# Context Sensing using Speech and Common Sense

**Nathan Eagle**
20 Ames St., Cambridge, MA 02139
nathan@media.mit.edu

**Push Singh**
20 Ames St., Cambridge, MA 02139
push@media.mit.edu

## Abstract

We present a method of inferring aspects of a person's context by capturing conversation topics and using prior knowledge of human behavior. This paper claims that topic-spotting performance can be improved by using a large database of common sense knowledge. We describe two systems we built to infer context from noisy transcriptions of spoken conversations using common sense, and detail some preliminary results. The GISTER system uses OMCSNet, a commonsense semantic network, to infer the most likely topics under discussion in a conversation stream. The OVERHEAR system is built on top of GISTER, and distinguishes between aspects of the conversation that refer to past, present, and future events by using LifeNet, a probabilistic graphical model of human behavior, to help infer the events that occurred in each of those three time periods. We conclude by discussing some of the future directions we may take this work.

## 1 Introduction

Can we build computers that infer a speaker's context by summarizing the conversation's gist? Once computers are able to capture the gist of a conversation, an enormous number of potential applications become possible. However, current topic-spotting methods have met with little success in characterizing spontaneous conversations involving hundreds of potential topics (Jebara *et al.*, 2000). This paper claims that performance can be greatly improved by making use of not only the text of a speech transcription, but also perceptual and commonsensical information from the dialogue.

We have enabled a suite of wearable computers with the ability to provide the perceptual information necessary for a human to infer a conversation's gist and predict subsequent events. To take the human fully out of the loop we have infused the system with two commonsense knowledge bases that enable the computer to make educated inferences about the user's context.

## 2 Implementation

Our system incorporated a Zaurus Linux handheld computer, with an 802.11b CF card and a wireless Bluetooth headset microphone. Applications were written to enable the Zaurus to stream high quality audio (22 kHz, 16-bit) to an available 802.11b network, or to store the audio locally when no network is detected. Besides streaming audio, packets in this wireless network could be 'sniffed' by the PDAs interested in determining who else is in the local proximity. Information regarding access point signal strength information was correlated with location using a static table look-up procedure. The system is typically kept in a participant's pocket, or for those with Bluetooth headsets, stored in a briefcase, purse, or backpack.

### 2.1 Audio Processing and Transcription

ViaVoice, a commercial speech recognition engine, is used to transcribe the audio streams, however typically word recognition rates fall below 35% for spontaneous speech recognition (Eagle & Pentland, 2002). This inaccuracy poses a serious problem for determining the gist of an interaction. However, a human can read through a noisy transcript and with adequate perceptual cues, still have an impression of the underlying conversation topic.

> ***Speaker 1****: you do as good each key in and tell on that this this printers' rarely broken key fixed on and off-fixes and the new nine-month London deal on and then now take paper out and keep looking cartridges and then see if we confine something of saw someone to fix it but see Saddam out of the system think even do about it had tools on is there a persona for the minister what will come paper response to use the paper is not really going to stay in the printer for very much longer high is Chinese college and shredded where inks that inks is really know where the*

> *sounds like a Swiss have to have played by ear than*
> ***Speaker 2**: a can what can do that now I think this this seems to work on which side is working are in*
> ***Speaker 1**: an hour riderless I E fix the current trend the Stratton practice page of the test casings to of printed nicely I think jacking years ago that is paid toes like a printed Neisse*

Additional context, such as information that the conversation is occurring in an office, or more precisely, by a printer, may help many people understand that the conversation is about fixing a printer jam. Prior knowledge about the conversation participants and the time of day may also significantly augment a person's ability to infer the gist of the interaction, for example, one of the speakers could be a printer repairman. Our work suggests that the additional contextual and commonsensical information a human can employ for inference on the transcript above is equally helpful to a probabilistic model.

As will be shown, this additional contextual and commonsense information can be used to form probabilistic models relating observed keywords to conversation topic. Thus by combining audio and information from a mobile device with a commonsense knowledge network, we can determine the gist of noisy, face-to-face conversations. In the above example, for instance, our system correctly labeled the conversation as 'printing on printer'.

