greatly from a larger data set from which to create testing and training sets. Though this EPA study was limited and the data set size was small, either similar data could be found from other studies of this type, or example utterances, representative of real diary entries, could be generated based on the current database. A larger corpus would provide more diary entries from which the system can build probabilities. With a larger, more diverse corpus, the system could be utilized and tested under the same parameters to ascertain the impact the larger data set had.

REFERENCES

Akland, G.G., Harwell, T.D., Johnson, T.R., and Whitmore, R.W. "Measuring human exposure to carbon monoxide in Washington, D.C., and Denver, Colorado during the winter of 1982-1983." *Environ. Sci. Technol*. 19:911-918. 1985.

Brauer, M., Hirtle, R.D., Hall, A.C., and Yip, T.R. Monitoring personal fine particle exposure with a particle counter. *J. Expos. Anal. Environ. Epidem.* 9:228-236. 1999.

Burton, R. R. *Semantic Grammar: a Technique for Efficient Language Understanding in Limited Domains*. Doctoral Thesis. UMI Order Number: AAI7629270. 1976.

Deerwester, S., Dumais, S., Landauer, T., Furnas, G., and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science, 41*(6):391--407, 1990.

Freeman, N.C.G., Waldman, J.M., Lioy, P.J. Design and evaluation of a location and activity log used for assessing personal exposure to air pollutants. *J Expos Anal Environ Epi,* 1: 327-338. 1991.

Grosz, B., Joshi, A., and Weinstein, S., Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics, 21*(2):203--225, 1995.

Guinn, C., Crist, D, and Werth, H., A Comparison of Hand-Crafted Semantic Grammars Versus Statistical Natural Language Parsing in Domain-Specific Voice Transcription, *Proceedings of Computational Intelligence*, Ed. B. Kovalerchuk, San Francisco, CA, pp. 490-495. 2006.

Guinn, C., and Rayburn Reeves, D, Using a Spoken Diary and Heart Rate Monitor in Modeling Human Exposure to Airborne Pollutnats for EPA's Consolidated Human Activity Database, *Third National Conference on Environmental Science and Technology*, to be published, 2008.

Hipp, D.R., *Design and development of spoken natural-language dialog parsing systems*. Ph.D. thesis, Duke University, available as Technical Report, CS-1993-15, 1992.

Jelinek, F., Self-organized language model for speech recognition, Readings in Speech Recognition, A. Waibel and K.F. Lee eds., pp. 450-506, Morgan-Kaufmann, San Mateo, CA, 1990.

Johnson, T. *A study of human activity in Cincinnati, Ohio.* Contractor report for Electric Power Research Institute, Palo Alto, CA, 1987.

Johnson, T., Long, T., and Ollison, W. *A survey of prospective technologies for collecting human activity data.* Contractor report for the American Chemistry Council. 2001.

D. Jurafsky and J. Martin. *Speech and Language Processing*. Prentice-Hall, 2000.

Klepeis, N.E., Nelson, W.C., Tsang, A.M., and Robinson, J.P. The national human activity pattern survey (nhaps): data collection methodology and selected results.) *J. Expos. Anal. Environ. Epidem.* 1998.

Lioy, P.J. Assessing total human exposure to contaminants. *Environ. Sci. Technol*. 1990. 24:938-945.

Mann, W. and Thompson, S. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text 8*(3): 243-281, 1988.

Mayrhofer, R., Radi, H. and Ferscha, A., Recognizing and Predicting Context by Learning from User Behavior, *The International Conference On Advances in Mobile Multimedia (MoMM2003)*, pp. 25-35, 2003.

McCurdy, T., Glen, G., Smith,L., and Lakkadi,Y. The National Exposure Research Laboratory's Consolidated Human Activity Database. *J. Exp. Anal. Environ. Epidem*. 10:566-578, 2000.

National Academy of Sciences (NAS). Human exposure assessment for airborne pollutants; advances and opportunities. National Academy of Sciences, Washington, DC, 1991.

Oliver and Horvitz, A Comparison of HMMs and Dynamic Bayesian Networks for Recognizing Office Activities, *User Modelling*, 2005.

Ott, W.R. "Concepts of human exposure to air pollution." Environ. Intl. 7:179-196. 1982.

Robinson, J.P., Wiley, J.A., Piazza, T., Garrett, K., and Cirksena, K. *Activity Patterns of California Residents and Their Implications for Potential Exposure to Pollution*. Sacramento, CA: California Air Resources Board (CARB-A6-177-33), 1989.

RTI, Longitudinal activity data for selected susceptible population subgroups. Final Report for EPA Contract No. 68-D-99-012 (Task Order 0012). 2001.

Ruiz, M. and Srinivasan, P., Hierarchical Text Categorization Using Neural Networks, *Information Retrieval 5*(1): 87-118, 2004.

U.S. Environmental Protection Agency (USEPA). Guidelines for exposure assessment. *Federal Register*, 57(104):22888-22938, 1992.

Zartarian, V.G., Ott, W.R., and Duan, N. A quantitative definition of exposure and related concepts. *J. Expos. Anal. Environ. Epidem*. 7:411-437, 1997.

Zelon, H.S. Capture of activity pattern data during environmental monitoring, in: T.H. Starks (ed.) *Proceedings of the Research Planning Conference on Human Activity Patterns.* Las Vegas NV: U.S. Environmental Protection Agency, pp. 8-1 to 8-12, 1989.