Sub-lexical Dialogue Act Classification in a Spoken Dialogue System Support for the Elderly with Cognitive Disabilities

Ken Sadohara[1], *Hiroaki Kojima*[1], *Takuya Narita*[2], *Misato Nihei*[2], *Minoru Kamata*[2], *Shinichi Onaka*[3], *Yoshihiro Fujita*[3], *Takenobu Inoue*[4]

[1]National Institute of Advanced Industrial Science and Technology (AIST), Japan
[2]The University of Tokyo, Japan
[3]NEC Corporation, Japan
[4]Research Institute of National Rehabilitation Center for Persons with Disabilities, Japan
ken.sadohara@aist.go.jp

Abstract

This paper presents a dialogue act classification for a spoken dialogue system that delivers necessary information to elderly subjects with mild dementia. Lexical features have been shown to be effective for classification, but the automatic transcription of spontaneous speech demands expensive language modeling. Therefore, this paper proposes a classifier that does not require language modeling and that uses sub-lexical features instead of lexical features. This classifier operates on sequences of phonemes obtained by a phoneme recognizer and exhaustively analyzes the saliency of all possible sub-sequences using a support vector machine with a string kernel. An empirical study of a dialogue corpus containing elderly speech showed that the sub-lexical classifier was robust against the poor modeling of language and it performed better than a lexical classifier that used hidden Markov models of words.

Index Terms: dialogue acts, support vector machines, string kernels, spontaneous speech, elderly speech, dementia

1. Introduction

This paper presents an information support system for elderly subjects with cognitive disabilities. The target users have difficulties maintaining their attention and absorbing new information, so this system tries to maintain conversations with them to deliver information necessary for their independent and autonomous life in a similar way to their caregivers. Thus, this system needs to recognize colloquial speech and to understand the intentions of utterances so that it can respond sufficiently correctly to sustain conversations. The assignment of an utterance with a predefined functional tag that represents the communicative intentions behind the utterance is referred to as dialogue acts (DAs) classification, which is considered to be a useful first step in dialogue processing. This paper proposes a DA classification method for the colloquial utterances made by the elderly to facilitate the production of an appropriate correct response.

Many studies of DA classification have shown that word n-grams are effective features for determining DAs [1, 2, 3, 4]. To obtain the lexical feature, the automatic transcription of colloquial speech is required. However, some difficulties of the speech recognition have been discussed in previous studies of spontaneous speech recognition [5, 6]. Spontaneous speech includes disfluencies (e.g., filled pauses, repairs, hesitations, repetitions, false starts, or partial words) [7], pronunciation variation [8, 9], and speaking rate variation [10, 9]. For the colloquial speech considered in this paper, its casual style of speech, the speech characteristics of the elderly subjects and the noisy room environment create additional difficulties in terms of the acoustics and language modeling. Among these difficulties, this paper focuses on the difficulty of language modeling.

As pointed out before in many studies, individuals differ not only in their acoustics but also in their lexical patterns. The difference is particularly great in spontaneous speech, so speaker-dependent language modeling has been considered a potential approach to cope with the variation. For example, the quantity of disfluencies varies depending on the speaker, so different models of different classes of speakers are effective for removing disfluencies [7]. Disfluency removal is useful because disfluencies cause problems during subsequent higher-level natural language processing such as DA classification. Another study [8, 9] showed that the lexical pattern used during lecture speech is quite variable among speakers, so language model adaptation to a specific speaker is effective for lecture speech recognition. This can be achieved provided a relatively long speech is available for each lecturer. Unfortunately, the cost of speaker-dependent language modeling is prohibitive in our application because it is difficult to obtain sufficient data to build speaker-dependent language models.

The limitation of the lexicon itself has also been noted. During spontaneous speech, the actual pronunciation of a word can vary greatly from its canonical pronunciation because of sloppy pronunciation, word contractions, or co-articulation between words. To address this variation, a previous study [11] proposed a data-driven dictionary adaptation that adds new entries for words that correspond to the actual pronunciations appearing in given corpora that are obtained using a phoneme recognizer. Another study [9, 12] also found that the use of multiple surface forms for each word baseform is effective for reducing the word error rate during the recognition of spontaneous Japanese speech. In the Japanese language, the different surface forms can be represented as different words, which ensures that they are faithful to the actual pronunciation. Thus, these words can be included as different baseforms in a dictionary. The existence of different representations of a single morpheme can have a harmful effect on DA classification, so it is necessary to normalize the recognized text by replacing the different representations with the corresponding baseform. Unfortunately, the normalization process is not straightforward, unlike word stemming.

