Accepted to the Artificial Intelligence, Information Access, and Mobile Computing Workshop at the 18th International Joint Conference on Artificial Intelligence (IJCAI). Acapulco, Mexico. August 2003.

Common Sense Conversations: Understanding Casual Conversation using a Common Sense Database

Nathan Eagle, Push Singh, Alex (Sandy) Pentland

MIT Media Lab
20 Ames St.
Cambridge, MA 02139
nathan@media.mit.edu
push@mit.edu
sandy@media.mit.edu

Abstract

We introduce a method for situation understanding in natural, face-to-face conversation. Our method combines a network of commonsense knowledge with keyword spotting and contextual information automatically obtained from a wearable device such as a PDA or cell phone. Using this method we demonstrate the potential for high accuracy, detailed classification of conversation topic.

Introduction

We describe a method for developing a situational understanding of face-to-face conversations using speech recognition, commonsense knowledge about the structure of ordinary activities, and contextual information from mobile devices recording and transmitting the conversation. We wish to leverage the ubiquitous wearable computing infrastructure within the workplace, using Personal Digital Assistants (PDAs) and cellular phone headsets in non-conventional ways, to glean information about conversation participants and location. By combining this data with commonsense knowledge about the activities people engage in and the topics they care about, we can infer a clearer picture of the content of conversations and the context of their participants.

Topic spotting is an important area in language understanding. Most research approaches it as a 'bag of words' problem: modeling the word statistics associated with each topic, and then classifying the stream of words against the topic models. In our group's previous work (Jebara *et al.* 2000), we were able to accurately classify a few general subjects like politics, sports, and religion. The disadvantage of this approach is that it requires a relatively large number of observed words for accurate classification, and is capable of only coarse topic classification. Our goal now is to more quickly recognize the gist of conversations, and to and identify finer-grained aspects of the participants' situation For example, if Sarah is talking about how hard it is to obtain tickets to the football game, we would like to move beyond recognizing '*sports*' as a topic to recognizing '*buying tickets to a sports event'*.

Traditionally, recognizing such fine-grained events required employing word-by-word parsing and semantic interpretation, which in turn requires a quiet environment, head-mounted microphone, and familiar discussion format. Our approach is to instead 'skim' a conversation stream and identify aspects of the situation given only spotted keywords. By leveraging prior information regarding the relationship between keywords, topics, and contextual information, conversation understanding can be substantially augmented.

Our approach may be described as using a common sense database network to condition or regularize the results of keyword spotting. This is necessary because keyword spotting noisy and unreliable, but more importantly because natural conversations often do not include the keywords needed to identify the situation: the speakers have shared context and background knowledge, so there is no need to mention these commonly shared items. Our method is to use a common sense database to reintroduce all of the commonly shared knowledge, thus regularizing the sampled keywords and permitting more robust and efficient estimation of conversation topic.

The practical value of directly combining perceptual inferences of keywords and context with propositional representations is that the two representations complement each other. It would take a tremendous quantity of sensory data to allow perception to infer commonsense rules like "weddings have a bride and a groom", or "the parents of the bride and groom usually attend the wedding". At the same time, the kinds of sensory models that are easy to learn from modest quantities of perceptual data are usually very difficult to supply via traditional knowledge engineering methods.

Wearable Computing: Information Gathering

PDAs and cellular phones currently have been adopted as part of corporate attire, yet few applications take advantage of the mobile computing infrastructure that they create. It seems inevitable that the processors currently in the pockets of millions (and similar to those in our desktop computers just a few years ago) will no longer remain relegated to calendar and address book functionality. We propose alternate applications for Mobile Computing: to collect and process interaction information on individuals and groups. As described in (Eagle and Pentland 2003), our approach incorporates linux-based PDAs and cellular phone headsets to collect data on complex social systems over extended periods of time.

Background

Much of the research on mobile devices has focused on context to interpret a user's situation. Recent research within our group include Clarkson's iSensed project, where an individual wore two 180-degree video cameras and a microphone for 100 days. A probabilistic model was trained to classify this context into one of 20 possible situations (such working in lab, eating in a restaurant, climbing up stairs) with an accuracy rate of 97%. (Clarkson 2002) A similar project incorporated two months of audio. Using only the noisy output of a speech recognition engine, it was shown that a modified Markov model could classify broad contextual categories such as 'home' or 'office' with over 90% accuracy. (Eagle 2002) Choudhury's ShortCuts project had multiple participants wear a shoulder-mounted computer that recorded audio energy, body motion and IR tagging to accurately quantify the interactions between the participants and model their social network. (Choudhury and Pentland 2003, Choudhury *et al.* 2003)

Reality Mining Hardware

A system of linux-based PDAs, the Sharp Zaurus, and a variety of both wireless (Bluetooth) and standard wired cellular phone headsets were used for preliminary data collection. The PDAs were equipped with CF 802.11b wireless cards that allowed the audio to be streamed to a central sever.

Access point and wireless traffic information are also collected at thirty-second intervals. Information such as access point name and signal strength can be used to infer approximate location, while wireless packet sniffing can detect other participants streaming audio nearby.

