Topic Spotting Common Sense Translation Assistant

Jae-woo Chung
M.I.T. Media Lab
20 Ames Street,
Cambridge, MA 02139-4307
jaewoo@media.mit.edu
617-253-0666

Rachel Kern
M.I.T. Media Lab
20 Ames Street,
Cambridge, MA 02139-4307
rachelk@media.mit.edu
617-253-7120

Henry Lieberman
M.I.T. Media Lab
20 Ames Street,
Cambridge, MA 02139-4307
lieber@media.mit.edu
617-253-0315

ABSTRACT

Our Translation Assistant applies common sense logic to the problem of translating speech in real time from one language to another. Using speech recognition combined with a software translator to do word-by-word translation is not feasible because speech recognition is notorious for poor results. Word-by-word translation requires grammatically correct input to translate accurately. Therefore, translation of speech that is potentially already fraught with errors is not expected to be good. Our Translation Assistant works around these problems by using the context of the conversation as a basis for translation. It takes the location and the speaker as input to establish the circumstances. Then it uses a common sense knowledge network to do topic-spotting using key words from the conversation. It only translates the most likely topics of conversation into the target language. This system does not require perfect speech recognition, yet enables end-users to have a sense of the conversation.

Author Keywords

Language Translation, Topic-spotting, Common Sense Reasoning

ACM Classification Keywords

Computer-Mediated Communication, Agents and Intelligent Systems, Speech I/O, Internationalization/ Localization

INTRODUCTION

Common sense reasoning is a relatively new field in the realm of artificial intelligence, which posits that machines need to know mundane facts about the workings of the world in order to reason about everyday life in much the same way that humans do [5]. Research in this domain shifts the focus of artificial intelligence from designing rule-based expert systems, in which the computer has a great deal of knowledge about one subject, to designing common sense applications, in which the computer knows a little bit about a wide variety of topics, and is able to exploit its knowledge about a vast array of subjects to solve real-world problems. The focus is on breadth of knowledge, rather than depth.

Topic spotting in conversation is one application of common sense reasoning. The goal of topic spotting is for a computer to identify the gist of a conversation based only on keywords, rather than using a more exhaustive grammar-based parsing mechanism [3]. The input is made via regular, casual speech, and speech recognition software is responsible for transcribing the audio into text. Although speech recognition technology is notoriously poor, when it is coupled with a database of facts of common sense knowledge and given some context information, it is possible to take just the words that the recognizer correctly identifies and infer from these words the most likely topics of conversation. For example, if the speech recognizer in a topic spotting system only recognizes the words “bride,” “ring,” “white,” and “dress,” it may draw upon its common sense knowledge database to guess that one of the most likely topics of conversation is a wedding.

Conversational topic spotting actually has many similarities to how a non-native speaker may understand a new language. Although native speakers generally talk too fast for the non-native speaker to grasp each word, if he can comprehend even 50-60% of the words and combine that knowledge with what he already knows about the world from common sense, then he can have a good sense of the gist of the conversation. This similarity between common sense topic spotting and understanding of a foreign language provided the motivation behind the current research. The idea was that we can use speech recognition to recognize at least 50% of the words that are said, translate just these recognized words into another language, and combine this with context information and a database of common sense facts to give end-users a fairly good understanding of the conversation. Therefore, our system should not be thought of as a word-for-word translator, but more of an aid that translates just enough speech, that when combined with context and common sense logic, can enable end-users to figure out roughly what is being discussed.

CHI 2005, April 2–7, 2005, Portland, Oregon, USA.
ACM 1-59593-002-7/05/0004.

Our system also contains an additional component, which is essentially a dynamically generated phrasebook, enabling the end-user to respond to the conversation in the same language in which the conversation is taking place. Similar to GloBuddy [4], another common sense-based translation tool, our system takes the topics guessed by the topic spotting mechanism, and uses them to generate phrases that an end-user might say in response. These phrases are displayed on the screen in the end-user's native language, but when he or she selects a phrase, a text-to-speech engine speaks it aloud in the language in which the conversation is taking place, so that the other participants can understand it. Thus, with the help of this phrasebook and text-to-speech engine, our system ultimately enables a two-way conversation between participants who do not speak the same language.

SYSTEM AND INTERFACE

Domain and Language Restriction

Our first task in building a prototype of our conception of the translation assistant was to narrow the domain of possible conversation topics. Because topic spotting via common sense reasoning is not yet sophisticated enough to work well across many domains, we restricted our domain to a particular situation, in order to optimize our results. We chose the domain of a sick individual seeking help in a hospital setting, interacting first with a receptionist, and then with a doctor. We also had to select two languages to translate between. We only had an English speech recognizer available to us, so we decided to build a system that translated from English to Korean. Korean was chosen because one of the authors of this work is fluent in it.