SLPAT 2013, 4th Workshop on Speech and Language Processing for Assistive Technologies, pages 93–98,
Grenoble, France, 21–22 August, 2013.

Elaborate language modeling is required to transcribe spontaneous speech faithfully but faithful transcription without normalization is not necessarily useful for our immediate goal of DA classification. To explore the utilization of lexical features in a more cost-efficient manner, this paper proposes a sub-lexical DA classifier that does not require language modeling and that operates on the sequences of phonemes obtained using a phoneme recognizer. The central hypothesis of this study is that if word n-grams are effective indicators for determining DAs, then sub-sequences of phonemes, which are fragments of words in a sense, should also be effective indicators. If this hypothesis is true, then even when effective language modeling is impossible and some salient words are misrecognized, it is expected that their fragments should be preserved, so a more robust form of DA classification based on fragments is possible. Furthermore, the use of phonemes facilitates the analysis of the saliency of the fragments based on the actual pronunciation while considering the patterns of misrecognition for each speaker. Other features such as prosodic features have been investigated [13] to compensate for inherently useful but unreliable and costly lexical features in colloquial speech, but this paper investigates the utilization of lexical features in a more robust and computationally inexpensive manner.

This paper is organized as follows. After describing our information support system and the DAs used in our elderly speech corpus in the next section, Section 3 presents the sub-lexical DA classifier. Section 4 presents an empirical study of the effectiveness of the classifier.

2. DAs used by our assistive system

People with mild dementia, who exhibit memory impairment, disorientation, and an impaired executive function, may use assistive devices [14, 15, 16] to compensate for their problems with absorbing or retaining new information, which have been shown to be effective in their independent and autonomous life. Our information support system is another general-purpose assistive device that was designed to provide information about schedules, times, or dates during conversations [17, 18]. The target users have difficulties maintaining their attention and absorbing information, so the system tries to maintain a conversation with a user based on the following protocol: (1) *attention-seeking* captures the user's attention, which is diminished by dementia; (2) *pre-sequence* prepares the user's mind for absorbing new information; (3) *distributing information* delivers the necessary information; and (4) *end of interaction* closes the conversation. During each stage of the conversation, the system can ask whether the user is following the conversation and can go back to a previous stage if necessary.

To facilitate the computational modeling of the transition of dialogue states, we designed the 12 dialogue acts (DAs) described in Table 1. DAs, which are representations of the communicative intention of each utterance, have been considered integral to the understanding and production of natural dialogue, and they are useful for various forms of speech and language processing, such as speech retrieval, summarization, resolution of ambiguous communication, or the improvement of speech recognition. This paper defines the specific set of DAs used by our application, although efforts to develop domain-independent sets of DAs exist such as DAMSL [19]. Each user utterance is classified as one of the 12 DAs and the system produces an appropriate response based on the classification.

Thanks to the cooperation of 20 single people who were living in nursing homes, we built a dialogue corpus between the system and the users. The details of the participants are as follows: 3 were male and the other 17 were female, the average age was 82.9 ± 7.2 (ranging from 67 to 97), and the average MMSE score [20] was 21.4 ± 5.8 (from 9 to 30). In total, 7,123 utterances were transcribed and annotated, of which 4,080 were user utterances. The total length of user utterances was about 115 hours and the average length of them is about 1.7 ± 1.6 seconds (from 0.2 seconds to 14.8 seconds). The DAs were annotated by two labelers and the inter-labeler agreement was 81.9%, while κ was 0.782.

Table 1: *Dialogue acts and their frequency of occurrence (percentages). The inter-labeler agreement was* 81.9% *and* $\kappa = 0.782$.

Tag	Example	%
Question	What did you eat for dinner?	0.2
Confirmation	Can you understand?	8.2
Request Action	Would you like to go to the bathroom?	4.3
Request Attention	May I ask a question?	15.1
Request Repeat	Pardon?	2.1
Affirmative Answer	Yes, I can.	26.9
Negative Answer	No, I can't.	0.2
Statement	I ate fish.	60.0
Greeting	How are you?	15.1
Affirmative Backchannel	Sure it is.	19.9
Negative Backchannel	Really?	0.2
Other	`Laughter`, `Filler`	5.4

3. Classification of DAs

The automatic classification of DAs comprises two important components: features and modeling methods. The features investigated previously used various types of knowledge, e.g., lexical [21, 1, 2, 3, 22, 4, 23], syntactic [22, 24], prosodic [13, 1, 3, 22, 23], and discourse structural [25, 1]. In this study, sub-lexical features, i.e., sequences of phonemes, were considered together with the DA of the preceding utterance as contextual knowledge. To examine the effectiveness of the sub-lexical feature, typical lexical features, i.e., word n-grams, were also considered together with the contextual knowledge.

