Meeting Assistant Application

Michel Assayag[1], *Jonathan Huang*[2], *Jonathan Mamou*[1], *Oren Pereg*[1], *Saurav Sahay*[2], *Oren Shamir*[1], *Georg Stemmer*[3], *Moshe Wasserblat*[1]

[1]Intel Corporation Israel, PO Box 1659, Matam Industrial Park, Haifa 31015, Israel
[2]Intel Labs, Santa Clara , CA , 95054 USA
[3]Intel Corporation Germany, Dornacher Strasse 1, D-85622 Feldkirchen/Muenchen, Germany
firstname.lastname@intel.com

Abstract

This paper describes the Meeting Assistant application developed at Intel. Unlike existing human-to-machine solutions, the challenges induced by human-to-human conversations are currently poorly addressed by the industry. In this paper, we describe the capabilities of in-house speech and NLP technologies: online automatic speech recognition, speaker diarization, keyphrase extraction and sentiment detection. These technologies has been adapted to conversational speech domain and integrated into the Meeting Assistant.

Index Terms: automatic speech recognition, ASR, speaker diarization, NLP, sentiment detection, keyphrase extraction, meeting assistant

1. Introduction

In the last decade, several solutions have been developed for human-to-machine interaction. However, the challenges induced by human-to-human conversations are less dealt with than human-to-machine. State-of-the-art transcription systems typically achieve 70% accuracy in transcription for conversational speech. In other words for this kind of data, approximately one out of three words is mis-recognized. In some circumstances like noisy environments, foreign accent, undertrained engines etc., accuracy may fall to 50% or even less. The low accuracy of the transcription can have a dramatic effect on the performance of the speech and NLP (Natural Language Processing) technologies. In this paper, we describe the Meeting Assistant application developed at Intel that demonstrates how speech and NLP technologies have been used on low accuracy conversational speech data. The Meeting Assistant application allows the user to transcribe free speech directly from microphone, transcribe a meeting with multiple participants, save and retrieve saved transcriptions and recordings. During the transcription, keyphrases are extracted; negative and positive sentiments are detected and marked. We use also third party software to translate the transcript of the conversation, and to display the extracted keyphrase as a word cloud.

The paper is organized as follows: we describe briefly the speech and NLP technologies in Section 2; in Section 3, we present how these technologies have been integrated into the Meeting Assistant.

2. Speech and NLP Technologies

In this section, we describe briefly speech and NLP technologies used by the Meeting Assistant.

Automatic Speech Recognition (ASR): the ASR engine used in the Meeting Assistant is a conventional continuous large-vocabulary speech recognizer. The frontend is based on a feature vector of Mel-frequency Cepstral Coefficients. Stacked feature vectors are scored by a Deep Neural Network (DNN). The decoding algorithm is based on dynamic composition of Weighted Finite State Transducers (WFST). The recognition models have been trained for US English speech. The acoustic models are trained using the Kaldi open source toolkit [1] on several thousands of hours of transcribed wideband speech data recorded under different noise conditions. The language model has been trained on about one billion words of web-scraped text data. The recognition vocabulary is prepared from the 200,000 most frequent words that occur in the language model training data. So far, no meeting-specific data has been used yet for optimizing the acoustic or language models.

Speaker Diarization: in the current implementation of the system we assume several minutes of enrolment speech is available for each speaker. Speech is collected from a single microphone without any source separation or preprocessing. The output of the voice activity detection is split into two-second segments, which are then fed into an ivector-PLDA pipeline [2, 3]. Each segment produces a PLDA scores for all participants in the meeting. The person corresponding to the highest score is assumed to be the active speaker for the segment.

Keyphrase Extraction: the keyphrase extraction system is implemented as a lightweight, single term and multi-term extraction algorithm that combines several strategies to extract terms and phrases from documents. The algorithm performs sentence splitting, part of speech tagging and noun phrase chunking to collect candidate keyphrases as potential significant terms. After acquiring a large candidate list of terms, it runs a few filters to remove stopwords and perform morphological analysis to normalize the words where possible. After this step, we apply two ranking strategies and combine the results to get the final ranked list of terms. The Statistical Ranking strategy uses notions of Domain Relevance, Domain Consensus and Lexical Cohesion to rank multi-word terms [4]. This algorithm is basically a variant of the Term Frequency-Inverse Document Frequency (tf-idf) algorithm where terms are heuristically weighted with reference to a large standard background corpus of words. The co-occurence based ranker is similar to an unsupervised, domain independent and language-independent method [5] that essentially creates a word co-occurrence graph to compute scores based on word frequencies and degrees of words. We have combined the above two methods together to rank the keyphrases from candidates.