## 3 GISTER

GISTER is a system that infers the most likely topics under discussion in a conversation stream by using a commonsense semantic network called OMCSNet. More details about the GISTER system are available in (Eagle *et al.*, 2003) but we summarize its operation in this section.

### 3.1 OMCSNet

We built the OMCSNet commonsense semantic network (Liu & Singh, 2004) by aggregating and normalizing the contributions from nearly 14,000 people from across the web (Singh *et al.*, 2002). Its semantic network structure resembles that of WordNet (Fellbaum, 1998) but its content is motivated by the range of knowledge in commonsense knowledge bases such as Cyc (Lenat, 1995). As in WordNet, the nodes of OMCSNet are natural language terms and the links are drawn from a fixed ontology of semantic relationships. But as in Cyc, the nodes include not just single words, but also compound expressions such as 'at the zoo', 'eat a sandwich' or 'fix a printer', and the links are drawn from a broader range of semantic relationships than are available in WordNet; OMCSNet goes beyond simple 'is-a' and 'part-of' relations to include spatial, temporal, causal, affect, and other types of relations. At present OMCSNet employs the 20 binary semantic relations shown below in Table 1.

| Relation Type | Semantic Relation |
|---|---|
| Things | KindOf, HasProperty, PartOf, MadeOf |
| Events | SubEventOf, FirstSubeventOf, LastSubeventOf, HappensAfter |
| Actions | Requires, HasEffect, ResultsIn-Want, HasAbility |
| Spatial | OftenNear, LocationOf |
| Goals | DoesWant, DoesNotWant, MotivatedBy |
| Functions | UsedInLocation, HasFunction |
| Generic | ConceptuallyRelatedTo |

**Table 1.** Semantic relations currently in OMCSNet

Prior research in text summarization has recognized the need for general world knowledge—in SUMMARIST (1997), Hovy & Lin describe how the words "gun", "mask", "money", "caught", and "stole" together would indicate the topic of "robbery", but they note that that WordNet and other dictionary-like resources did not contain enough such knowledge. However, OMCSNet contains precisely this type of knowledge. It contains a wide variety of knowledge about many aspects of everyday life: typical objects and their properties, the effects of ordinary actions, the kinds of things people like and dislike, the structure of typical activities and events, and many other things. A small excerpt of OMCSNet is show in Figure 1 below.

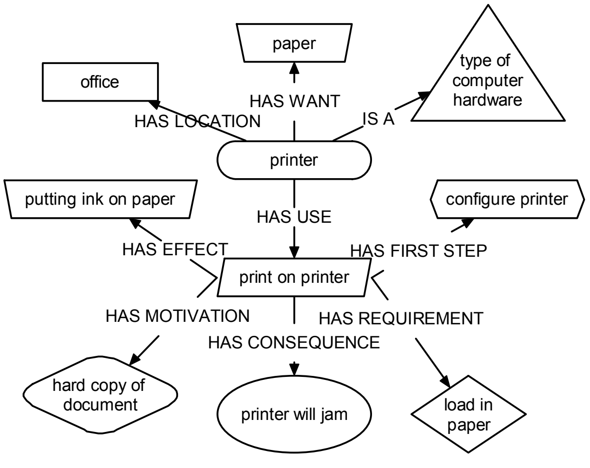

**Figure 1.** A selection of OMCSNet's 250,000 relations

OMCSNet has been used in a variety of applications to date (Lieberman *et al.*, 2004).

### 3.2 GISTER infers fine-grained topics

The purpose of the GISTER system is to infer the 'fine grained topic', or gist, of a conversation. A gist is the class of event that most accurately summarizes the current subject of the conversation. For example:

- Buying a ticket to a baseball game
- Looking for a restaurant
- Scheduling a meeting
- Canceling a meeting

These gists are represented within OMCSNet as the nodes of the semantic network containing simple verb phrases. For our set of target gists, we use the 700 most richly defined verb phrase nodes within OMCSNet (those for which at least 10 facts are asserted.)

GISTER infers gists using a two step process. First, the transcriptions are preprocessed to reduce the noise of the speech recognition engine. To do this the transcriptions are lemmatized and filtered for stop words (such as 'like', 'the', 'a', etc.), and a filtering process is performed using a clustering metric to reduce the number of weakly connected words. These outliers, words with very sparse links to the rest of the transcription, are removed from the data set.