These features are used by various modeling methods, e.g., decision trees [13], transformation-based learning [26], hidden Markov models (HMMs) [1], maximum entropy models [22], conditional random fields [27], and support vector machines (SVMs) [3, 4]. To facilitate an exhaustive analysis of all the sub-sequences of phonemes, an SVM with a string kernel based on phonemes was used in this study. Before describing the sub-lexical classifier, we describe a typical classifier based on HMMs of words using a simpler formalization that was obtained by restricting the formalization in [1] to our problem.

3.1. Lexical DA classifiers with HMMs

In a previous study [1], based on the assumption that each observation E_i is emitted from an unobservable DA U_i and the prior distribution of U is Markovian, the optimal sequences U^* of DAs were obtained as follows:

$$U^* = \operatorname*{argmax}_U \prod_{i=1}^{n} P(U_i|U_{i-1})P(E|U_i). \quad (1)$$

In our application, the preceding DA U_{i-1} is observable because the corresponding utterance is given by the system. Therefore, given an utterance of the system with a DA U_R, it

is sufficient to maximize the following equation to obtain the optimal DA U^* of the subsequent utterances E of users,

$$U^* = \operatorname*{argmax}_U P(U|U_R)P(E|U). \quad (2)$$

When the observation E is a text, i.e., a sequence $W_1, \ldots, W_n$ of words and W_j is i.i.d.,

$$U^* = \operatorname*{argmax}_U P(U|U_R)\prod_{j=1}^{n} P(W_j|U). \quad (3)$$

When the observation E is a speech signal A represented in spectral features and is conditioned on the N-best texts $W^{(1)}, \ldots, W^{(n)}$ hypothesized by a speech recognizer, U^* is obtained as follows.

$$U^* = \operatorname*{argmax}_U P(U|U_R)P(A|U) \quad (4)$$

$$= \operatorname*{argmax}_U P(U|U_R)\sum_{n}^{N} P(A|U, W^{(n)})P(W^{(n)}|U) \quad (5)$$

$$= \operatorname*{argmax}_U P(U|U_R)\sum_{n}^{N} P(A|W^{(n)})P(W^{(n)}|U), \quad (6)$$

where the last equality holds under the assumption that $P(A)$ depends only on the words $W^{(n)}$, although this is not true in general because U affects the pronunciation of $W^{(n)}$. Although $P(A|W^{(n)})$ can be computed based on the acoustic likelihood of the speech recognizer, it tends to be a very small value. To avoid underflow, the maximization is computed using the maximum acoustic likelihood $M = \max_n P(A|W^{(n)})$ as follows.

$$U^* = \operatorname*{argmax}_U \frac{P(U|U_R)}{M}\sum_{n}^{N} P(A|W^{(n)})P(W^{(n)}|U) \quad (7)$$

$$= \operatorname*{argmax}_U P(U|U_R)\sum_{n}^{N} \exp(L(n)) \quad (8)$$

$$L(n) = \ln(P(A|W^{(n)})) - \ln(M) + \ln(P(W^{(n)}|U)) \quad (9)$$

In the rest of the paper, N is set as 10.

3.2. Sub-lexical DA classifiers with SVMs

The DA classifier presented in this paper operates on sequences of phonemes obtained using a phoneme recognizer. For any sequence of phonemes, the DA classifier analyzes whether any noncontiguous sub-sequence is salient to the discrimination of a particular class. The analysis is performed using an SVM [28, 29] by computing the optimal hyperplane that separates positive samples from negative samples in the feature space spanned by all possible sub-sequences of phonemes. Although the dimension of the feature space is exponential in terms of the length of sub-sequences, the analysis can be performed efficiently using string kernels [30].

