The SpeDial Datasets:
Datasets for Spoken Dialogue Systems Analytics

José Lopes[1], Arodami Chorianopoulou[2]
Elisavet Palogiannidi[2], Helena Moniz[3], Alberto Abad[3]
Katerina Louka[4], Elias Iosif [5,6], Alexandros Potamianos[5,6]
[1] KTH Speech, Music and Hearing, Stockholm, Sweden
[2] School of ECE, Technical University of Crete, Greece
[3] INESC-ID, Lisboa, Portugal
[4] Voiceweb S.A., Greece
[5]"Athena" Research and Innovation Center, Greece
[6]School of ECE, National Technical University of Athens, Greece
jdlopes@kth.se

Abstract

The SpeDial consortium is sharing two datasets that were used during the SpeDial project. By sharing them with the community we are providing a resource to reduce the duration of cycle of development of new Spoken Dialogue Systems (SDSs). The datasets include audios and several manual annotations, i.e., miscommunication, anger, satisfaction, repetition, gender and task success. The datasets were created with data from real users and cover two different languages: English and Greek. Detectors for miscommunication, anger and gender were trained for both systems. The detectors were particularly accurate in tasks where humans have high annotator agreement such as miscommunication and gender. As expected due to the subjectivity of the task, the anger detector had a less satisfactory performance. Nevertheless, we proved that the automatic detection of situations that can lead to problems in SDSs is possible and can be a promising direction to reduce the duration of SDS's development cycle.

Keywords: Spoken Dialogue Systems, Multi-lingual Data, Emotions, Sentiment Analysis

1. Introduction

The speech services industry has been growing both for telephony applications and, recently, also for smartphones (e.g., Siri). Despite recent progress in Spoken Dialogue System (SDS) technologies the development cycle of speech services still requires significant effort, expertise and time. Developers often have to wait until the system goes live to detect potential sources of errors that could not be predicted in the design phase. The SpeDial consortium (www.spedial.eu) worked to create a semi-automated process for spoken dialogue service development and speech service enhancement of deployed services. Our main goal was to provide a tool where incoming speech for the system is semi-automatically analyzed (human-in-the-loop) to identify sources of problems in the dialogue. The first step towards this goal was to build tools that automatically identify problematic dialogue situations or as we will call hereafter miscommunications.

The automatic detection of miscommunications in SDSs has been extensively investigated in the literature (Walker et al., 2000; Swerts et al., 2000; Schmitt et al., 2010; Paek and Horvitz, 2004). This problem is vital in the development cycle of speech services. However, very little data is publicly available to perform research on this topic. One of the exceptions is (Swerts et al., 2000), but even in this case the dataset does not contain interactions with real users or annotations. In (Schmitt et al., 2012) a dataset collected with real users is described. This data was annotated for interaction quality (Schmitt et al., 2011), emotions and contains also a variety of automatically extracted features. This dataset was built with dialogues from CMU Let's Go (Raux et al., 2005) system from 2006, which performance and architecture are substantially different than the current Let's Go system. In addition, the interaction quality might not be the most suitable measure for identifying problematic dialogue situations, namely if severe problems occur in the very first exchange of the interaction. Recently, more Let's Go data was made available for the Spoken Dialogue Challenge (Black et al., 2010). Although part of the dataset was transcribed, no other manual annotations were provided with it. Therefore, the creation of new resources with more recent data and more annotations was need for SpeDial. The multi-lingual nature of the project also required datasets in languages other than English.

Thus, we are making two datasets publicly available: 1) the Let's Go 2014 collect from Let's Go collected during 2014 and 2) The Movie Ticketing (MT) dataset collected with the MT system developed by Voicweb S.A.. Both datasets include audios for every user turn with respective manual transcriptions, together with gender, task success, anger, satisfaction and miscommunication annotations in the SPDXml format (SpeDial, 2015b).

The shared datasets were evaluated on relevant to SDSs tasks like anger, miscommunication detection and gender recognition. The promising results confirm the usefulness of the datasets that we are releasing for future research in SDS Analytics.

The following section will described the two datasets shared. In Section 3. the annotation scheme used will be explained. Section 4. describes the experimental procedure and the results achieved by the detectors developed with these datasets. The paper closes with final remarks in Section 5. and future work in Section 6..