Second, the OMCSNet network is flattened into a bipartite network that incorporates all ties from words in the OMCSNet lexicon to gists. The probability of a specific gist can be modeled as proportional to the gist's links to the selected words:

$$P(g_i|k) = \frac{k_i}{\sum_{i=1}^{G} k_i}$$

$$GistScore = k_i$$

where $k_i$ is the number of links between a gist, $g_i$ , and the observed transcript, and $G$ is the number of potential gists (approximately 700). This simple method is often capable of identifying a small group of potential gists, frequently with one dominating the others.

Once the probable topics of conversation have been identified and ranked, contextual information about the conversation is incorporated into the model. In many instances, information such as location or participant identity can identify the gist from the small subsection of topics. In our initial tests we incremented a gist's score for each of its links to a keyword related to the given context.

### 3.3 Experiments

We ran a series of experiments on a testing set of 20 speech segments ranging from 50 to 150 words and taken from a single individual on a wide range of topics. No prior knowledge about the participant was assumed, but the 802.11b networks were used to give general locations such as office and cafeteria when appropriate. In one test we captured conversations from the student center cafeteria – streaming data to an access point mapped as 'restaurant'. Using this contextual information to condition the model, our results significantly improved:

Transcription:

*Store going to stop and listen to type of its cellular and fries he backed a bill in the one everyone get a guess but that some of the past like a salad bar and some offense militias cambers the site fast food them and the styrofoam large chicken nuggets son is a pretty pleased even guess I as long as can't you don't have to wait too long its complicity sunrise against NAFTA pact if for lunch*

Selected Keywords:

*wait type store stop salad past lunch long long listen large fry food fast chicken cellular bill big bar back*

Top Ten Scores:

| Without Location Context | | With Location Context | |
|---|---|---|---|
| 5 | talk with someone far away | 27 | eat in fast food restaurant |
| 5 | buy beer | 21 | eat in restaurant |
| 5 | Eat in restaurant | 18 | wait on table |
| 5 | eat in fast food restaurant | 16 | you would go to restaurant because you |
| 5 | buy hamburger | 16 | wait table |
| 4 | go to hairdresser | 16 | go to restaurant |
| 4 | wait in line | 15 | know how much you owe restaurant |
| 4 | howl with laughter | 12 | store food for people to purchase |
| 4 | eat healthily | 11 | sitting down while place order at bar |
| 4 | play harp | 11 | cook food |

**Table 2.** Results of using Context for Gist Differentiation

Actual Situation:

*Deciding what to get for lunch while standing in line at the cafeteria.*

## 4 OVERHEAR

The OVERHEAR system is a newer system, built on top of GISTER, and distinguishes between aspects of the conversation that refer to past, present, and future events. The system relies on LifeNet, a probabilistic graphical model of human behavior, to infer the events occurring in each of those three time periods.

We have two reasons for trying to distinguish between past, present, and future events. First, using additional sensory context (such as addition information about the speakers' location) to bias the results of gist sensing only works when the conversation is referring to the present context. Often, people's conversations referred to things that happened in the past, or things they were planning to do in the future, and in those cases sensory context only hurt GISTER's performance. However, one could imagine making use of recorded, time-stamped sensory data to bias the gisting of conversations that were talking about events that happened earlier.

Second, our long term goal is to use context sensing from speech to build new types of context-aware applications for wearable computers and other mobile devices. An application that knew that the speaker was referring to past events could perform tasks like retrieve documents and e-mails that referred to those past events. However, if the speaker was referring to the current situation, the application could know to make use of sensory information to improve its understanding of the current context. And if the speaker was referring to potential future events, like 'going to a movie this weekend', the application could assist the user by making plans to help make those events happen (or not happen, as the case may be), for instance by offering to purchase movie tickets on-line.

### 4.1 LifeNet

LifeNet is a probabilistic graphical model that captures a first-person model of human experience. It relates 80,000 'egocentric' propositions with 415,000 temporal and atemporal links between these propositions, as shown in Figure 2. More details about how the LifeNet model is generated are given in (Singh & Williams, 2003).