A string kernel modified for the analysis of sequences of phonemes was investigated in a previous study [31] for topic segmentation. Given two sequences of phonemes, s and t, the string kernel computes the similarity between s and t efficiently in $O(p|s||t|)$, where p is the maximum length of sub-sequences. The similarity is computed based on the number of occurrences of any non-contiguous sub-sequence, where the occurrence count is decayed according to λ^g $(0 \leq \lambda \leq 1)$ for the number g of gaps in each sub-sequence. In the occurrence count, a soft-matching method is used between phonemes, which assigns 1 if they are identical and a value between 0 and 1 otherwise. Based on this definition of the similarity, the classifier is expected to be robust against insertion, deletion, and substitution errors of phonemes.

The kernel function is normalized and extended to consider the contextual DA of the preceding utterance as follows:

$$K_\ell(s,t) \stackrel{\text{def}}{=} \delta_{c(s),c(t)}\frac{\kappa_\ell(s,t)}{\sqrt{\kappa_\ell(s,s)}\sqrt{\kappa_\ell(t,t)}} \quad (10)$$

where κ_ℓ is the kernel function with the length ℓ of sub-sequences, as defined in [31], $c(s)$ is the DA of the preceding utterance of s, and $\delta_{c(s),c(t)} = 1$ if $c(s) = c(t)$, but 0 otherwise.

The string kernel is extended further to consider the weighted contributions of different lengths ℓ of sub-sequences as follows.

$$K^{\leq p}(s,t) \stackrel{\text{def}}{=} \sum_{\ell=1}^{p}\gamma_\ell K_\ell(s,t). \quad (11)$$

We can see that the kernel function satisfies the Mercer condition [29] required for SVM optimization because it is actually the inner-product of the feature space spanned by the sub-sequences of phonemes, although it is computed implicitly. In the rest of the paper, we assume $\gamma_k = 1$, $\lambda = 0.7$, and $p = 4$.

Using the string kernel, an SVM is trained to discriminate a particular class. Because an SVM is fundamentally a binary classifier, various methods have been considered for extending multiple SVMs to a multi-class classifier. In this study, the simple one-versus-the-rest approach is adopted, i.e., for each DA U, an SVM f_U is trained that discriminates U from the other DAs, and the optimal DA U^* for any sequence s of phonemes is obtained as $U^* = \operatorname{argmax}_U f_U(s)$.

In the same way as the previous section, the N-best hypotheses of a phoneme recognizer are considered as follows

$$U^* = \operatorname*{argmax}_U \sum_{n}^{N} P(A|s_n)f_U(s_n) \quad (12)$$

$$= \operatorname*{argmax}_U \sum_{n}^{N} \exp(\ln(P(A|s_n)) - \ln(M))f_U(s_n) \quad (13)$$

where M is the maximum acoustic likelihood $M = \max_n P(A|s_n)$.

4. Empirical study

The aim of the empirical study was to verify the effectiveness of the sub-lexical DA classifier. In the experiments described below, several classifiers were trained for 4,080 user utterances and DAs of the utterances were predicted. For each user, the utterances from the first several days were used for training while the rest were used for testing. As a result, 1,920 utterances were used for training and 2,160 utterances were used for testing. Because the training data contain a small number of samples with the following four tags: *Request Action*, *Request Attention*, *Negative Answer*, and *Negative Backchannel*, for the remaining eight tags, eight classifiers were trained and tested.

The transcriptions were obtained manually and automatically, where the latter was conducted using a large-vocabulary continuous speech recognizer, Julius [32]. Its dictionary and

Table 2: *Accuracy and F-measure of a lexical classifier using HMMs and a sub-lexical classifier using SVMs for manual (MT) and automatic (ASR) transcription.*

	word-HMM		phone-SVM	
	Accuracy	F1	Accuracy	F1
MT	0.800	0.521	0.817	0.624
ASR	0.758	0.521	0.789	0.563

word trigram model were built from the training data. The number of entries in the dictionary was 1,008, the test set perplexity of the language model was 17.97, and the OOV rate was 6.37%. Its acoustic model was a gender-independent PTM triphone model of elderly speech [33] distributed by CSRC [34], which was adapted to each speaker using the MLLR method [35]. The word error rate in the test data were 58%. Phonetic transcriptions were obtained using the same decoder, except a phoneme trigram model was trained and used where the phoneme error rate was 46%.

