Using Statistical Techniques and WordNet to Reason with Noisy Data

Rakesh Gupta and **Mykel J. Kochenderfer***
Honda Research Institute USA, Inc.
800 California Street, Suite 300
Mountain View, CA 94041
rgupta@hra.com, m.kochenderfer@ed.ac.uk

Abstract

We collected data from non-experts over the web to create a common sense knowledge base for indoor home and office environments. In this paper, we discuss how we use statistical data dimension reduction and clustering techniques to determine consensus in the knowledge base. We explain the use of Latent Semantic Indexing in finding consensus. These statistical techniques make our system robust to noisy data in the knowledge base. Our work contrasts with traditional AI systems which are typically brittle as well as difficult to extend due to handcrafted pieces of knowledge in their knowledge bases. We then discuss how the WordNet hypernym hierarchy is used to generalize knowledge and perform inference about objects not in the knowledge base. WordNet also makes the reasoning system robust to vocabulary differences among people by using synonyms.

Introduction

The dominant view of intelligent reasoning, derived from mathematical logic, approaches intelligent reasoning as a form of calculation, typically deduction. These symbolic systems are domain dependent. As knowledge increases it becomes increasingly difficult to keep the database consistent with prior handcrafted knowledge. Systems like Cyc (Guha *et al.* 1990) have resorted to partitioning the database in consistent knowledge chunks based on context (called microtheories). However, such systems are handcrafted by a relatively small number of experts, and construction of the system requires examination of all the existing items in the relevant piece of the database for consistency. This consistency check requires exponential time based on the number of items in the database.

Statistical techniques have been used extensively in text retrieval and data mining to work with large collections of text documents (Baeza-Yates & Ribeiro-Neto 1999). We use these statistical techniques at the sentence and phrase level to address the issue of making these noisy databases consistent.

In our work, statistical data clustering techniques such as Latent Semantic Indexing (LSI) are used to remove inconsistencies in the knowledge base to find consensus data. The derived database can then be used to perform practical reasoning using Belief-Desire-Intention (BDI) architectures (Rao & Georgeff 1995; Wooldridge 1999).

It is important to emphasize that our system is not attempting to find the *right* answer but the one that reflects the majority consensus opinion. Without performing any consistency checking, we handle inconsistencies in two ways. Firstly, we do not allow users to explicitly enter negations (e.g. "a toilet is not found in the kitchen") into the knowledge base. Thus, we do not have assertions of a proposition and its negation in our knowledge base. Secondly, we focus on resolving high level inconsistencies in the knowledge base by consensus knowledge. For example, 15 people say that refrigerator is used to store food and one person says that it is used to provide filtered water. The later is not the primary use of the refrigerator, and is discarded by consensus even though it is valid knowledge.

One of the strengths of our approach is in the restriction of the domain (to indoor home and office environments), which makes our knowledge base dense enough to be statistically usable for inferencing. In the past, the available data has not been dense enough to leverage statistical techniques. Knowledge bases including the *Open Mind Common Sense* project at the MIT Media Lab (Liu, Lieberman, & Selker 2003; Eagle, Singh, & Pentland 2003), ThoughtTreasure (Mueller 1998) and Cyc (Guha *et al.* 1990) have attempted to capture much broader but sparse human common sense knowledge.

Objective

The objective of our work is to make indoor mobile robots that work in environments like homes and offices more intelligent through common sense. We want to endow these robots with some amount of general knowledge to use as a basis to accomplish their tasks and interact with people and their environment.

Mobile robots in homes and offices will be expected to accomplish tasks within their environment to satisfy the perceived desires and requests of their users using common sense. Common sense does not require expert knowledge, and hence it may be gathered from non-specialist netizens

*Currently at Institute of Perception, Action and Behaviour, School of Informatics, University of Edinburgh, Edinburgh, United Kingdom.

in the same fashion as the projects associated with the rather successful *Open Mind Initiative* pioneered by David Stork (Stork 1999; 2000). The raw data collected over the web is processed using statistical techniques and WordNet to populate an initial robot knowledge base. In the future, these robots will be able to learn through human interaction and other methods and add to their knowledge bases online.

In the next section, we describe our framework for capturing template data using Open Mind and our database schema. We then explain how we use Latent Semantic Indexing to prune our database to data reflecting consensus. We then discuss how we use WordNet to generalize our knowledge and to further make it independent of vocabulary. Finally, we discuss our conclusions and future work.