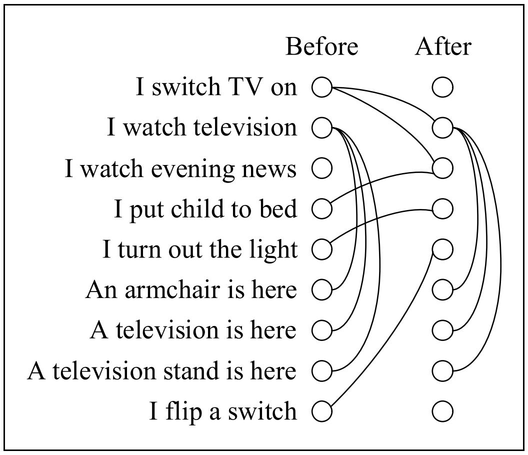

**Figure 2.** A sample of LifeNet

The structure of LifeNet is represented by a Markov network whose structure resembles a Dynamic Bayesian Network (DBN). Although lacking the 'explaining away' power of true Bayesian inference, the model is not constrained to directed acyclic graphs. LifeNet is designed to support the same kinds of temporal inferences as a DBN, including predicting future states and guessing prior states from the current state.

LifeNet was built as a probabilistic graphical model because stochastic methods can be more tolerant than traditional logical reasoning methods to the uncertainty in our knowledge of the situation, as well as to the uncertainty in the reliability of the rules themselves. Additionally these methods have efficient and well-known inference procedures for generating approximate solutions.

Our early experiments reasoning with LifeNet treat it as Markov network, an undirected graphical model where the nodes represent random variables and the edges joint probability constraints relating those variables. We convert LifeNet into a series of joint probabilities (the details of this process are described later this paper), and we reason with the resulting network using local belief updating techniques. We engage in 'loopy' belief propagation as described by Pearl (Pearl, 1988). Belief propagation in a Markov network is straightforward. We use the following belief updating rules, as described in (Yedidia *et al.*, 2000):

$$m_{ij}(x_i) \leftarrow \alpha \sum_{x_i} \psi_{ij}(x_i, x_j)\psi_i(x_i) \prod_{k \in N(i)\backslash j} m_{ki}(x_i) \quad (1)$$

$$b_i(x_i) \leftarrow \alpha\psi_i(x_i) \prod_{k \in N(i)} m_{ki}(x_i) \quad (2)$$

In these rules $x_i$ represents the random variable at node $i$. The current belief in node $i$ is denoted by $b_i$, the local evidence for node $i$ by $\psi_i$ and the joint probability of a pair of linked nodes $i$ and $j$ by $\psi_{ij}$. The message sent from node $i$ to $j$ is denoted by $m_{ij}$. $N(i)$ is the set of all neighbors of node $i$, and $N(i) \backslash j$ represents the set of all neighbors of node $i$ except for node $j$. $\alpha$ is a normalization constant.

These simple updating rules run fairly quickly even on a knowledge base as large as LifeNet. In our optimized Common Lisp implementation, on a 1.8 GHz Pentium 4 with 512 MB ram, a single iteration of belief updating runs in 15 seconds. Inference is further sped up by restricting the belief updating to run only on nodes within a fixed distance from the evidence nodes. Given a single evidence node and using only those nodes within a depth of three edges away, a single iteration of belief updating runs in as little as 0.5 seconds for some nodes; on average it takes about 3 seconds.

### 4.2 Model Integration and Implementation

GISTER leverages the commonsense facts within OMCSNet to generate discrete conceptual topics from a given transcript segmented into twenty-word-long observations, with each twenty-word observation independent from the others. We extended GISTER to infer the average tense of the text within the observation by detecting verb tenses, auxiliary verbs like *did* and *will*, and also specific temporal expressions like *yesterday* and *tomorrow*. LifeNet then allows us to calculate the transition probabilites to a given specific propositional representation based on previous states. By using the independent output of GISTER as input into LifeNet, we are able to improve the inferences of a user's context which subsequently can be used to training data for improved models of human behavior.