During the training of classifiers for manual transcriptions, transcribed texts or sequences of phonemes converted from the texts were used. On the other hand, during the training of classifiers for automatic transcriptions, the five best hypotheses of the output of the speech recognizers for the training data were used as well as the texts or the sequences of phonemes obtained from the manual transcriptions. For the parameters of SVMs, we used $\lambda = 0.7$, $\gamma = 1.0$, $C = 10.0$, and p was set as $p = 4$ because the average lengths of the words were 4.8 phonemes.

During the evalution of the classifiers, texts or sequences of phonemes obtained form manual or automatic transcriptions for the test data were used. Especially for automatic transcriptions, the 10 best hypotheses of the output of the speech recognizers with the acoustic likelihood of them are used.

Table 2 summarizes the results of the experiments where the accuracy indicates the ratio of correct predictions and F1 indicates the average harmonic mean of the precision and recall, i.e., the F-measure averaged across DAs. We can see that the phone-SVM, i.e., the sub-lexical DA classifier with SVMs, performed better than the word-HMM, i.e., the lexical DA classifier with HMMs in both the manual and automatic transcriptions. The difference in the manual transcription was significant ($p < 0.05$) according to McNemar's test and the difference in the automatic transcription was also significant ($p < 0.01$). In the following section, these results are discussed in more detail.

4.1. Robustness of the sub-lexical DA classifier

Figure 1 depicts the accuracy of the two classifiers during manual and automatic transcription for the convenience of the reader. It also shows the result of the word-HMM for another automatic transcription, which was obtained using a cheating language model built from all of the data including the test data. The number of entries in the dictionary for the cheating model was 1,587, the test set perplexity was 4.70, and the word error rate was 32.6%. There was no significant difference between ASR and ASR(CHEAT), which suggests that the accuracy would not be improved even if a better language model could be obtained from a larger amount of training data. Thus, it is unlikely that that the accuracy of the word-HMM would improve without an elaborate language modeling and text normalization for spontaneous speech. Furthermore, the accuracy of word-HMM would become worse as the mismatches between

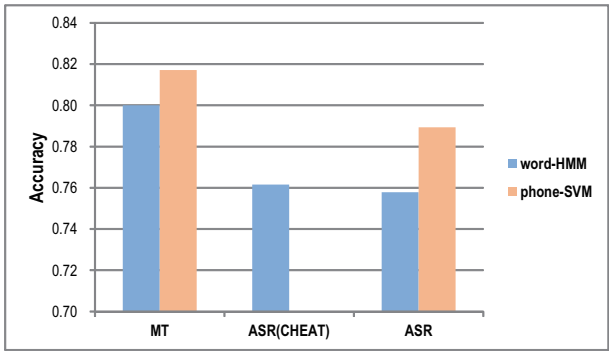

Figure 1: *The lexical classifier vs. the sub-lexical classifier.*

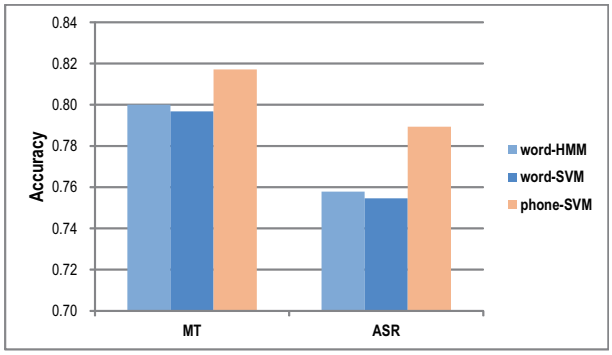

Figure 2: *Lexical features vs. sub-lexical features.*

the language model and the corpus increased. On the other hand, the accuracy of the phone-SVM would not decline because the sub-lexical classifier does not depend on the language model.

4.2. Effectiveness of the sub-lexical features

Figure 2 shows another result of a SVM (word-SVM) using the bag-of-words feature that operates on the feature spaces spanned by the frequency of each word appearing in the training data. The performance of word-SVM was worse than that of phone-SVM, and its performance was not significantly different from that of word-HMM. This suggests that the superior performance of phone-SVM was not attributable to the SVM-based modeling method, but instead it was due to the sub-lexical features.

In particular, the difference between the word-SVM and phone-SVM results with manual transcription was due only to the difference between the lexical feature and the sub-lexical feature. A possible explanation for this difference is that the existence of multiple surface forms of a baseform degraded the performance of word-SVM. Using phone-SVM, however, the common fragments of the different surface forms allowed us to capture salient properties for DA classification.