Knowledge Capture Framework

In this section, we describe the common sense knowledge base, the various implications used, and the database schema. We also provide examples of template-based forms for collecting common sense knowledge.

To capture relations, it would be easiest to have the users enter statements of the form $(object, property)$, but this would be unnatural for the user. The other extreme would be to accept phrases in natural language and later parse them using lemmatization and part-of-speech tagging. We decided to take an intermediate approach using sentence templates where the user is prompted with a natural language sentence with some blanks. Users enter words or phrases to complete these sentences (as in many of the other *Open Mind Initiative* projects).

Knowledge Base Implications

The framework of this work is *object-centric*. It is assumed that the desires of the users and the actions taken by the robot are grounded in the properties of objects in the world.

In our system, a statement is a pair $\phi = (o, p)$ where o is the object and p is a verb or adjective. For example, a user may wish that her coffee be heated. In this case, the object of the desire is *coffee* and the desired property is *heated*. The representation of this statement would be $(coffee, heated)$. Statements can also be thought of as actions. For example, $(coffee, heated)$ represents the action that causes coffee to be heated. The robot is capable of explicitly executing some set of actions represented in our system as statements.

One statement implying (or causing) another is represented as: $\phi_1 \rightarrow \phi_2$. For example, we might have the implication $(trash, in\_trash\_can) \rightarrow (house, clean)$. We shall denote the collection of all implications of this form by F.

One statement can indicate a desire represented by the second statement. Symbolically, this implication is represented as: $\phi_1 \rightarrow_d \phi_2$. For example, we might have the implication $(stomach, growling) \rightarrow_d (human, fed)$. We shall denote the collection of all such implications used to anticipate human desires by I.

This knowledge base can be used for common sense and practical reasoning using Belief-Desires-Intentions (BDI) theory. BDI was originally developed by Bratman (1987) and focuses on the role that intentions play in practical reasoning. It is founded upon established theory of rational action in humans.

Beliefs correspond to information the agent has about the world. Given F, sensor observations, and a belief revision function, current beliefs can be determined. *Desires* represent states of affairs that the agent would (in an ideal world) wish to be brought about. Current beliefs and I provide a list of current desires. In addition, humans may also explicitly assign goals (desires) to the robot. *Intentions* represent desires that it has committed to achieving. Desires can be filtered by a deliberation function to give current intentions that are then acted upon by the robot.

Relational Common Sense Knowledge

We first collected the names of objects commonly found in homes and offices. We started with a collection of photographs of such objects, and users identified them by entering their corresponding names into a form. Users were also prompted to enter new object names by forms like *An object found in the office is* _____.

Once some objects were collected, we asked users to identify various properties of an object. For example, the user might be prompted *A microwave is often* _____. Or, the user might be prompted *People sometimes desire that a cup of coffee is* _____. From this data, the system built a collection of object attributes.

From these statements, the system constructed questions about causality. For example, a form asked *A plant is healthy when a* _____ *is* _____. The response of the user was then transformed into an implication and added to the set F and stored in the system's database.

We gathered information about indicators of desires. For example, a form asked, *You might want a dvd player to be plugged into a television if you notice that your* _____ *has become* _____. The entries for the blanks were stored as an implication in the set I.

Common sense knowledge about the location of common objects is required to perform actions on objects in the home or office. For example, if we want the robot to make a cup of coffee, it would be useful for the robot to know that coffee makers tend to be found in kitchen and that milk is usually stored in the refrigerator. We therefore had activities where netizens could indicate typical object location and pairs of objects that are generally found together.

Other common sense knowledge we collected included knowledge about uses of different objects and activities of people. All this data was collected over the web from non-experts through forms hosted on the Open Mind website. More details about Open Mind Indoor Common Sense data collection is described in Gupta & Kochenderfer (2004).

Database Schema

Open Mind data collection is schematically shown in figure 1. All knowledge entered into the website goes through a review process by an administrator. Once accepted, these structured responses are transformed into relations and saved in the database. These relations are then used to generate new sentence templates.