We propose a variation to the Markov network implementation of LifeNet described in section 4.1. Noisy transcript and signal data is still used as initial input into the system; GISTER then processes this data, semantically filters the speech, and calculates the likely subjects of conversation and their tenses. Highly ranked output from GISTER is then used as temporal observations for inference on the LifeNet model, as shown in Figure 3. These observations are linked to specific nodes within LifeNet that correspond to the given tense (past, present, future). We used multiple root nodes with weights proportional to the rank generated from the gister. This belief weighting system accounts for the uncertainty of the gister's output while starting with multiple roots enables much richer inference.

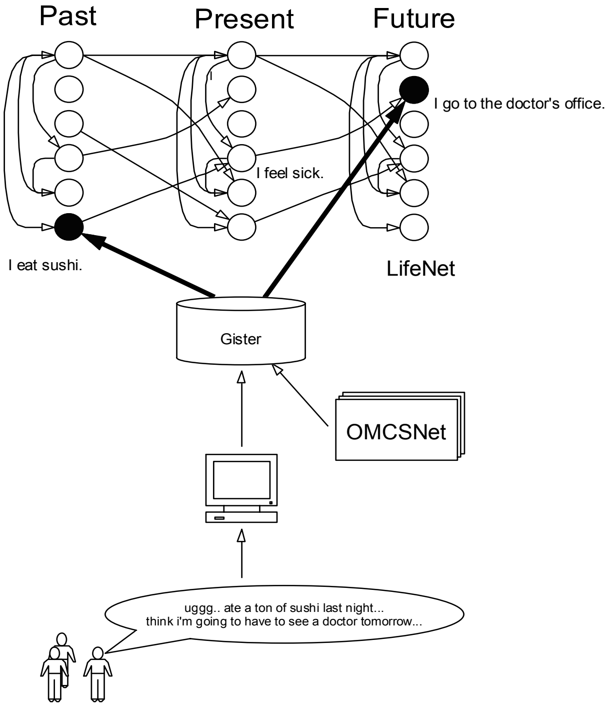
Past
Present
Future
I go to the doctor's office.
I feel sick.
I eat sushi.
LifeNet
Gister
OMCSNet
uggg.. ate a ton of sushi last night...
think i'm going to have to see a doctor tomorrow...

**Figure 3.** The OVERHEAR System

As shown in Figure 3, by using the output of GISTER for inference in LifeNet, additional insight can be gained about the user's situation. If the output from GISTER is 'eating sushi' and was assigned a past tense, while 'going to the doctor' was assigned a future tense, LifeNet can make educated inferences about what happened to the user. This inference can be fed back into the lower level of the model by weighting words like 'sick, full, tired', and rerunning the semantic filtering technique. By incorporating this feedback into the system, the filtering technique would be much less likely to exclude words related to being sick despite them being initally filtered from the transcript. If the gister's output changes, the process continues until the two systems converge on a solution.

### 4.3 Preliminary Results

The system was initially tested on an office worker's conversation regarding how she had eaten too much the day before and that she will have to go to the doctor's office during the next day. The following transcripts were input into GISTER:

> *PAST: had sushi for lunch could then have thought so he and then so yesterday's the sushi I its while I was at the Senate Committee lunch it tasted good sign yet was expenses over a cost me $7 to buy six roles and they lead to much of its in the rules were not a very good and I ate too many roles half so after words about six hours later I wasn't feeling very well of this so more Matsushita I never bought some sugar before usually advised chicken sandwich usual and normal food there I thought that this issue would be a good deal I also bought some seltzer water was so worked well and silence*

> *FUTURE: of debt reduction appointment tomorrow they can see mental tomorrow to clock will meet Dr. Smith and he's going to put my stomach because of what I a yesterday bomb I'm hoping that when I'll feel better so looking forward to going*

In this experience an overall tense was assigned to each passage. GISTER correctly inferred that the first passage referred to past events and the second to future events, and output potential topics of the conversation for each of those time periods:

| Past | Present | Future |
|---|---|---|
| eat lunch | | fall |
| eat | | have examination |
| have lunch | | eat cookie |
| get in shape | | go for run |
| get job | | have physical exam |
| get fit | | eat lunch |
| eat breakfast | | go on vacation |
| cook dinner | | give assistance |
| taste something sweet | | take walk |
| lose weight | | walk |

