Curtis Kularski

Dr. R.F. Lunsford

English 6161, Introduction to Linguistics

9 December 2015

Artificial Intelligence Speech Recognition with Paralinguistic Features and Absent Context

Introduction

Artificial intelligence is one of the most promising fields to aid human-computer interaction. Creating methods to give computers the functionality of imitating intelligence by adapting their own responses to information they are given or utilizing machine-learning techniques to expect or anticipate needs has the intended benefit of simplifying human interactions with computers (Dinsmore and Moehle 13). One of the more troubled areas of artificial intelligence is voice recognition. Computers function in specific binary (digital) commands, and human language is laden with ambiguous lexical terms, internal contradictions, cultural references and tonal and inflectional differences. How are artificial intelligences to cope with these complexities to become fluent in human verbal communication? This paper will examine how artificial intelligences cope with tonal and inflectional differences and the absence of a contextual[1] awareness in speech recognition.

The Turing test has been a much debated pseudo-standard for artificial intelligence since the 1950s. In the test, human subjects are asked to communicate, by typed language, to conversational

[1] Context in this paper shall refer to social and cultural contexts and not the linguistic context (surrounding words), unless otherwise noted.

"partners". One conversational partner is an artificial intelligence. The objective is for the human subject to be unable to prove which partner is a computer. The test itself is more of a test of language comprehension than of a functional intelligence (Proudfoot 952). This illustrates how strongly factors of communication are tied to the concept of intelligence. The Turing test is normally administered in only a text-based medium, which frees the computer from needing to speak in a way that is convincing to human auditory processing and from any need to be able to understand the nuances of human speech (Dinsmore and Moehle 13). Despite the Turing test being administered only in text, it does capture the objective of the artificial intelligence field: to create a general artificial intelligence which can interact with humans in a natural way.

From a human-computer interaction perspective, communication through speech has become a requirement for artificial intelligences and assistive computing systems in recent years. Ideally an artificial intelligence would be able to understand and comprehend language in the same way as an organic intelligence and then generate a response or perform an action based on the comprehended meaning. At this point computers are restricted to reacting to either simple spoken command given from a limited set of known commands or to limited conversation based in a narrow topic area (Mesnil, Dauphin and Yao 530). Essentially the restriction is that computers can verbally interact with humans within the limited confines of a specific programmatic scope. General artificial intelligences which can hold conversations across broad topics do not yet exist. Humans perform natural language comprehension and interpretation with ease (Voskuhl 393).

Problems in Natural Language Processing by Artificial Intelligence Agents

Primary speech recognition processing for artificial intelligences processes auditory data, reduces background noise and 'normalizes' the verbal information to reduce tonal variations and other

cues that may inhibit transcription into a text-based representation. This approach is quite effective at producing nearly perfect transcription when provided with a well-formed sample (Turk and Arslan 441-442). The problem is that in the process of cleaning the auditory data in search of content, the algorithmic processing eliminates a large amount of data that human recipients of verbal communication would find to be important. Many in the fields of artificial intelligence, computer speech recognition and linguistic technology have noted this limitation and have begun developing ways to cope with the current inability of computers to handle that type of auxiliary speech/auditory data (Voskuhl 393-395).

Reductionist techniques such as the normalization of verbal information are common in human-computer interaction because humans produce a large amount of information and there are technical limitations to the amount of information that a computer can store and process. Most of the reduction occurs through a process called pattern recognition, which allows for the encoding of complex input, such as natural human language, into less complex digital information that has meaning to the artificial intelligence/computing system. For natural human language transcription this is most often accomplished by comparing patterns in spoken language to previous encoded language samples for which there is already a known translation. The pattern recognition process is independent of speakers, but can be impacted by localized pronunciation differences (Voskuhl 396-398). Pattern recognition in which natural language is compared to language prototypes eliminates distinctive voice qualities such as tone, inflection and volume. All that is left is a symbolic interpretation, either as textual words or reference to a specific entry in a lexical database, which represents the original utterance (Voskuhl 395). This leaves out information for non-tonal languages, but is especially problematic for processing languages that are tonal.