Furthermore, the difference between word-SVM and phone-SVM in ASR was larger than in MT. The results for both classifiers were obtained using the same decoder and the same acoustic model, so the bigger difference may have been because some salient word features were lost by the poor modeling of language, whereas some fragments of the salient features were still preserved with phone-SVM.

5. Conclusion and future work

This paper proposed a sub-lexical DA classifier for use as a dialogue management module in a spoken dialogue system that provides necessary information to elderly users with cognitive disabilities. To avoid costly and difficult language modeling when transcribing the colloquial utterances of the elderly users in a faithful manner, the classifier determines the DAs based on the sequences of phonemes obtained using a phoneme recognizer. Instead of searching for salient word features used by many lexical classifiers, the sub-lexical classifier searches for salient sub-sequences of phonemes while considering possible misrecognitions, i.e., insertion, deletion, and substitution errors. To search the space spanned by the exponentially many features efficiently, the proposed method uses an SVM with a string kernel based on sequences of phonemes. An empirical study was conducted using a dialogue speech corpus collected from elderly subjects with mild dementia. The sub-lexical classifier was found to be robust against the poor modeling of language, while it performed better than a lexical classifier using HMMs.

These results are now limited to our small and simple dialogue corpus, which contains only four thousands short (1.7 seconds on average) user utterances, and only 8 of 12 DA tags have been tested. The effectiveness of the sub-lexical DA classifier should be investigated for larger and well-studied corpora. The DA classification itself does not essentially need any faithful transcription of spontaneous speech. We believe the analysis of the frequency of sub-sequences of phonemes instead of the frequency of words is effective especially when the faithful transcription is hard to obtain. Furthermore, the robust and cost-efficient use of the sub-lexical feature without language modeling could be more effective when it is used together with other non-lexical features, e.g., prosody.

6. Acknowledgements

We would like to thank Seikatsu Kagaku Un-Ei Co. Ltd. This work was supported partially by the Japan Science and Technology Agency, JST, as part of the Strategic Promotion of Innovative Research and Development Program.

7. References

[1] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. V. Ess-Dykema, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational Linguistics*, vol. 26, no. 3, pp. 1–34, 2000.

[2] N. Webb, M. Hepple, and Y. Wilks, "Dialogue act classification based on intra-utterance features," in *Proceedings of the AAAI Workshop on Spoken Language Understanding*, 2005.

[3] D. Surendran and G. A. Levow, "Dialog act tagging with support vector machines and hidden Markov models," in *Proceedings of the Annual Conference of the International Speech Communication Association*, 2006.

[4] B. Gambäck, F. Olsson, and O. Täckström, "Active learning for dialogue act classification," in *Proceedings of the Annual Conference of the International Speech Communication Association*, 2011.

[5] S. Furui, "Spontaneous speech recognition and summarization," in *Proceedings of the Baltic Conference on Human Language Technologies*, pp. 39–50, 2005.

[6] E. Shriberg, "Spontaneous speech: how people really talk and why engineers should care," in *Proceedings of the European Conference on Speech Communication and Technology*, 2005.

[7] M. Honal and T. Schultz, "Automatic disfluency removal on recognized spontaneous speech - rapid adaptation to speaker-dependent disfluencies," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2005.

[8] H. Nanjo and T. Kawahara, "Unsupervised language model adaptation for lecture speech recognition," in *Proceedings of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003.

[9] H. Nanjo and T. Kawahara, "Language model and speaking rate adaptation for spontaneous presentation speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 4, pp. 391–400, 2004.

[10] H. Nanjo and T. Kawahara, "Speaking-rate dependent decoding and adaptation for spontaneous lecture speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 725–728, 2002.

[11] T. Sloboda and A. Waibel, "Dictionary learning for spontaneous speech recognition," in *Proceedings of the International Conference on Spoken Language Processing*, pp. 2328–2331, 1996.

[12] Y. Akita and T. Kawahara, "Statistical transformation of language and pronunciation models for spontaneous speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp.1539–1549, 2010.

[13] E. Shriberg, A. Stolcke, D. Jurafsky, N. Coccaro, M. Meteer, R. Bates, P. Taylor, K. Ries, R. Martin, and C. V. Ess-Dykema, "Can prosody aid the automatic classification of dialog acts in conversational speech?," in *Language and speech*, vol. 41, no. 3-4, pp. 443–492, 1998.