2. Datasets

2.1. Let's Go

The Let's Go dataset[1] is composed of dialogues between real users and the Let's Go system that provides bus schedule information for public buses in the city of Pittsburgh. Initially 105 dialogues were randomly selected from all the dialogues collected during the first half of 2014. Dialogues shorter than 4 turns were then excluded from the dataset since this is the minimum number of turns needed to get schedule information. The final 85 dialogues correspond to 1449 valid user turns (average 17.1 turns per dialogue).

The dataset was annotated following the scheme that will be described in Section 3. for Let's Go data. The preprocessing of the logs allowed the extraction of live features from several system modules. Features derived from transcriptions and their parsing, such as Word Error Rate and Concept Error Rate were also included in the data that were are releasing.

2.2. Movie ticketing

The movie ticketing dataset[2] consists of 200 dialogues in Greek collected through a call center service for retrieving information about movies, e.g. show times information and ticket booking. The dataset includes two data types for each dialogue: audio recordings and the corresponding transcriptions. The annotation of dialogues was performed by an expert annotator, while the selected dialogues were balanced with respect to three factors: (i) gender of caller, (ii) call success, (iii) emotional content according to the annotation scheme that will be described in Section 3..

3. Annotation scheme

As the existing datasets lacked in some annotation that we thought might be useful in the context of the SpeDial project, in this section we will describe in detail the annotation scheme adopted.

The first annotation step was to manually transcribe the user utterances in both datasets. The system prompts in the MT dataset were also transcribed since only the audio files from user and system turns were available.

To perform the miscommunication annotation on Let's Go data, annotators were given snippets of four turns, two system and two user turns as shown in Table 1. The speech recognition output and transcription were presented to the annotator when performing the task, with the respective parsing when available. The annotators had access to the audio from the utterances when they were annotating.

There were several reasons to provide only four turns to the annotator. Initially some of the data was annotated using a crowd-sourcing platform which required the number of slots filled per task to be fixed. Therefore we picked the minimum number of turns that we thought sufficient to make a decision whether turn S3 would be an appropriate system answer or not. We assumed that if humans could perform this task with this amount of data, it should be straightforward to develop an algorithm that could perform the same task with the same amount of data automatically. In addition, by having a limited number of turns in the analysis, we would reduce the computational complexity of the problem, thus improving efficiency. Label 0 was used when system answer was not considered problematic, 1 when the system answer was problematic and 2 when the annotator could not decide from the context whether the system answer was problematic or not. An example of a snippet provided for annotation is shown in Table 1.

During the MT miscommunication annotation the annotator could see the whole dialogue, instead for the four turn snippet provided for the Let's Go annotation.

As mentioned before, the presence of anger, the user satisfaction and the presence of repeated content in the utterances could be indicators that a miscommunication occurred.

In Let's Go 1 was used when anger was detected and 0 otherwise. The labels used the Movie Ticketing data were discrete scores that lie in the $[1-5]$ interval capturing very angry user utterances (1) to friendly utterances (5). In order to adopt the same scheme across datasets the values $[1-3]$ were mapped into 1 and values in the interval 4 and 5 were mapped into 0. The presence of anger was always signaled by the shouting or use of offensive language. However, there are other ways of user's to express their (dis)satisfaction towards the system. Therefore, Satisfaction was also annotated as a subjective measure of the user experience. As expected, all the subset angry turns are part of dissatisfied turns as well For Let's Go 0 when the user was satisfied and 1 when she was not. In the MT data, the data was annotated in a five point scale from 1 very unsatisfied to 5 very satisfied.

In the MT dataset, 1 was used for user utterances in which repetition was observed and 0 otherwise. In Let's Go 1 was used for complete repetitions (TOTAL), 2 for partial repetitions, that is when all the content provided partially matches another turn in the dialogue (PARTIAL), 3 when there is some content repeated between turns, but there is no complete match between them (MIXED) and 0 when no repetition was observed. The annotation scheme for Let's Go was already same adopted in (Lopes et al., 2015).

While listening to the dialogue the annotators were asked to be aware of gender. As soon as they were confident they would assign the gender label to the whole dialogue.

To annotate task success, the annotators should listen to the whole dialogue and verify if the intention of the user was correctly answered by the system. The label 1 was used for successful dialogues and the 0 for unsuccessful dialogues.

The Let's Go dataset was annotated by two expert annotators. One of them annotated the whole dataset, whereas the other annotated 10% of it. The Cohen's Kappa agreement observed for the two annotators was 0.79 for miscommunication (substantial agreement), 0.38 for anger (fair agreement), 1.0 for task success and 1.0 for gender annotations (perfect agreement). We have computed the agreement between the majority annotation for task success and the estimated task success. The Cohen's kappa found was 0.44, which is seen as fair agreement.