Humans output a lot of data in their use of natural language, but it is insignificant when compared to the amount of information that humans use when interpreting natural language spoken by another human. Humans use tone of voice, word order, visual cues, situational context, cultural background and a variety of other pieces of data to make sense of spoken utterances. Take two possible utterances: “How do humans recognize speech?” and “How do humans wreck a nice beach?” (Allen 256) Each utterance has a distinct meaning, or in this case asks a distinct question. The utterances can be phonetically pronounced in a similar way. If there is any reason for auditory interference, background noise perhaps, the distinction between the two questions can be lost. If the undistinguished utterance is spoken in a linguistics classroom, it is likely to be interpreted as the first version, whereas if it is heard in an ecology classroom it is likely to be interpreted as the second. Humans have little difficulty with processing partial recognition data because they can often fill in with auxiliary information or context. In essence human speech has built-in redundancy to protect the signal (Allen 252). Computers have difficulty with partial recognition because they either do not perceive the redundant data or they discard it as part of their reduction and encoding efforts to make language processing more efficient (Voskuhl 395).

Paralinguistic Features for Artificial Intelligences

The current state of literature on artificial intelligence interpretation of human language with regard to special auditory paralinguistic features (tonality and inflection) is centered on teaching artificial intelligences how to determine the emotional state of the speaker, which can give the artificial intelligence an additional component of information to use in deciphering the meaning that the human speaker is attempting to convey. Emotion is a feature of interest for researchers in computational linguistics and artificial intelligence because there is a pattern that can be more easily prototyped than

more subtle tonal cues (Batliner, Steidl and Schuller 5), although some researchers are working in the area of tonal differences for vowel sounds in some languages (Plonkowski 187; Rotovnik, Maucec and Kacic 438).

Computer speech recognition of tonal and strongly-inflectional languages is processed using techniques similar to the basic pattern recognition used for non-tonal languages and languages with limited inflections. The density of inflectional languages can present a problem for computational systems as they require a larger prototype database. For example, Slovenian, a strongly inflectional language, requires a computational vocabulary that is 10 times larger than that required for English, a language with minimal inflections. The vocabulary is composed of word families, essentially lexical items, and various words that represent the inflected forms. The larger database of prototypes reduces speech recognition efficiency and increases the rate of errors in the recognition. Tonal languages, such as Mandarin, present a similar problem as words must be matched against several prototypes to select the correct tonal variation. With increases in the capability of database technology as well as overall improved computing hardware capability additional methods for handling large vocabularies have been developed. One such method is matching speech patterns at the morpheme level instead of the lexeme or word family level. This technique requires the computer to make multiple passes at speech recognition, but using fewer language units for comparison improves accuracy and speed of the initial conversion to symbols that are usable by the artificial intelligence. During the second interpretive pass the artificial intelligence must reassemble the morphemes into vocabulary terms for the process of parsing meaning (Rotovnik, Maucec and Kacic 436-438). There are differences between real time analysis (continuous analysis) and transcription analysis. Real time analysis is most often used for conversational interactive artificial agents, whereas transcription analysis is used for transcribing spoken language in a situation where the timing of the results is not critical. Real time analysis often relies on a best guest interpretation of what has been spoken. Best guess interpretation allows the computational

system to quickly compare the auditory input to the prototypes[2]. The comparison results in the selection of one or more language units that are a good fit for the input. Robust speech recognition systems will also perform a statistical analysis given the context in the sentence to determine the likelihood of each of the selected language units being present in the given context (Rotovnik, Maucec and Kacic 338-441). The number of variations possible in strongly inflectional and tonal languages presents a challenge for this method of speech recognition. This method is often slow and inaccurate because of the number of possible choices and the relatively minimal information available to the artificial intelligence. English is considered to be computationally efficient because it is not tonal or highly inflectional, but attempting to account for the various intonation possibilities for each word or phrase would increase the complexity of the language from the computational perspective.

A technique that is available to non-tonal languages that are highly-inflected is the neural network. Neural networks are software constructions that are seeded with initial data, but are then given the capability of creating connections between the various pieces of seed data and also with new network nodes that an artificial intelligence may decide to add to that network based upon frequency of use. The operational purpose of a neural network is to simulate the human brain's network of neurons and synapses to produce an efficient computational structure. Neural networks used for speech recognition are seeded with lexical components, such as nouns and verbs. In inflectional languages they are also given suffixes, affixes and other grammatical modifiers. The network is not given combined forms of the lexical components and grammatical components. It is the task of the artificial intelligence to deconstruct the natural language using the nodes provided in the network and then utilize known rules to extrapolate meaning. The network relies on connections between nodes to perform the speech analysis. The neural network method is not a perfect solution for speech recognition as the process

[2] Prototype matching is carried out by searching an indexed database. Indexing is crucial to retrieval speed.

slows and the number of errors grows as the network becomes larger (Mirkovic, Seidenberg and Joanisse 654-656).

Emotional recognition can be performed either together with or apart from speech recognition. Separate from speech recognition emotional recognition provides an artificial intelligence with valuable information regarding the interaction situation (Guido, Pereira and Slaets 431). Recent research revealed that while emotional recognition can be performed outside of speech recognition, accuracy is greatly improved when both the quality of the language and the content of the speech are taken into account (Batliner, Steidl and Schuller 5). One of the simplest ways in which emotional recognition is conducted is a differential pitch analysis. The artificial intelligence can monitor speech for a period of time to determine an average pitch for a particular speaker and then compare future utterances to that pitch standard. The pitch alone is not particularly revealing of the emotional state of the speaker, but combined with simple speech recognition an artificial intelligence can create an emotional estimate (Batliner, Steidl and Schuller 25).

The addition of voice intensity, formants and temporal characteristics of speech provide an even richer data source from which to determine emotion. Having more data than just pitch allows for emotional interpretation from spontaneous speech acts, compared with needing a baseline measure for pitch alone. The most accurate determinations for emotional recognition based on recent research are for emotions in the categories of neutral, irritation, resignation and excitement. Pitch, formants and temporality of speech utterances are measured acoustically and then analyzed against previously coded speech data. The analysis is handled in a method distinct from the prototype matching techniques used for speech recognition. Individual differences in emotional state made such a prototyping exercise inefficient. Instead a variety of mathematical formulas were created based on natural language data to

model the relationships between the various features in each of the emotional states which the artificial intelligences were expected to observe (Laukka, Neiberg and Forsell 98-99).

Detecting and identifying intonation in natural speech is more difficult than recognizing emotion because intonation is more localized in the moves between word sets. One way intonation has been incorporated into speech recognition is by measuring tilt. Tilt refers to a combination of parameters of speech which includes speech event auditory contour, amplitude of the event and duration of the event. For English speech recognition, event timing is measured by breath spaces and a rapid reduction in sound density (Taylor, King and Isard 501-503). Intonation use can vary by individual speaker and therefore the artificial intelligence must be trained[3] on each individual speaker for highly accurate recognition. If an artificial intelligence has been trained on the speaker, then the intonation information can improve the accuracy of word recognition. Intonation functions as a type of context for the words to improve individual word interpretation. This works as additional information for the statistical pattern matching discussed previously (Taylor, King and Isard 493). There are also meaning-making uses for intonation in speech recognition. In some languages, including English, it is possible to use the same words and word sequence to either give information or ask for information, but the intention of those utterances can only be determined by reading the intonation. Historically this has been a use of language that computers could not decipher, but by measuring tilt and moves in spoken language artificial intelligences can minimally use intonation.

Filling in Contextual Gaps for Natural Language Processing

[3] In speech recognition training refers to the process of a speaker performing a repetitive task demonstrating a language trait so that the recognition system can generate a model or prototype.