In our imagined scenario, the patient is the Korean-speaking end-user of the device, trying to communicate with an English-speaking receptionist and doctor. The device takes the location and the identity of the person to whom the user is speaking as input before the conversation even begins, in order to establish context. In our situation, we tell the system that we are in a hospital, speaking with either a doctor or a receptionist. Providing context, in addition to aiding in topic spotting, is also necessary for optimizing speech recognition results. For example, if the speech recognizer makes substitution errors that do not make sense in the context of a hospital setting, our system will ideally know enough to throw these words away instead of trying to incorporate them into the conversation.

Speech Recognition and Text-to-Speech Engines

We are using IBM's ViaVoice as our speech recognition tool. Because this software is speaker-dependent, the system will not work well for speakers who have not previously trained ViaVoice to recognize their voice. For the initial prototype, however, ViaVoice was easily incorporated into our architecture and was thus a logical option for testing purposes. For the text-to-speech engine, we used Microsoft Reader.

System Details

Our system works as follows. First the context is established by identifying the location and the speaker. Then the speaker starts talking in English. A microphone captures the speech and the speech recognition software transcribes the audio into text. From this text, irrelevant or insignificant words are filtered out (for example, is, a, the, etc.) and the rest of the words are fed into a topic-spotting tool based on ConceptNet [1], a large-scale semantic knowledge base built upon the Open Mind Common Sense database [6]. ConceptNet is optimized for making practical context-based inferences about real-world situations. The topic-spotter produces related concepts based on the results of the speech recognition, and these concepts are translated into Korean and displayed on the end-user's screen in order of greatest likelihood, with the most likely topics at the top of the list. The end-user is now able to read through this list to have a general understanding of what the speaker said. The speaker's facial cues and gestures should convey additional meaning to the end-user.

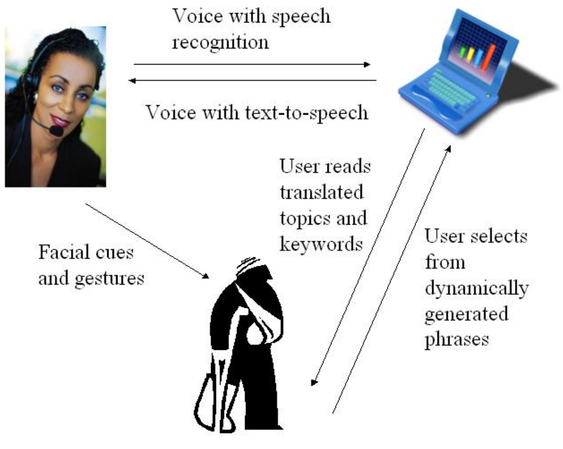

Voice with speech recognition
Voice with text-to-speech
User reads translated topics and keywords
User selects from dynamically generated phrases
Facial cues and gestures

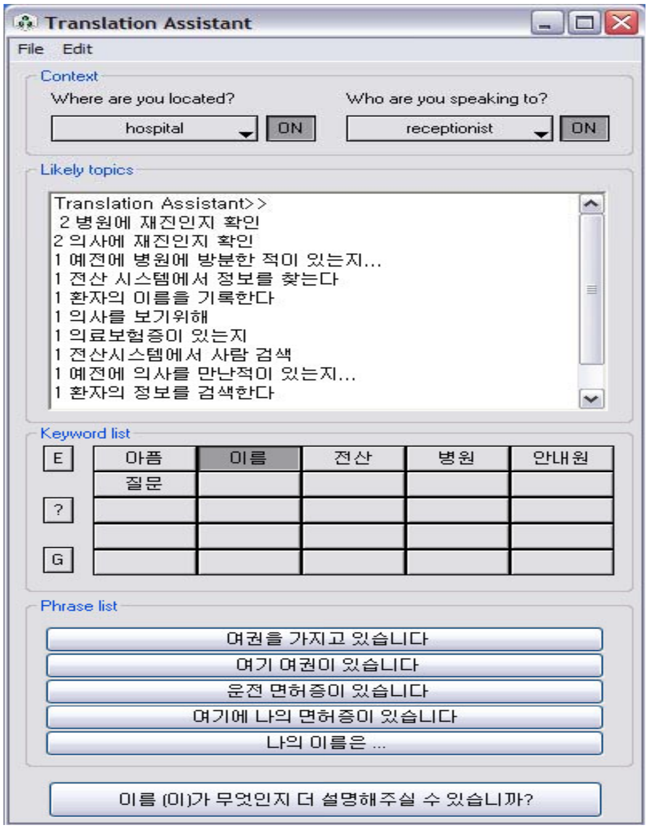
Figure 1. Communication flow.

The filtered, speech-recognized words are also displayed in a grid below the list of likely topics. The end-user can select one or more of these recognized words. Based on which words are selected, relevant phrases are generated and displayed in Korean. These phrases are stored in a phrasebook database that we created specifically for this system. The phrasebook was inspired by GloBuddy [4], and is made up of keywords associated with related phrases. Several general expressions are also available to the end-user. Once the end-user selects a phrase, a text-to-speech engine will say this phrase aloud in English, so that the speaker can understand it. The speaker can then respond verbally, and the conversation proceeds in this manner. Figure 1 provides a diagram of how the system works.

Interface Details

Figure 2 shows a screenshot of the end-user's interface. The context is established at the top of the screen, in the section marked "Context". The first drop-down menu lets the user indicate the location, which in this case, is a hospital. In the second drop-down menu, the user specifies the person to whom they are talking. A receptionist is selected in the screenshot.

The second section of the interface is the "Likely topics" section, which displays a list of the ten most likely topics of conversation. The numbers to the left of each phrase indicate the likelihood of that phrase being a topic of conversation. The greater the number, the more likely the topic is. In Figure 2, the most likely topic shown "see if person has been to our hospital before". The second most likely topic is "look up person in database". As the speaker continues to talk and the speech recognition software feeds new words into the Translation Assistant, this list of likely topics will change in accordance with the newly added information.



Where are you located? hospital ON

Who are you speaking to? receptionist ON

Translation Assistant>>
2 병원에 재진인지 확인
2 의사에 재진인지 확인
1 예전에 병원에 방분한 적이 있는지...
1 전산 시스템에서 정보를 찾는다
1 환자의 이름을 기록한다
1 의사를 보기위해
1 의료보험증이 있는지
1 전산시스템에서 사람 검색
1 예전에 의사를 만난적이 있는지...
1 환자의 정보를 검색한다

아픔	이름	전산	병원	안내원
질문				

여권을 가지고 있습니다
여기 여권이 있습니다
운전 면허증이 있습니다
여기에 나의 면허증이 있습니다
나의 이름은 ...

이름 (이)가 무엇인지 더 설명해주실 수 있습니까?

Figure 2. Interface screenshot.

The third section of the interface is entitled "Keyword list," and is used in conjunction with the fourth section, the "Phrase list". The Keyword list is a grid of the speech-recognized words with the insignificant words filtered out. The end-user selects from these recognized words in order to display appropriate phrases in the Phrase list section.

This section also has three special buttons on the left-hand side. The top button, "E," stands for English and lets the user change the whole display from Korean to English. This may be useful for end-users who have some understanding of the English language; the ability to switch the display from Korean to English may further aid in their comprehension of the dialogue. The second button, labeled with a question mark, generates question phrases, as opposed to statement phrases. If the question mark button is not selected, only statements will be generated in the Phrase list. If it is selected, only questions will be generated. The third button, labeled with a "G," stands for general, and brings up a list of general phrases that are always available to the end-user, regardless of the situation. The general phrases include the following: "I don't understand, can you please repeat that?" "I do not speak English." "Can someone help me?" "Can you please write that down for me?" and "I don't understand, can you please rephrase that?"

There is one more question phrase at the very bottom of the screen. This phrase says, "Can you please elaborate about the word ___?" The word that fills in the blank is the most recent word that the end-user selected from the Keyword list. The user can employ this tool to prompt the speaker to further describe any word that the user may not be familiar with, or that may not have been translated correctly.

EVALUATION

We conducted an informal study of our prototype system with two native Korean speakers. The first part of the evaluation task was for the Korean user to have a conversation with the hospital receptionist and to schedule an appointment with a doctor. The second part was to have a conversation with the doctor and to receive a diagnosis. One of the authors of this work played the roles of both the English-speaking receptionist and the doctor in the study. Because the Korean subjects also understand English, we had to locate them in a different room from the receptionist/doctor in order to obtain useful results. To still achieve the benefits of added meaning from visual and facial cues, we used a Webcam so that the Korean end-user and the receptionist role-player could see each other while using the system.

We divided each user session into two sections. In one section, we used the speech recognition software to capture the English speaker's input. In the other section, the speaker entered input by typing, rather than speaking.

Results

Though our study was informal, it did yield some interesting results. First of all, both subjects thought that overall, the interface was very simple and easy to use. They initially required some time and explanation to learn the system, and had to play with it to understand the mechanism of generating phrases by selecting different combinations of keywords. After the first use of the system, however, they commented that our interface could be learned relatively quickly. The subjects also commented that seeing the facial cues and gestures of the speaker was beneficial to them in understanding what the other person was saying. This implies that if the device were actually used in a real-life situation with two people communicating face to face, rather than over a Webcam, these visual cues would be further pronounced, and potentially even more valuable in conveying meaning than in our contrived study setting. Another interesting result was that there was not much difference in the quality of the conversation, regardless of whether the input was made by typing or speaking. When asked afterwards, the users said they did not notice a difference between the two sections, and could not tell that the input was entered in different ways. This implies that using speech recognition did not detract from the usability of the system.

The study also highlighted some areas in which we could make improvements. Both subjects commented that after each English utterance, so many likely topics were presented that it became too time-consuming to read through all of them; this resulted in slowing down the whole interaction. The subjects also expressed the wish that more keywords had been available to select from when choosing words for the phrase generation. They felt that they did not have a wide enough array of alternatives when picking response phrases. Finally, the subjects complained that it was confusing to have to read through both the list of likely topics as well as the grid of keywords used to generate response phrases. They tended to focus more on the keywords to generate an appropriate response, at the expense of concentrating on the likely topics. Despite these suggestions for improvement, both subjects were able to complete the tasks given, and successfully communicate to the speaker; this proves that our system can actually work, and that our idea is indeed viable. With further enhancements to the system, we would expect an even greater success rate and more positive feedback from users in the future.

FUTURE WORK

This research introduces several implications for future work to be done with this system, and in the area of common sense-based translation in general. Specifically in our current system, our informal evaluation suggests that we should limit the number of likely topics presented. Currently, our interface displays a maximum of ten likely topics. In our next version, we would like to limit this to show only five likely topics at most. This should aid both in the speed of the interaction and in the quality of the end-user's comprehension. The evaluation also implies that our phrasebook should be expanded. The subjects requested that more keywords be available to choose from in order to generate more phrases. This would give the end-users more choices so that they can express themselves more accurately. Another implication of our evaluation is that perhaps the list of likely topics and the grid of phrase-generating keywords should be merged into one window, rather than separated into two. This would narrow the end-user's focus to one area, and might also aid in making the interaction a bit faster overall.

Once the system is robust enough, another obvious consideration for the future would be to expand it to encompass more domains and more languages. In addition to expanding it, we would also like to explore the possibility of porting it to a hand-held, mobile device, so that end-users will be able to take the system with them and use it in all types of situations. This will also require the system to use speaker-independent speech recognition.

CONCLUSION

While conducting our user evaluation, it became apparent that we had built a system in which one person was speaking and responding only in English, while the other person was using only Korean to speak and respond. There was no overlap between the uses of the two languages, and yet the system still permitted a feasible two-way communication. Thus, using common sense as the only common ground, our system enables two people, who normally could not sufficiently communicate, to carry on a reasonable conversation, albeit within a narrow domain. This is the most powerful result of our work.

REFERENCES

1. Liu H. and Singh P. ConceptNet – a practical common sense reasoning tool-kit, *BT Technology Journal 10* (2004)
2. Lieberman H., Liu H., Singh P. and Barry B. 'Beating some common sense into interactive applications', *AI Magazine* (Winter 2005)
3. Eagle N., Singh P., and Pentland A. 'Common Sense Conversations: Understanding Casual Conversation using a Common Sense Database', In *Proc. IJCAI 2003*, ACM Press (2003), 1163 - 1166.
4. Musa R., Kulas A., Anguilete Y., Scheidegger M. GloBuddy, A Broad-Context Dynamic Phrasebook. In *Proc. CONTEXT '03*, AAAI (2003), 467-474.
5. Minsky M. Commonsense-based Interfaces. *CACM*, 43,8 (2000), 67-73.
6. Singh P., Lin T., Mueller E.T., Lim G., Perkins T., Tompkins M., and Zhu W.L. Open Mind Common Sense: Knowledge Acquisition from the General Public. In *DOA/CoopIS/ODBASE 2002 Confederated International Conferences* (2002), 1223-1237.