The MT dataset was originally annotated by one expert an-

[1] For data access, please contact `jdlopes@kth.se`.

[2] For data access, please contact `info@voiceweb.eu`.

Annotation	Turn Id	Turn [TRANSCRIPTION, Parse]
	S1	Where would you like to leave from?
	U2	WEST MIFFLIN [WEST MIFFLIN AND, DeparturePlace = MIFFLIN]
NOT PROBLEMATIC	S3	Departing from MIFFLIN. Is this correct?
	U4	SHADY EIGHT [EXCUSE ME, (DeparturePlace = EIGTH, DeparturePlace = SHADY)]

Table 1: Example when label 3 was attribute to turn S3 in Let's Go data.

notator. Two additional annotators labeled a subset of the 60 dialogues from the original dataset for anger. The agreement between annotators found was 58% with 0.4 Kappa value –computed as the average pairwise agreement– according to the Fleiss coefficient, which can be interpreted as a moderate agreement.

Table 2 summarizes the distribution of each of the categories annotated in each dataset.

4. Experimental Procedure and Results

In this section, we briefly present a series of indicative experimental results for a variety of different detectors developed using the datasets previously described.

4.1. Anger detection in the Movie Ticketing dataset

The experimental results for the movie ticketing dataset are briefly presented with respect to two different systems performing speech– and text–based analysis.

4.1.1. Speech-based system.

Here, the goal is to capture the speaker's emotional state using exclusively the speaker's speech signal. Hence, we utilize a set of low-level descriptors (LLDs) in order to describe the emotional content. Such LLDs have been widely used and include prosody (pitch and energy), short-term spectral (Mel Frequency Cepstral Coeficients, MFCCs) and voice quality (Jitter) features (Ververidis et al., 2004). The LLDs were extracted in a fixed window size of 30 ms with a 10 ms frame update and were further exploited via the application of a set of functions, in order to map the speech contours to feature vectors. The following functions (statistics) computed at the utterance-level for each of the LLDs were used for the speech analysis: percentiles, extremes, moments and peaks. A detailed system description is provided in (SpeDial, 2015a).

4.1.2. Text-based system.

The goal is to estimate the emotional content of the transcribed speaker utterances. The affective content of a word w can be characterized in a continuous space consisting of three dimensions, namely, valence, arousal, and dominance. For each dimension, the affective content of w is estimated as a linear combination of its semantic similarities to a set of K seed words and the corresponding affective ratings of seeds (Turney and Littman, 2002). Here, we provide a brief description of the underlying model, while more details can be found in (Palogiannidi et al., 2015).

$$\hat{u}(w) = \lambda_0 + \sum_{i=1}^{K} \lambda_i \; u(t_i) \; S(t_i, w), \qquad (1)$$

where $t_1...t_K$ are the seed words, $u(t_i)$ is the affective rating for seed word t_i with u denoting one of the aforementioned dimensions (i.e., valence or arousal or dominance). λ_i is a trainable weight corresponding to seed t_i. $S(t_i, w)$ stands for a metric of semantic similarity between t_i and w. The model of (1) is based on the assumption that *"semantic similarity can be translated to affective similarity"* (Malandrakis et al., 2013). The $S(\cdot)$ metric can be computed within the framework of (dataset-based) distributional semantic models (DSMs) that rely on the hypothesis that *"similarity of context implies similarity of meaning"* (Harris, 1954). In DSMs, a contextual window of size $2H+1$ words is centered on the word of interest w_i and lexical features are extracted. For every instance of w_i in the (text) dataset the H words left and right of w_i formulate a feature vector. For a given value of H the semantic similarity between two words, w_i and w_j, is computed as the cosine of their feature vectors.

4.1.3. Fusion of speech and text analysis.

The main idea for the fusion of the two systems is motivated by the hypothesis that each system exhibits different types of errors. For example, cases of offensive language may be missed by the speech-based system, while cases of anger are likely to be missed by the text-based one. In an attempt to improve the performance of the speech-based system, we employed a late fusion scheme. Specifically, the posterior probabilities of the two systems were combined in an algebraic scheme, i.e., the mean. The final decision was the class with the maximum mean posterior probability.