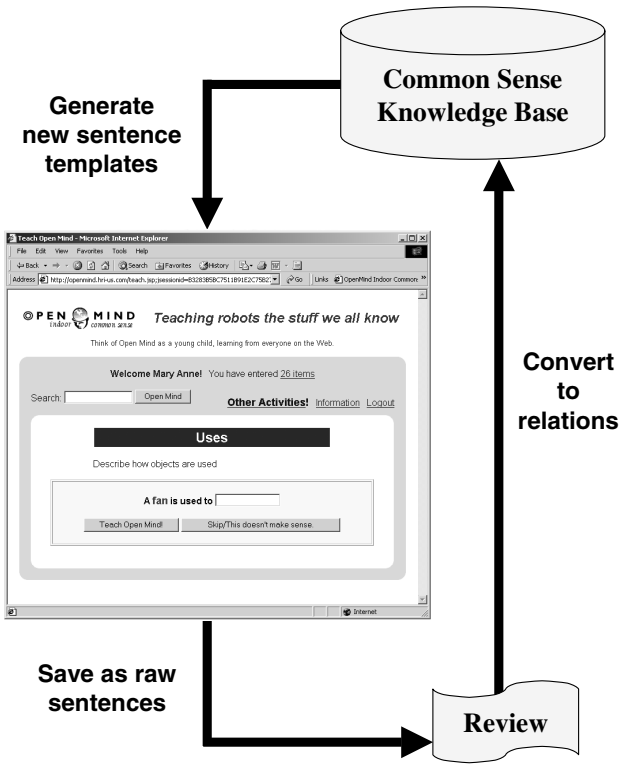

Figure 1: Schematic for Open Mind data collection.

A *MySQL* database is used to store the data collected from the website. Raw sentences are archived in a table with the following schema.

```
Entries(id, userid, date, sentence,
status, activity, form)
```

When a user enters new knowledge, its status is uncommitted. An administrator is later able to commit (or reject) the entry to the database. Committing an entry transforms the sentence and populates one or more of the following relations:

```
Objects(id, name)
Statements(id, obj, prop, desire)
Uses(id, obj, vp)
Causes(id, obj1, prop1, obj2, prop2)
Desires(id, obj1, prop1, obj2, prop2)
Rooms(id, name)
Proximity(id, obj1, obj2)
Locations(id, obj, room)
```

We use statistical techniques like LSI to clean up that raw knowledge base. Once data is cleaned up and compacted using WordNet, these relations are used to answer user posed text queries about object location and their use in various tasks.

Latent Semantic Indexing for finding consensus

Luhn (1961) proposed the idea of notional families to group together words of similar and related meaning. We can view Latent Semantic Indexing (LSI) as a similarity measure that makes judgements based on word co-occurrence. Such co-occurrence has been used as an indicator of topic relatedness in document analysis (Deerwester *et al.* 1990; Scott & Matwin 1999), but in our work, we use it to analyze phrases. In latent semantic space, phrases that are semantically similar according to the co-occurrence analysis are clustered together. We select the primary cluster as the one reflecting consensus.

To prepare the data for LSI, we first run a spell-check on the web-collected data and correct spellings. We then analyze the sentences to remove frequent words using a stop-word list to reduce the feature set size, and then we perform stemming. Stemming maps word variants to the same feature (e.g. cooking, cooked, cooks maps to same word cook). We then apply LSI via Singular Value Decomposition (SVD) to find the most relevant sentence cluster. The computation of SVD is quadratic in the rank of the term by sentence matrix and cubic in the number of singular values that are computed (Manning & Schuetze 2001).

For example, in our database following is the raw data collected for the template *A cup is used to* _____.

get the right amount of an ingredient when cooking
drink tea
hold tea
drink out of
hold a drink
serve tea
hold coffee
drink coffe from
drink coffee
drink from
drink
drink and hold tea
holding tea
drink from
hold coffee
store items
wake up in the morning
store dishes
keep awake

The entry *coffe* is replaced by the word *coffee* in the spell-check phase. The stop words *of*, *a*, *from*, *when*, *up*, *in*, *the* and *an* get filtered out. Stemming maps *hold* and *holding* to the same root word *hold*. Thus, after preprocessing steps we get the following terms:

1. hold
2. tea
3. drink
4. out
5. serve
6. coffee
7. get
8. right
9. amount
10. ingredient
11. cook
12. store

0 0 0 0 0 0 1 1 1 1 1 0 0 0 0 0 0 0
0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0
1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0
0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0
0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0
0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1

Figure 2: 19 phrase $\times$ 18 term matrix for SVD. Each row corresponds to a particular phrase (e.g. "get the right amount of an ingredient when cooking"), and the 1's and 0's correspond to the presence or absence of a particular term (e.g. "hold").