Audio Processing and Transcription

ViaVoice, a commercial speech recognition engine, is used to transcribe the audio streams, however typically word recognition rates fall below 50% for spontaneous speech recognition. (Eagle 2002)

The inaccuracy of the speech recognition engine poses a serious problem for determining one of the most critical features in this dataset: the gist of the interaction.

A human can read through a noisy transcript and still have an impression of the underlying conversation topic, as shown by the following example:

> **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 riderlessI 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 two conversation participants are in an office, or more precisely, by a printer, would help many people understand that the participants are trying to fix a jammed printer. Assuming some prior knowledge about the participants (for example, if one of them speaks about repairing printers on a daily basis because he is a printer repairman), may significantly augment a person's ability to infer the gist of the interaction.

As will be shown, this additional contextual and commonsense information can be used to form a probabilistic model relating observed keywords to conversation topic. Thus by combining audio and contextual 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*'.

OMCSNet

We make use of OMCSNet, a large-scale semantic network built at the Media Lab by aggregating and normalizing the contributions from over 10,000 people from across the web. (Singh 2002) It presently consists of over 250,000 commonsensical semantic relationships of the form 'a printer is often found in an office', 'going to a movie requires buying a ticket', and so forth. OMCSNet 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. OMCSNet has been used in a variety of applications to date. (Liu & Singh 2002)

OMCSNet uses a hybrid knowledge representation strategy where individual concepts are represented linguistically (lexically and phrasally), and are related by a small set of about twenty specific semantic relationships such as LocationOf, SubeventOf, HasEffect, and so on. At present OMCSNet employs the 20 binary semantic relations shown below in Table 1. A small section of OMCSNet is show in Figure 1.

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

Table 1. Semantic Relation Types currently in OMCSNet

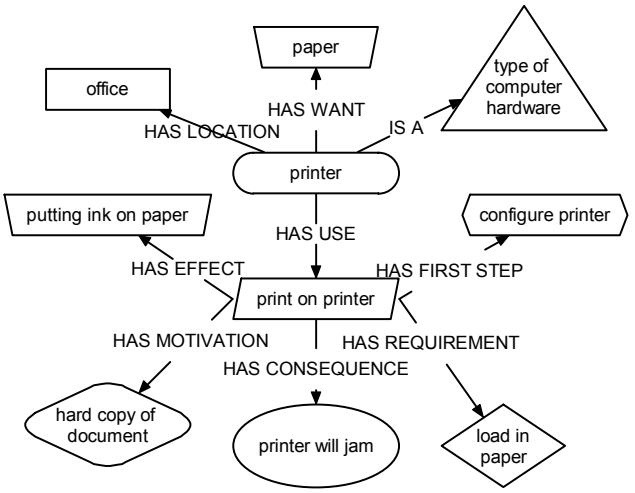

Figure 1. A Selection of OMCSNet's 250,000 Relations

Fine-grained topics: Gists

Our system's goal is to infer the 'fine grained topic', or gist, of the 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 simple verb phrases. For our set of target gists, we use the 700 most richly defined situational aspects within OMCSNet (those for which at least 10 facts are asserted.)

One feature that distinguishes commonsense reasoning from other forms of reasoning is that it involve making inferences using many different kinds of knowledge: about objects, events, goals, locations, and so forth. Accordingly, our system uses a probabilistic model that incorporates different types of knowledge, as well as contextual data from the mobile devices.

Inference in OMCSNet

Inference over the OMCS network can be done with varying levels of complexity, ranging from simple network analysis metrics to probabilistic modeling using Bayesian networks.

Preprocessing the Transcriptions

Before the inference, the transcriptions are preprocessed to reduce the noise of the speech recognition engine and improve inference performance. The transcriptions are first lemmatized and filtered for stop words (such as 'like', 'the', 'a', etc). A second filtering process is then 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.

Inference using the Word/Gist Bipartite Network

By flattening the networks of the different relationship types, a bipartite network can be formed to incorporate all ties from words 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). As will be shown in the Experiments section, this method is often capable of identifying a small group of potential gists, frequently with one dominating the others.

Context for Gist Differentiation

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.

A more sophisticated model was developed to incorporate the conditional probability distributions determined by training data.

Bayesian Inference in OMCSNet

We can represent the system with a Bayesian network of observations (keywords) and the gist as the hidden state of the network. To determine the probability of the gist given the observations (words: W, locations: L, and participants: P) we need to define our probability distributions conditioned on each gist (for G total gists), as shown below:

$$P(g|W,L,P) = \frac{P(g)P(W|g)P(L|g)P(P|g)}{\sum_{i=1}^{G} P(W|g_i)P(L|g_i)P(P|g_i)}$$

If information about the participants identities were available, their conditional probability distribution (CPDs), $P(P|g)$, could be obtained either from surveys about their common topics of conversation, or learned given adequate training data. The CPDs involving words can be a function of the network topology within OMCSNet:

$$P(W|g) = \sum_{w=1}^{n} \frac{k_w}{d_g}$$

where n is the total number of words W, k_w is the non-zero number of links between the gist and the word, w, and d_g is the `degree' of the gist (ie: its total number of links). As discussed in Future Work section, an extension would be to calculate the CPD as being inversely proportional to the degree of separation between the two nodes in the graph. A similar method can be implemented for location CPDs, $P(L|g)$, using only the location ties within the OMCSNet.