4.1.4. Experiments and evaluation results.

The goal is the detection of "angry" vs. "not angry" (i.e., 2-class classification problem) user utterances. The anger annotations were used both for training and evaluation purposes. Specifically, the *friendly* and *neutral* labels were mapped to the "not angry" class, while the *slightly angry*, *angry* and *very angry* labels were mapped to the "angry" class. Both speech- and text-based systems were developed adopting a leave-one-dialogue-out scheme and aiming to the prediction of anger on utterance level. The unweighted average recall (UAR) and the classification accuracy (CA) were used as evaluation metrics. For the speech–based system the used feature set consisted of statistics over the first ten Mel-frequency cepstral coefficients (MFCCs) (Lee et

[3] No distiction was made between different types of repetitions for this dataset.

[4] Two speakers with different gender interacted with the system.

		Datasets	
Annotation Type	Labels	Let's Go	MT
Miscommunication	NOT PROBLEMATIC	0.51	0.61
	PROBLEMATIC	0.42	0.32
	PARTIALLY PROBLEMATIC	0.07	0.07
Anger	ANGRY	0.05	0.17
	NOT ANGRY	0.96	0.75
	NO ANNOTATION	-	0.08
Satisfaction	SATISFIED	0.81	0.44
	NOT SATISFIED	0.19	0.48
	NO ANNOTATION	-	0.08
Repetition	TOTAL	0.11	0.02[3]
	PARTIAL	0.07	-
	MIXED	0.04	-
	NO REPETITION	0.78	0.98
Gender	MALE	0.47	0.48
	FEMALE	0.52	0.52
	MIXED	0.01[4]	-
Task success	SUCCESSFUL	0.35	0.47
	NOT SUCCESSFUL	0.65	0.53

Table 2: Distribution of the data with the respect to the annotations performed.

al., 2004) extracted via OpenSmile (Eyben et al., 2010). In order to reduce the feature vector's dimensionality a forward selection algorithm was applied to the original feature set using the WEKA toolkit. Statistics of dominance scores were used as features for the text–based system. Different classifiers were used regarding each modality, JRip for speech and Random Forest for text.

System	**UAR**	**CA** (%)
Speech	0.67	67
Text	0.61	59
Fusion of speech and text		
Mean of posterior probabilities	0.67	68

Table 3: Movie ticketing dataset: "angry" vs. "not angry" classification.

The results of the affective analysis on the MT dataset are presented in Table 3. All the systems exceed the performance of the majority–based classification regarded as naive baseline (0.5 UAR for binary problems and 59% CA). The speech–based affective system outperforms the text–based system with respect to both evaluation metrics. The best performance, with respect to CA, was obtained by the fusion of the speech– and text–based systems suggesting that the performance of the speech-based system can be (slightly) benefited by the indatasetstion of the text-based analysis. The affective speech analysis was also applied over the Let's Go dataset for the task of anger detection achieving 0.88 UAR. We used the leave-one-dialogue-out technique and two features, namely energy and the first mel-frequency coefficient. The attempts to use the affective text analysis on Let's Go were in vain, since only three utterances in the whole dataset include lexical anger markers.

4.2. Gender detection

A brief description of the gender classification module developed in the context of the SpeDial project is here presented, followed by gender classification results for the Let's Go and Movie Ticketing datasets.

4.2.1. Gender classification module.

The SpeDial gender classification module used in these experiments is a modified version of the frame-level gender classifier based on artificial neural network modelling described in (Meinedo and Trancoso, 2011), which was mainly optimized for media and broadcast news captioning and considered 3 target output classes: *male*, *female* and *child*. Like in (Meinedo and Trancoso, 2011), the SpeDial gender module is based on an artificial neural network model of the multi-layer perceptron (MLP) type with 9 input context frames of 26 coefficients (12th order PLP coefficients (Hermansky et al., 1992) plus deltas), two hidden layers with 350 sigmoidal units each and the appropriate number of softmax output units (one per target class). In this case, only *male* and *female* output classes were considered. Moreover, the MLP model has been re-trained using 70 hours of multi-lingual telephone speech data corresponding to a sub-set of the SpeechDat datasets. In particular, the training dataset is composed of speech utterances from the Portuguese version of SpeechDat(II), and additional speech from the English, Spanish, Italian, French and German versions of SpeechDat(M). The strategy followed to classify a speech segment is to compute the average posterior probability for the 2 target classes and select the one with the highest probability. Given the characteristics of typical dialogue turns, which contain a considerable amount of silence, a frame-level non-speech removal stage has been indatasetsted to avoid decision making over the whole audio segment. Thus, this component performs frame-level classification of each speech utterance exploiting an MLP trained with Modulation-Filtered Spectrogram (MSG) (Kingsbury et al., 1998) features (28 static) corresponding to approximately 50 hours of downsampled TV Broadcast News data and 41 hours of varied music and sound effects at 8kHz sampling rate. Hence, the SpeDial gender module computes gender average posterior probabilities for each sentence based only on the frames previously labelled as speech.