13. item
14. wake
15. morning
16. dish
17. keep
18. awake

This leads to the 19 phrase $\times$ 18 term matrix in figure 2.

The first phrase indicates a use for a cup, and the last three are nonsense options. SVD decomposition clusters phrases from 2–16 in the clusters represented by the first two singular values. The first phrase shows up in the cluster with the third highest singular value and the remaining nonsense phrases in the clusters with the remaining smaller singular values. Thus LSI helps us select one of the relevant phrases from 2–16 range. A phrase from this range can be selected randomly.

It is important to point out that even though we used LSI for determining consensus, any of the other techniques for data dimension reduction and clustering such as Spectral Clustering and Principal Component Analysis (PCA) could be similarly applied.

While the LSI method deals nicely with the synonymy problem by clustering them together, it offers only a partial solution to the polysemy problem. That is, a word with more than one entirely different meaning (e.g. bank) is represented as a weighted average of the different meanings. If none of the real meanings is like the average meaning this may cause a severe distortion. In our case polysemy is not a serious issue because the restriction of the domain heavily cuts down the number of terms with multiple relevant meanings in the domain.

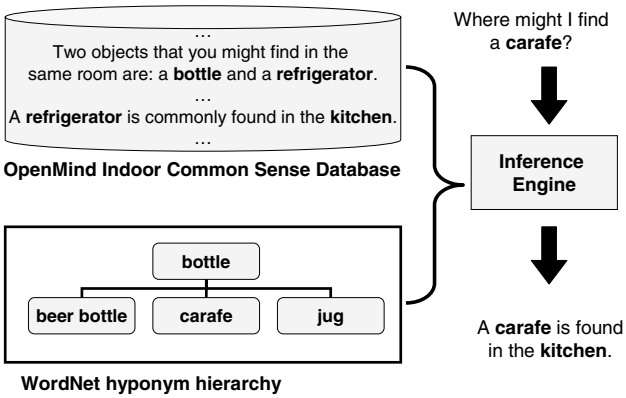

Figure 4: Inference using Open Mind and WordNet.

Using WordNet for knowledge compaction and vocabulary independence

The WordNet lexical database is used to compact the knowledge base to allow the making of inferences about objects that do not themselves exist in the knowledge base but their hypernyms or synonyms do.

To illustrate knowledge compaction, suppose our database contains the fact that spoons are usually found in the kitchen. Hypernymy is a linguistic term for the *is-a* relationship, e.g. since a spoon is a cutlery, cutlery is a hypernym of spoon as shown in the WordNet hierarchy for the noun *tableware* in Figure 3. In this example, we would ask if all types of cutlery are usually found in the kitchen. If the consensus answer is yes, we would replace the knowledge about spoon location with the knowledge about cutlery. We use the disambiguated sense to determine the appropriate hypernym to query the user.

Knowledge compaction not only saves space but also makes our knowledge more general. For example, initially our knowledge base may have the knowledge that "forks are usually found in kitchen" and "spoons are usually found in the kitchen". With knowledge compaction, we remove knowledge about forks and spoons and replace it with cutlery knowledge. This saves storage space as well as allows us to infer that table knives are found in the kitchen. We could not have inferred anything about table knives before this knowledge compaction.

Since the relations are structured and interconnected, rule-based inference may be applied to produce sentences that were not explicitly entered into the database. For example, if the database contains the facts that a spoon is found near a bowl and a bowl is generally found in the dining room, the inference engine would be able to infer that a spoon may be found in the dining room with a certain degree of confidence. Our knowledge base has been used in conjunction with WordNet to infer the location of objects that were not found in the database. Users may issue commands using objects that do not explicitly exist in the knowledge base but are in the WordNet hypernym hierarchy of the object.