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.

Example - Inference with Context

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	4 play harp	11	cook food

Table 2. Results of using Context for Gist Differentiation

Actual Situation:

Deciding what to get for lunch in the cafeteria

Results

Each row in Table 3 represents a distinct interaction that was later characterized by the participant. Using the method described above, a ranking of the top ten gists for each interaction was created. The model gave a correct gist the number 1 ranking in 40% of the tests. [1] In 70% of the tests, a correct gist was one of the top ranking three. However in 25% of the tests, a correct gist was ranked outside the top ten, represented by Ø in the Table 3.

[1] A correct gist of the conversation does not need to be the exact topic, but rather the general idea, as determined by a human judge. For example, in the conversation regarding buying fresh vegetables – a correct gist could be 'grocery shop', 'going to the market', or even 'buy food'.

TOPIC	Correct Gist Ranking
Stock market questions	1
Divorce complications	1
Marathon training	1
Studying for upcoming final	1
Trying to remember something	Ø
Playing basketball	Ø
Discuss love interest	Ø
Buying a present for a girlfriend	Ø
Learning how to surf	1
Driving a car	2
Having to pay a bill	1
Needing to buy fresh vegetables	2
Making a shopping list	5
War on Iraq	2
Reading newspapers	3
Vacation	1
Feeling sick	3
Questioning a waiter	Ø
Writing a paper	2
Trying to use the printer	1

Table 3. Accuracy Results of the Gist Inference

Discussion

The amount of help from location information seems to vary significantly depending on the topic of conversation. Office conversations regarding current love interests occur just as much, if not more, than office conversations about broken printers. Once information on the participants is incorporated, we expect the model's performance to increase considerably.

Future Work

We have shown that inference of conversational topic by combining commonsense and mobile computing is a promising research direction. As this research matures over the next few months we expect significant improvements to the probabilistic model, the additions of participant information, as well as compelling applications enabled by a real-time implementation on the mobile devices.

Much of the data within OMCSNet has yet to be fully leveraged by our probabilistic model. Using Probabilistic Relational Models, we hope to better exploit the information inherent within different link structures. Inverse distance metrics, perhaps derived from Dijkstra's algorithm, could be used to model the CPDs between words and topics in the bipartite graph. The probability of a specific observation given a gist would thereby be a function of the distance between the two nodes in OMCSNet. The CPD would be modeled as a power law:

$$P(w|g) = \frac{1}{N}\sum_{w=1}^{N}\frac{1}{2^{\frac{d(w,g)}{S}}}$$

where $d(w,g)$ is the shortest path between word w and gist g, N is the number of selected words, and S is the average tie strength.

Conclusions

We believe it is possible to build commonsense-enabled wearable systems that are knowledgable about face-to-face conversations. We hope to benefit from integrating both perception and commonsense reasoning. In the one direction, commonsense knowledge can greatly expand the range of inferences one can draw from a small amount of perceptual and contextual information. In the other direction, perceptual systems can situate our commonsense reasoning systems in real-world contexts so that they can reason about situations and events as they happen. It is our hope that this approach will enable an entirely new class of context-aware and context-dependent applications that can make a broad range of commonsensical inferences in order to aid us in ways they reason will help us achieve our goals.

Acknowledgements

This work was partially supported by the NSF Center for Bits and Atoms (NSF CCR-0122419).

References

Choudhury, T., Pentland, A. 2003. The Sociometer: A Wearable Device for Understanding Human Networks, P. 96, Sunbelt XXIII, Cancun MX. Also Technical Report, TR-554, The Media Lab, MIT.

Choudhury, T., Clarkson, B., Basu, S., and Pentland, A. 2003. Learning Communities: Connectivity and Dynamics of Interacting Agents. To appear in the *Proceedings of the International Joint Conference on Neural Networks - Special Session on Autonomous Mental Development*. 2003. Portland, Oregon.

Clarkson, B. 2002. Life Patterns: structure from wearable sensors. Ph.D. diss., The Media Lab, MIT.

Eagle, N. 2002. Information Explication from Computer-Transcribed Conversations: The Weighted Markov Classifier, Technical Report, TR-560, The Media Lab, MIT.

Eagle, N. 2003. Reality Mining: Quantifying knowledge networks with sensors, P. 46, Sunbelt XXIII, Cancun, MX, Also Technical Report XXX, The Media Lab, MIT.

Liu, H. and Singh. P. (2003). OMCSNet: A commonsense inference toolkit. In submission.

Jebara, T., Ivanov, Y., Rahimi, A., Pentland, A. 2000. Tracking Conversational Context for Machine Mediation of Human Discourse. *Proceedings of AAAI Fall Symposium on Socially Intelligent Agents* : AAAI.

Singh, P. 2002. The public acquisition of commonsense knowledge. *Proceedings of AAAI Spring Symposium on Acquiring (and Using) Linguistic (and World) Knowledge for Information Access.* Palo Alto, CA: AAAI.