Alternatively, we have also explored novel approaches for gender classification based on segment-level features (Eyben et al., 2010) in combination with neural network modelling and i-vector (Dehak et al., 2009) based classifiers. However, these systems are not reported in this work, since none of them provided significant improvements with respect to the frame-level MLP classifier.

4.2.2. Experiments and evaluation results.

Gender classification has been conducted separately for each speaker turn of the Let's Go and Movie Ticketing datasets in order to obtain automatic turn gender classification. Also, the complete speaker side of both datasets have been processed to obtain per dialogue results. In the case of turn-level classification, gender accuracies obtained are 79.6% and 89.8% in the Let's Go and Movie Ticketing datasets respectively, when considering all the speaker turns. Notice that in both datasets not only most of the turns are extremely short, but there is also a significant number of turns without speech content. In particular, around 12% of the Let's Go turns do not contain useful speech, which affects negatively the performance of the classifiers. When considering only the turns annotated as containing speech, the performance increases up to 84.9% in the Let's Go dataset (speech content annotation is not available in the Movie dataset). Regarding dialogue level evaluation, a 91.7% and 98% classification accuracy is attained in the Let's Go dataset and Movie Ticketing datasets, respectively. Overall, the module seems to perform consistently in both datasets, independently of the language (notice that Greek data was not included in the training set). We consider these results quite satisfactory, particularly considering the reduced amount of actual speech in most of the speaker turns.

4.3. Miscommunication detection

We performed miscommunication detection for both datasets using a similar supervised approach to the one described in (Meena et al., 2015).

4.3.1. Features

The feature set used is highly dependent on the information that could be extracted from the system logs for each dataset. The best case scenario, where all features are available, includes the following features.

Speech Recognition: the best hypothesis, confidence score and the number of words.

Language Understanding (LU): user dialogue act (the best parse hypothesis), the best parse hypothesis obtained from the manual transcription, number of concepts in both user dialogue act and best parse transcription, concept error rate and correctly transferred concepts.

Language Generation (LG): system dialogue act, number of concepts in system act, system prompt and number of words in the prompt.

Features derived from transcriptions: manual transcriptions, number of words in the transcription, word error rate, correctly transferred words and fraction of words observed both in the transcription and the ASR hypothesis.

Discourse features: fraction of turns completed up to the decision point, fraction of new words/concepts[5] used in successive speaker utterances; cosine similarity between consecutive turns and the number of repeated concepts in consecutive turns; whether the user response to a system explicit confirmation request was a 'no' (at the semantic level); various features indicating the number of slot values mentioned in previous turns that were given a new value either by the speaker or the system.

4.3.2. Method

Given that the main purpose of Spoken Dialogue Systems analytics is to provide tools to detect problems off-line, we will report the results of the *offline model*. This model was trained with all possible features available (including those derived from manual annotations) from each four turn snippet taken into account. The method developed was only applied for turns annotated either as PROBLEMATIC or NON-PROBLEMATIC. Given the skew of the distribution and to obtain comparable results to those reported in (Meena et al., 2015), we report the results in terms of UAR (Schmitt et al., 2011), for this the baseline majority will be 0.5 for all the datasets regardless of the distribution of the data. Several machine learning algorithms were explored both in Weka (Hall et al., 2009) and sklearn toolkit (Pedregosa et al., 2011). For validation purposes, instead of the 10-fold cross-validation scheme used in (Meena et al., 2015) leave-dialogue-out validation scheme was adopted to avoid that samples from the same dialogue could fall both in train and test sets for a given fold. The following sections report the results both for Let's Go and MT datasets.

4.3.3. Let's Go

Using the JRip classifier implemented in Weka (Hall et al., 2009) and the complete set of features, including both online and off-line features, we have obtained an Unweighted Average Recall (UAR) of 0.88, a very similar result to the one achieved in (Meena et al., 2015) for an older Let's Go dataset. The same performance was obtained in the Random Forest classifier from sklearn toolkit (Pedregosa et al., 2011). Using the libSVM classifier (Chang and Lin, 2011) the UAR was 0.73. Similarly to previous experiments when the SVM classifier was used for this task, the performance decreased. It seems that SVMs are less effective when dealing with bag of words/concepts features, generated from the dialogue acts.