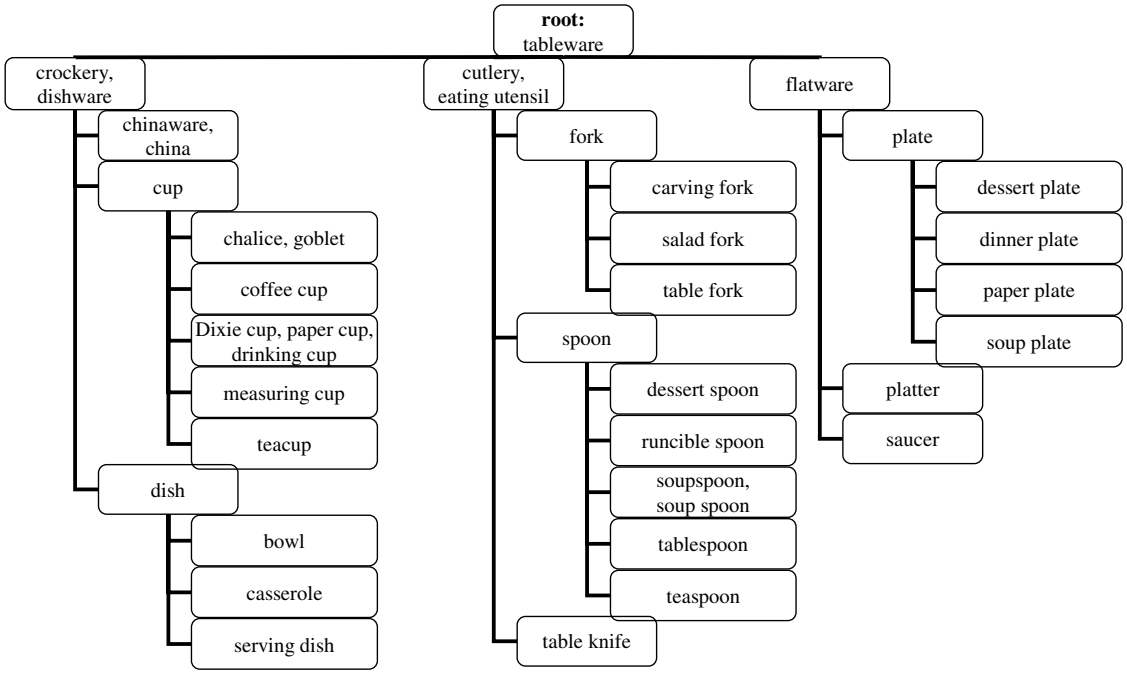

Figure 3: Example of a WordNet hierarchy for the noun *tableware*.

Figure 4 shows an example of inferencing where the query is: *Where might I find a carafe?* From the Open Mind Indoor Common Sense database, after LSI we have:

- Bottle and refrigerator are found together.
- A refrigerator is found in the kitchen.

We can use the WordNet hierarchy and knowledge entered by netizens to prune irrelevant senses of the word bottle in the context of indoor environments. From this pruned set of senses, we can use the information that bottle is a hypernym for carafe from WordNet. We can therefore infer that a carafe is found in the same place as a bottle, which is the kitchen.

As another example, the object *patchwork* was never entered into the knowledge base. However, our inference engine was able to observe that a patchwork is a hyponym of *quilt*. The knowledge base contained the fact that quilts are typically found in bedrooms and living rooms, and so the system was able to infer that patchworks are found in bedrooms and living rooms.

There is a distinction between hypernyms and hyponyms in inferencing with WordNet. If a queried word is not present in the database but its hypernym is, then we can proceed using the information we have in the database about the hypernym. However, if the queried word is not present in the database but its hyponym is, it is not necessarily sound to use the information contained in the database about the hyponym. For example: if we do not know anything about *pets* but we know that its hyponym *dogs* has the property of *has four legs*, we cannot simply deduce that *pets have four legs*—only that *pets can have four legs* (but they might have two legs) etc.

Besides knowledge compaction, it is important that our common sense system provide an amount of vocabulary independence since two people typically choose the same name for a single well-known object less than 20% of the time (Deerwester *et al.* 1990). Voorhees (1994) expanded queries manually to include synonyms of terms to improve performance in text retrieval tasks. An example of using synonyms is the query *Where can I find chalice*. The word chalice does not exist in our knowledge base but its synonym goblet does. This allows the inference that chalice can be found in the kitchen.

Conclusions and Further Work

We extracted common sense knowledge as relations from structured sentence inputs obtained over the web. Focusing on a restricted indoor home and office environment domain supplied us with dense common sense knowledge base that was processed using data clustering statistical techniques to clean up the noisy data as well as to determine consensus.

We further processed the knowledge base using the WordNet lexical database to generalize data by replacing implications associated with existing objects by implications associated with their hypernyms. WordNet was also used to provide vocabulary independence in queries.

This framework is general enough to be extended to other restricted environments like hospitals and airports. Outdoor common sense knowledge might be used to determine when it is safe to cross roads. In hospitals and airports, common

sense may be used to determine when to offer help to people in carrying objects or answer queries about locations of objects or places in the environment.

Further work includes refinements to LSI including phrase finding and methods to handle negation and disjunction in queries. Another improvement would be to incorporate low-frequency words that are informative but are filtered out. We are also interested in expanding this work to learn new knowledge online.

On the interface side, future work includes replacing the text interface with a speech recognition system. We will also explore human-robot dialogue so that the robot can ask questions if it is not able to understand new information. It will be an interesting challenge to extend the system to understand simple spatial descriptions such as, *Peter's room is to the right* or *the object on the table is the projector*.

This work can be extended to perform practical reasoning and action selection using the BDI architecture. Another possible path is the use of teleo-reactive programs (Nilsson 1994) in accomplishing tasks in dynamic environments.

Acknowledgments

This work was done while Mykel Kochenderfer was a summer intern at Honda Research Institute USA, Inc. Nils Nilsson reviewed an earlier version of this paper and provided invaluable feedback. Thanks are also due to anonymous reviewers and the users of the Open Mind Indoor Common Sense website for their data and feedback.

References

Baeza-Yates, R., and Ribeiro-Neto, B. 1999. *Modern Information Retrieval*. New York: Addison Wesley Longman.

Bratman, M. 1987. *Intention, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press.

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

Eagle, N.; Singh, P.; and Pentland, A. 2003. Common sense conversations: Understanding casual conversation using a common sense database. *Artificial Intelligence, Information Access, and Mobile Computing Workshop at the 18th International Joint Conference on Artificial Intelligence (IJCAI)*.

Guha, R. V.; Lenat, D. B.; Pittman, K.; Pratt, D.; and Shepherd, M. 1990. Cyc: A midterm report. *Communications of the ACM* 33(8):391–407.

Gupta, R., and Kochenderfer, M. J. 2004. Common sense data acquisition for indoor mobile robots. In *Nineteenth National Conference on Artificial Intelligence (AAAI-04)*.

Liu, H.; Lieberman, H.; and Selker, T. 2003. A model of textual affect sensing using real-world knowledge. In *Proceedings of the Seventh International Conference on Intelligent User Interfaces (IUI 2003)*, 125–132.

Luhn, H. P. 1961. The automatic derivation of information retrieval encodement from machine readable text. *Information Retrieval and Machine Translation* 3(2):1021–1028.

Manning, C. D., and Schuetze, H. 2001. *Foundations of Statistical natural language Processing*. Cambridge, Massachusetts: MIT Press. chapter Topics in Information Retrieval: Latent Semantic Indexing.

Mueller, E. T. 1998. *Natural language processing with ThoughtTreasure*. New York: Signiform.

Nilsson, N. J. 1994. Teleo-reactive programs for agent control. *Journal of Artificial Intelligence Research* 1:139–158.

Rao, A. S., and Georgeff, M. P. 1995. BDI-agents: from theory to practice. In *Proceedings of the First Intl. Conference on Multiagent Systems*.

Scott, S., and Matwin, S. 1999. Feature engineering for text classification. In Bratko, I., and Dzeroski, S., eds., *Proceedings of ICML-99, 16th International Conference on Machine Learning*, 379–388. Bled, Slovenia: Morgan Kaufmann Publishers, San Francisco, US.

Stork, D. G. 1999. The Open Mind Initiative. *IEEE Expert Systems and Their Applications* 14(3):19–20.

Stork, D. G. 2000. Open data collection for training intelligent software in the open mind initiative. In *Proceedings of the Engineering Intelligent Systems (EIS2000)*.

Voorhees, E. M. 1994. Query expansion using lexical-semantic relations. In *Proceedings of SIGIR 94*, 61–69.

Wooldridge, M. 1999. Intelligent agents. In Weiss, G., ed., *Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence*. Cambridge, MA, USA: The MIT Press. 27–78.