[14] T. Inoue, R. Ishiwata, R. Suzuki, T. Narita, M. Kamata, M. Shino, and M. Yaoita, "Development by a field-based method of a daily-plan indicator for persons with dementia," in *Assistive Technology from Adapted Equipment to Inclusive Environments: AATE 2009*, pp. 364–368, IOS Press, 2009.

[15] "Cognitive aids", Online: http://www.abilia.org.uk, accessed on 24 Mar 2013.

[16] "Automatic pill dispenser", Online: http://www.pivotell.co.uk/, accessed on 24 Mar 2013.

[17] M. Nihei, T. Narita, R. Ishiwata, M. Onoda, M. Shino, H. Kojima, S. Ohnaka, Y. Fujita, M. Kamata, and T. Inoue, "Development of an interactive information support system for persons with dementia," in *Proceedings of the International Technology and Persons with Disabilities Conference*, 2011.

[18] T. Inoue, M. Nihei, T. Narita, M. Onoda, R. Ishiwata, I. Mamiya, M. Shino, H. Kojima, S. Ohnaka, Y. Fujita, and M. Kamata, "Field-based development of an information support robot for persons with dementia," *Technology and Disability*, vol. 24, no. 4, pp.263–271, 2012.

[19] M. Core and J. Allen, "Dialogs with the DAMSL annotation scheme," in *Working Notes of the AAAI Fall Symposium on Communicative Action in Humans and Machines*, pp. 28–35, 1997.

[20] M. F. Folstein, S. E. Folstein, and P. R. McHugh, *Mini-mental state: a practical method for grading the cognitive state of patients for the clinician*, Pergamon Press, 1975.

[21] D. Jurafsky, E. Shriberg, B. Fox, and T. Curl, "Lexical, prosodic, and syntactic cues for dialog acts," in *Proceedings of the Workshop on Discourse Relations and Discourse Markers*, pp. 114–120, 1998.

[22] V. K. R. Sridhar, S. Bangalore, and S. Narayanan, "Combining lexical, syntactic and prosodic cues for improved online dialog act tagging," *Computer Speech & Language*, vol. 23, no. 4, pp. 407–422, 2009.

[23] N. G. Ward and A. Vega, "Towards empirical dialog-state modeling and its use in language modeling," in *Proceedings of the Annual Conference of the International Speech Communication Association*, 2012.

[24] J. O'Shea, Z. Bandar, and K. Crockett, "A multi-classifier approach to dialogue act classification using function words," *Transactions on Computational Collective Intelligence VII*, pp. 119–143, 2012.

[25] D. Jurafsky, R. Bates, N. Coccaro, R. Martin, M. Meteer, K. Ries, E. Shriberg, A. Stolcke, P. Taylor, and C. V. Ess-Dykema, "Automatic detection of discourse structure for speech recognition and understanding, " in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 88–95, 1997.

[26] K. Samuel, Sandra. Carberry, and K. Vijay-Shanker, "Dialogue act tagging with transformation-based learning," in *Proceedings of the International Conference on Computational Linguistics*, pp. 1150–1156, 1998.

[27] S. Quarteroni, A. V. Ivanov, and G. Riccardi, "Simultaneous dialog act segmentation and classification from human-human spoken conversations," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5596–5599, 2011.

[28] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.

[29] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*. Cambridge Press, 2000.

[30] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins, "Text classification using string kernels," *Journal of Machine Learning Research*, vol. 2, pp. 419–444, 2002.

[31] K. Sadohara, "Kernel topic segmentation for informal multi-party meetings and performance degradation caused by insufficient lexicon," in *Proceedings of the IEEE Workshop on Spoken Language Technology*, pp. 430–435, 2010.

[32] A. Lee, T. Kawahara, and K. Shikano, "Julius — an open source real-time large vocabulary recognition engine," in *Proceedings of the European Conference on Speech Communication and Technology*, pp.1691–1694, 2001.

[33] A. Baba, S. Yoshizawa, M. Yamada, A. Lee, and K. Shikano, "Elderly acoustic model for large vocabulary continuous speech recognition," in *Proceedings of the European Conference on Speech Communication and Technology*, pp.1657–1660, 2001.

[34] A. Lee, T. Kawahara, K. Takeda, M. Mimura, A. Yamada, A. Ito, K. Itou, and K. Shikano, "Continuous speech recognition consortium: an open repository for CSR tools and models," in *Proceedings of the International Conference on Language Resources and Evaluation*, pp. 1438–1441, 2002.

[35] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.