Copyright © 2015 ISCA

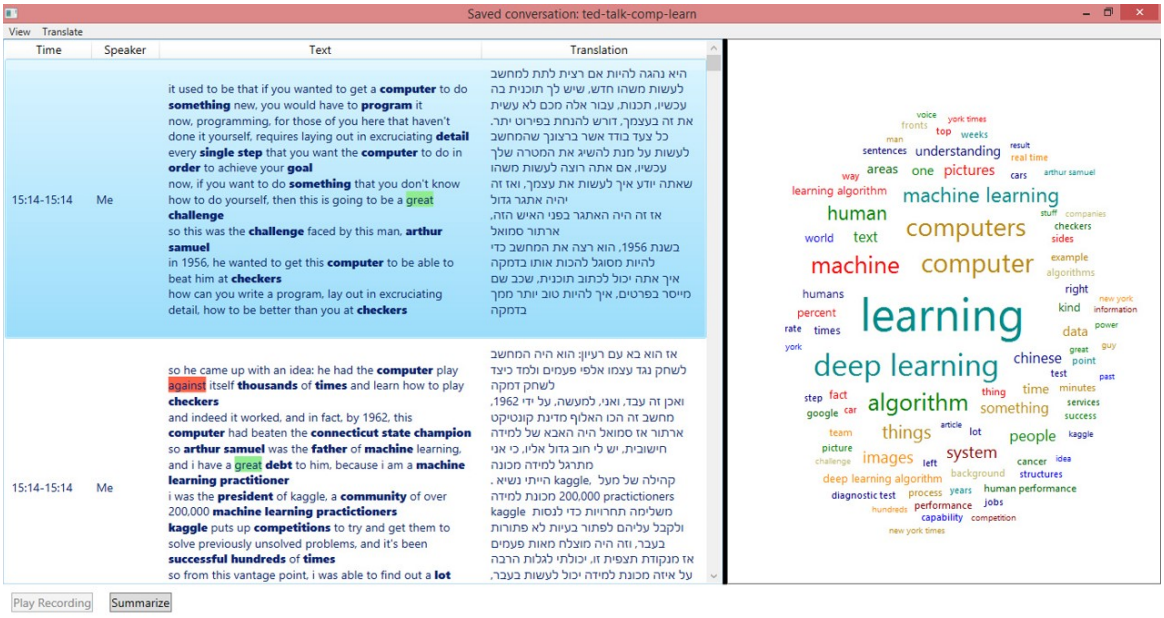


Figure 1: The User Interface of the Meeting Assistant Application.

Sentiment Detection: sentiment analysis is an NLP application that aims to identify and extract subjective information, namely attitudes and opinions form textual documents. The input to the sentiment analysis module in the Meeting Assistant is a textual document that originate from automatically transcribed spoken human to human meeting interaction. The quality of sentiment analysis algorithms in terms of recall and precision is highly dependent on the quality and coverage of the sentiment lexicon. The sentiment lexicon is essentially a list of sentiment words and phrases along with their sentiment polarities. The generation of a sentiment lexicon is a delicate task. Some sentiment terms may convey positive opinion in one topical domain but neutral or negative opinion in other topical domain. That is, lexicon based sentiment analysis is sensitive to the topical domain it operates in. The variety of topics that may be conveyed in human to human meetings is large. In order to achieve high sentiment analysis quality in every domain, there is a need to generate a domain specific sentiment lexicon per domain. Manual acquisition of sentiment lexicons is a costly labor intensive task and therefore impractical for the industry. Semi-supervised sentiment lexicon adaptation methods, such as the method proposed by [6], enable cost effective generation of domain specific sentiment lexicons. The Meeting Assistant uses such methods for achieving high precision and recall percentages of sentiment classification across different domains.

3. Integration

We show in Figure 1 the User Interface of the application. The Meeting Assistant transcribes the audio data and the transcript is displayed along with the NLP metadata. Keyphrases are marked in bold; sentiments are marked – red for negative, green for positive; the summary based on the extracted keyphrases is available. On the right hand side, a word cloud, based on Word Cloud Control[1], displays frequent extracted keyphrases. Translation service is provided by MS Translate[2] cloud service. The application can filter the transcript according to the NLP metadata, e.g., display only detected sentiments. If multiple speakers are recorded on the same audio stream, the speaker diarization process labels the transcribed speech according to its speaker. If multiple clients participate to the meeting, a client-server model is implemented for synchronization. Microsoft Lync is our meeting provider. Each client records its speakers. The new transcript with its NLP metadata is distributed to all the clients connected to the meeting and is displayed according to its timestamp.

4. References

[1] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011.

[2] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, 2011.

[3] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems." in *Interspeech*, 2011, pp. 249–252.

[4] P. V. Francesco Sclano, "Termextractor: a web application to learn the common terminology of interest groups and research communities," in *TIA Conference*, 2007.

[5] S. Rose, D. Engel, N. Cramer, and W. Cowley, "Automatic keyword extraction from individual documents," *Text Mining*, pp. 1–20, 2010.

[6] G. Qiu, B. Liu, J. Bu, and C. Chen, "Opinion word expansion and target extraction through double propagation," *Computational linguistics*, vol. 37, no. 1, pp. 9–27, 2011.

[1] http://www.codeproject.com/Articles/224231/Word-Cloud-Tag-Cloud-Generator-Control-for-NET-Win

[2] http://www.microsoft.com/translator/