4.3.4. Movie Ticketing

The feature set available for the MT dataset is fairly limited given that only the audios were provided initially. Given that we do not have access to the parser used by the live

[5] the term concept is used to refer to the pair slot type and slot value

system , as we had for Let's Go, we couldn't use dialogue acts from both system and user, nor extract the dialogue acts from the transcriptions. Therefore, we only used features that could be derived from the transcriptions, such as number of words or words repeated in consecutive turns, combined with the annotations for anger, valence, arousal, barge in, satisfaction and repetition.

The best performance achieved for miscommunication detection in the MT dataset was using the Random Forests implementation in Weka with a 0.75 UAR. Using JRip the performance drops to 0.72. As only numeric features are used in this case the performance achieved using SVMs levels the one achieved with JRip, 0.72.

4.3.5. Discussion

The results achieved show that the method presented in (Meena et al., 2015) is replicable to different datasets. The differences observed in performance between the different datasets are easily explainable by the lack of LU and some LG features that proved to be very helpful when detecting miscommunications in the datasets studied in (Meena et al., 2015).

We also compared the most relevant features to perform miscommunication detection in each dataset. The LU features and features that compare the speech recognition output with transcription, such as word and concept error rates, are the most relevant in the Let's Go data. Whereas in the MT dataset, the satisfaction is the highest ranked feature. This feature is ranked 36^{th} in the Let's Go ranking. We could hypothesize that the results would be similar on the MT dataset if the most relevant features were also available.

5. Conclusions

In this paper we have presented two different spoken dialogue system datasets that collected with real users, in two different languages (English and Greek) and two different domains of application. The datasets were aim to be useful resources to reduce the effort, expertise and time spent in the development cycle of new Spoken Dialogue Systems. The datasets were described in detail together with the annotation scheme adopted. To demonstrate the usability of the datasets we present three detectors developed with these datasets: anger, gender and miscommunication. The detectors achieved very good performance in tasks where humans have high agreement such as miscommunication (0.88 UAR for Let's Go) and gender (0.92 and 0.9 accuracies for Let's Go and MT respectively). The result achieved for anger using acoustic feature is very satisfactory (0.67 and 0.88 UARs for MT and Let's Go respectively), given the low agreement between annotators (0.38 Cohen Kappa agreement for Let's Go).

These results show that the development of automatic detectors for some tasks can be a reliable solution to shorten the cycle of creation of new dialogue systems, especially those detectors where humans have higher agreement when performing a similar task. The performance of the detectors is strongly limited not only by the amount of features that can be used to train them, as the results for miscommunication have shown, but also by the idiosyncrasies of each systems. System developers should take this into account when developing their systems so that systems log the necessary information to develop robust detectors.

6. Future Work

The next steps would be to proceed with the automatization process, reducing the human intervention and decreasing the time involved in the development process. By this we mean to identify more tasks that could be performed automatically, for instance to identify the causes for breakdown in communication (Martinovsky and Traum, 2003), identify behavioral patterns for specific users types and other tasks where humans have high agreement rate. As we iterate over the process, we hope to reduce the human intervention in the loop, as our detectors become more and more robust with more data being collected. The robustness of our platform, will necessarily require our detectors to be more system independent than they are today. Identifying more general features and normalize the data representation are still challenges to make spoken dialogue analytics more system independent.

7. Acknowledgements

The authors would like to thank Maxine Eskenazi and Alan W. Black for facilitating the access to the 2014 Let's Go data, Fernando Batista for helping with the gender detection results and Maria Vomva for annotating the MT dataset. This research is supported by the EU project SpeDial Spoken Dialogue Analytics, EU grant # 611396. A. Abad and H. Moniz are partially supported by Portuguese national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2013.

8. References

Alan W. Black, Susanne Burger, Brian Langner, Gabriel Parent, and Maxine Eskenazi. 2010. Spoken dialog challenge 2010. In Dilek Hakkani-Tür and Mari Ostendorf, editors, *SLT*, pages 448–453. IEEE.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Najim Dehak, Reda Dehak, Patrick Kenny, Niko Brümmer, Pierre Ouellet, and Pierre Dumouchel. 2009. Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification. In *Interspeech*, volume 9, pages 1559–1562.

F. Eyben, M. Wöllmer, and B. Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the international conference on Multimedia*, pages 1459–1462.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. 2009. The weka data mining software: An update. *SIGKDD Explorations*, 11(1).

Z. Harris. 1954. