Automatically Learning New Words From Spontaneous Speech

Sheryl R. Young and Wayne Ward

January 1993
CMU-CS-93-118

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Appeared in the *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP-93, 1993.

This research was sponsored by the Department of the Navy, Naval Research Laboratory under Grant No. N00014-93-1-2005; and by the Department of the Navy, Office of Naval Research under Grant No. N00014-93-1-0806.

Keywords: machine learning, spontaneous speech recognition and understanding, acoustic normalization algorithm, independent estimates of recognition accuracy

Abstract

This paper describes the design of a system to learn new words from spontaneous speech input, and presents an initial experiment in detecting the new words to be learned. Learning a new word involves detecting an out-of-vocabulary word in the input, determining its meaning, and adding the word to the system lexicon and grammars. Such learning would enable later recognition, parsing and interpretation of the new words. Most current continuous speech recognition systems map known lexical entries onto all parts of the input. However, in real applications, novel words are frequently encountered. Out-of-vocabulary words are not only misrecognized, but often cause misalignments that cause recognition failure in the surrounding regions of the input. While the ability to automatically detect novel words is important, it is also desirable to be able to understand the new words and add them to a system's language models. In this way they can be recognized and understood when encountered again. In this paper we describe the preliminary version of our automatic word learning system and present an experiment which evaluate the system's ability to detect unknown words.

1. Introduction

This technology is based upon a generic word model that is used for detecting novel words. Our generic unknown word model permits any triphone (context-dependent phone model) to follow any other triphone given n-gram phone transition probabilities. This is similar to the system of Asadi [1]. The triphones in this initial version come from the CMU dictionary of vocabulary independent English triphones [2]. The triphone transition probabilities were trained from a large amount of English text. The initial experiments reported here were based upon a triphone bigrams trained on a 20 million word corpus from the Wall Street Journal.

Our system permits new words to be recognized under two separate paradigms, designed and implemented for speech recognition and understanding systems respectively. For recognition alone, our system requires application specific open word categories to be predefined. These are the word categories that can be augmented with new words, incorporating them into the lexicon and class n-gram language model for future recognition. In the case of speech understanding systems, our system enables open word categories to be dynamically defined by using the existing MINDS-II spoken language dialog system [3]. We have modified MINDS-II and the RTN recognizer so that unknown words can be searched for in any of the word categories associated with the content predictions. Thus, in theory, all categories of words can be automatically extended and incrementally added to the system lexicon and grammar when performing speech understanding.

When new words are found, they are represented by the sequence of triphones matched in the associated region. This sequence is added to the system lexicon. We do not yet attempt to determine whether our newly learned words are actually single words or word strings. The words are then analyzed for most likely word category and semantic meaning using methods similar to and capitalizing upon the textual work of Carbonell [4] and Lytinen [5]. Word category and meaning are then used to add the new words to the system lexicon and recognition grammar and, when applicable, to the parsing grammar as well. Here we maintain representations of competing hypothesis when a unique word category and meaning cannot be determined from a single encounter with a novel word. These are later revised and updated as more information becomes available. In the interim, words are tentatively added to the system grammars. For recognition purposes, we use a statistical category or class grammar. Our system uses a robust semantically based caseframe parser specifically designed for parsing spontaneous speech called Phoenix [6]. The Phoenix parser uses semantic recursive transition nets. To add new words to this grammar, the appropriate nets are incremented.

2. Word Score Normalization

In an initial experiment we sought to determine how well the system is able to reject misrecognitions using acoustic information. Most speech recognizers produce a maximum likelihood word sequence using acoustic models and word-level language models. The scores assigned are a weighted sum of the log probabilities from the acoustic and language models. These scores are not normalized, and don't represent any absolute measure of the match, but are meaningful only in comparison to other hypotheses produced for the same utterance. The

score produced by the recognizer is therefore not really useful directly for rejecting utterances or regions of utterances as misrecognitions. In order to be able to detect misrecognitions we want to normalize the score to give a better measure of confidence in the recognition. We use our all-phone decoding model as the basis for the normalization. An all-phone trigram based decoding is run in parallel with the word-based search. In order to normalize a word score, the score from the all-phone path for the same set of frames is subtracted from the word score. This is in effect estimating the prior probability of the acoustics unconstrained by word or word-sequence models (the denominator of Bayes' equation). We found that this normalization was a good discriminator for some words but not for others and in general still doesn't provide a good confidence measure. In order to turn the normalized score into a confidence measure we use a Bayesian updating method. We attempt to estimate the probability that a word is correct when it has a given score. We do not have enough data to make such estimates for every word in the lexicon, so words were clustered into word classes. The words were grouped according to their broad-class pronunciations. We mapped the set of 49 phones used by or lexicon into a set of nine broad classes. Words with the same broad class pronunciations formed a class. For each class, we formed a set of 10 score bins (ranges). We then took recognizer output and accumulated histograms for each word class. For each bin in a word class we determined the percentage of the time words in the class with a score in the bin were correct. This gives us a direct measure of confidence that a word is correct when it has a given score. This method is often used in situations where several knowledge sources combine, to appropriately weight the contribution from each.

3. Results

In order to determine the performance of this method we trained the histograms on 1000 utterances from the DARPA ATIS2 training data and evaluated on on a separate set of 300 utterances. The ATIS database contains spontaneously spoken queries about airline information. The test set contains words never seen in training. We used the Sphinx I discrete HMM speech recognition system to generate the word hypotheses. The system has a lexicon of approximately 1800 words, including ten non-verbal events. The word-class bigram was trained on approximately 12000 utterances taken from the DARPA ATIS2 training set. It has a perplexity of 55. The 1800 words in the lexicon were clustered into 153 word classes. We set a rejection criteria to maintain 95% correct accepts and determined the ability to reject misrecognitions. On the test set of 300 utterances, with correct acceptance rate of 96%, 38% of the misrecognized words were rejected. In looking at the histograms for the word classes, some had almost perfect classification, while others had only slightly better than chance. So for some word classes, we can very reliably reject misrecognitions on acoustic evidence. The training data in the experiment was not really adequate to estimate the histograms. We will train on a much larger set of data and this should improve performance. However, we believe that certain classes of words will still be difficult to reject on acoustic evidence alone. Our intent is to combine the acoustic measures with other (higher-level) knowledge sources to make the final decision. The word class histograms provide a way of determining how much the acoustic evidence should be weighted in the decision.

References

1. Asadi, A., Schwartz, R., Makhoul, J., ''Automatic Modeling for Adding New Words to a Large-Vocabulary Continuous Speech Recognition System'', *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1991, pp. 305-308.

2. Hon, H.W., Lee, K.F., Weide, R., ''Towards Speech Recognition Without Vocabulary-Specific Training'', *Proceedings of the DARPA Speech and Natural Language Workshop*, 1989, pp. 271-275.

3. Young, S.R., Matessa, M., ''MINDS-II Feedback Architecture: Detection and Correction of Speech Misrecognitions'', Tech. report CMU-CS-92-119, Carnegie Mellon University, School of Computer Science, 1992.

4. Carbonell, J. G., ''POLITICS: Automated Ideological Reasoning.'', *Cognitive Science*, Vol. 2, No. 1, 1978, pp. 27-51.

5. Hastings, P.M., Lytinen, S.L., Lindsay, R.K., ''Learning Words from Context'', *Proceedings of the Eighth International Workshop in Machine Learning*, 1991, pp. 55-59.

6. Ward, W., ''Understanding Spontaneous Speech: The PHOENIX System'', *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1991, pp. .