**Table 3.** Potential Topics Separated by Tense

The topics generated by GISTER in Table 3 were subsquently used as observational inputs to the next section of the model. These topics were mapped to the past and future nodes within LifeNet, marked as 'true', and then we ran the loopy belief propogation algorithm described earlier. The solution converged on nodes representing the present state, in-between the first tier (past) and the third tier (future). The nodes deemed most likely by the system are listed in Table 4 below.

| Inferences on Present Situation | |
|---|---|
| 0.999 | I stop being hungry |
| 0.999 | I warm feeling |
| 0.982 | I satisfy hunger |
| 0.964 | I make appointment |
| 0.962 | I have energy |
| 0.957 | I schedule appointment with doctor |
| 0.956 | I feel worry |

**Table 4.** Inferences about Present Situation given Past and Future

### 4.4 Training Future Models of Human Behavior

When this system is deployed on many users over an extended period of time, information about people's behavior can begin to influence the initial priors from LifeNet. Although it has not been determined how additional links could be made, this represents an alternative method for increasing the common sense knowledge stored within LifeNet. Additionally, extensive observations on the same people could augment the original commonsense model by better reflecting an individual's behavior.

## 5 Conclusions

Combining common sense with speech and other types of sensory context presents abundant opportunities within a wide range of fields, from artificial intelligence and ubiquitous computing to traditional social science. By integrating two common sense knowledge bases, we have developed a method for inferring human behavior from noisy transcripts and sensor data. As mobile phones and PDAs become ever more embedded in society, the additional contextual information they provide will become invaluable for a variety of applications. This paper has shown the potential for these devices to leverage this additional information to begin understanding informal face-to-face conversations and inferring a user's context.

## Acknowledgements

The authors would like to thank to Hugo Liu for his work developing OMCSNet and Sandy Pentland.

## References

Nathan Eagle and Alex Pentland. 2002. Information Explication from Computer-Transcribed Conversations. MIT Media Lab Vision and Modeling Technical Report.

Nathan Eagle, Push Singh, and Alex Pentland. 2003. Common sense conversations: understanding casual conversation using a common sense database. *Proceedings of the Artificial Intelligence, Information Access, and Mobile Computing Workshop (IJCAI 2003)*. Acapulco, Mexico.

Christiane Fellbaum (Ed.) 1998. *WordNet: An electronic lexical database*. MIT Press.

Eduard Hovy and Chin-Yew Lin. 1997. Automated text summarization in SUMMARIST. In *ACL/EACL-97 Workshop on Intelligent Scalable Text Summarization*. Madrid, Spain.

Douglas Lenat. 1995. CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11).

Henry Lieberman, Hugo Liu, Push Singh, and Barbara Barry. 2004. Beating Some Common Sense into Interactive Applications. In submission. Draft available at http://web.media.mit.edu/~lieber/Publications/Beating-Common-Sense.pdf.

Hugo Liu and Push Singh (2004). OMCSNet: A commonsense inference toolkit. MIT Media Lab Technical Report SOM03-01. Available at: http://web.media.mit.edu/~push/OMCSNet.pdf

Tony Jebara, Yuri Ivanov, Ali Rahimi, and Alex Pentland. 2000. Tracking Conversational Context for Machine Mediation of Human Discourse. *Proceedings* *of AAAI Fall Symposium on Socially Intelligent Agents*. North Falmouth, MA.

Judea Pearl. 1988. Probabilistic reasoning in intelligent systems: networks of plausible inference. San Mateo, CA: Morgan Kaufman.

Push Singh, Thomas Lin, Erik T. Mueller, Grace Lim, Travell Perkins and Wan Li Zhu. 2002. Open Mind Common Sense: Knowledge acquisition from the general public. *Proceedings of ODBASE'02*. Irvine, CA.

Push Singh and William Williams. 2003. LifeNet: a propositional model of ordinary human activity. *Proceedings of the Workshop on Distributed and Collaborative Knowledge Capture (DC-KCAP)* at K-CAP 2003. Sanibel Island, Florida.

Jonathan Yedidia, William Freeman, and Yair Weiss. 2000. Generalized belief propagation. *Advances in Neural Information Processing Systems (NIPS)*, 13, 689-695.