Artificial intelligences are limited in their access to context by the available sensors on the computers on which they run as well as the types of sensory information that they are designed to incorporate into their decision procedures. Humans approach speech tasks with full contextual awareness. Humans are skilled at incorporating a variety of sensory and contextual information into information processing. Utilizing additional sensory information in artificial intelligence requires additional computing resources as well as more algorithmic complexity. It is for those reasons that it is not practical with available technologies to enable artificial intelligences with the same degree of sensory consideration as is used in human cognition. Context is important for natural language processing to resolve word ambiguity, correct partial recognition errors and to respond to situational inquiries.

Designers of artificial intelligences resolve the context problem through different means depending upon the intended use of the particular artificial intelligence system. This approach is referred to as “targeted understanding”. Its purpose is to provide enough information to the artificial intelligence to convert the speaker’s natural language into a task-specific representation of the user’s intention (Mesnil, Dauphin and Yao 530).

The most basic context that humans use in conversation is location. The location in both physical space and time provides cues to how to answer many questions. Artificial intelligences typically will have access to time through the computational system on which they operate. Physical location requires the use of other sensors, such as global positioning system (GPS) or a wireless radio that can identify the location of nearby wireless transmitters to allow for location approximation (Mesnil, Dauphin and Yao 530-531). These sensors are most commonly available to assistive artificial intelligences such as those found in mobile phones. Location data provides the artificial intelligence with the means to answer certain types of questions, with the assistance of outside data sources, such

as getting directions or providing a weather report that is relevant to where the speaker is located. Location data can also be used to remove ambiguity either in overloaded terms or in partial recognition by identifying the more probable form of a word that is to be used in a location.

Location is a special case of context because it can be reduced to coordinates; other factors of context such as cultural norms and situation are not able to be encoded into such simple forms that can be used quickly to aid in speech recognition because those separate contexts would also have to be interpreted and encoded into meaningful data.

In the absence of context information artificial intelligences compensate by using a variety of slot-filling techniques. In situations where the artificial intelligence is designed for use by either a single user or a small group of users the artificial intelligence can compile metadata about the user and then use that data to make informed decisions about the intent of the user's speech (Mesnil, Dauphin and Yao 531-532).

Slot-filling refers to the process by which an artificial intelligence attempts to make semantic sense of natural language utterances. Artificial intelligences typically contain prototypes and rules for sentence structures in the language in which they are intended to interact. Using those rules the artificial intelligence can analyze language input for recognition by assigning speech part definitions to each word, allowing the artificial intelligence to predict which word to use in uncertain cases based on more certain words in the phrase or sentence. A technique known as hot filling allows for the use of language rules to predict words based on words previously used. For example, if the speaker used an adjective in the first word position, it would be more likely to select a noun than a verb for the second position and then a verb over a noun for the third position. These techniques are usually applied with the use of a recurrent neural network. A recurrent neural network allows for loops that compare previous node connections as well as the sum of a semantic unit, therefore allowing for predictive

functions that are bidirectional. In essence this allows the artificial intelligence to change its assessment of a previous word based on new information. The artificial intelligence is therefore able to create meaning based on the whole of a phrase or sentence rather than the prediction of individual words (Mesnil, Dauphin and Yao 537).

Machine Learning and Recognition Failures

The reliability of speech recognition for semantic meaning making in artificial intelligences is dependent upon the ability of the artificial intelligence to adapt. A static model for human language, even if it were finely tuned to the language use of a specific user, will be unreliable. This is because of the variability in human language and ways in which social interactions and new experiences change individual human cognition over time. To accommodate this change it is necessary that artificial intelligences can learn. As discussed previously in this paper, neural networks are connected structures of nodes in artificial intelligences. As an artificial intelligence gains more experience, it adds more nodes and more connections between nodes. While the connections are usually not ever broken between nodes, unless proven to be invalid, each connection is given a weight. Connection weights are used in determining the statistical likelihood that given a series of connection, that a specific connection is likely to yield an accurate result (Yu, Varadarajan and Deng 436-438). Regardless of the techniques used for interpreting speech, the result of that interpretation must be considered in a network to produce semantic meaning.

Throughout this paper I have referred to “best-guess”, “probability” and “statistical determination” with regard to how artificial intelligences make speech recognition decisions. This reflects the fallibility of artificial intelligences with regard to their interaction with humans. When humans interact with computers, errors will be made (Charlesworth 447). Historically when this

occurred it was left to the human to resolve the problem, or the computer was allowed to discard the error. In artificial intelligence due to the learning capability it is possible for adaptations to be made based on new input. As long as a basic level of comprehension is reached between the computer and the user, it is possible for the user to make corrections when there are mis-interpretations of speech, just as humans do in conversations.

Conclusion

Paralinguistic features such as tonality, intonation and inflection are important to humans for the smooth conduction of conversations. Historically artificial intelligences have normalized these features to simplify the process of recognizing speech. New research has shown that if ample processing capability is present then these features can improve recognition accuracy and reduce cases of uncertainty. Current research is focused on experimental word sets and clean language samples, but progress is being made toward artificial intelligences that can use more robust features of language to interpret semantic meanings which may in time make verbal human-computer interaction less problematic. Problems of context are more difficult to defeat as the conventional methods available for providing context would provide more data than could be reasonably used in the rapid-interpretation situations that conversational artificial intelligences would be expected to perform.

Works Cited

Allen, Jont B. "How do humans recognize and process speech? ." Ramachandran, Ravi and Richard Mammone. *Modern Methods of Speech Processing*. Springer Science & Business Media, 2012. 251-276.

Batliner, Anton, et al. "Whodunnit - Searching for the Most Important Feature Types in Signalling Emotion-Related User States in Speech." *Computer Speech and Languages* (2011): 4-28.

Charlesworth, Arthur. "The Comprehensibility Theorem and the Foundations of Artificial Intelligence." *Mind & Machines* (2014): 439-476.

Dinsmore, John and Chistopher Moehle. "Artificial Intelligence Looks at Natural Language." *Petics* (1990): 13-35.

Guido, Rodrigo Capobianco, Jose Carlos Pereira and Jan Frans Willen Slaets. "Emergent Artificial Intelligence Approaches for Pattern Recognition in Speech and Language Processing." *Computer Speech and Language* (2010): 431-432.

Laukka, Petri, et al. "Expression of Affect in Spontaneous Speech: Acoustic Correlates and Automatic Detection of Irritation and Resignation." *Computer Speech and Language* (2011): 84-104.

Mesnil, Gregoire, et al. "Using Reccurent Neural Networks for Slot Filling in Spoken Language Understanding." *ACM Transactions on Audio, Speech and Language Processing (IEEE)* (2015): 530-538.

Mirkovic, Jelena, Mark Seidenberg and Marc Joanisse. "Rules Versus Statistics: Insights from a Highly Inflected Language." *Cognitive Science* (2011): 638-381.

Plonkowski, Marcin. "Using Bands of Frequencies for Vowel Recognition for Polish Language." *International Journal of Speech Technology* (2015): 187-193.

Proudfoot, Diane. "Anthropomorphism and AI: Turing's Much Misunderstood Imitation Game." *Artificial Intelligence* (2011): 950-957.

Rotovnik, Tomaz, Mirjam Sepesy Maucec and Zdravko Kacic. "Large Vocabulary Continuous Speech Recognition of an Inflected Language Using Stems and Endings." *Speech Communications* (2007): 437-452.

Taylor, Paul, et al. "Intonation and Dialog Context as Constraints for Speech Recognition." *Language and Speech* (19985): 493-512.

Turk, Oytun and Levent M. Arslan. "Robust Processing Techniques for Voice Conversation." *Computer Speech and Language* (2006): 441-467.

Voskuhl, Adelheid. "Humans, Machines and Conversations: An Ethnographic Study of the Making of Automatic Speech Recognition Technologies." *Social Studies of Science* (2004): 393-421.

Yu, Dong, et al. "Active Learning and Semi-Supervised Learning for Speech Recognition: A Unified Framework Using The Global Entropy Reduction Maximization Criteron." *Computer Speech and Language* (2010): 433-444.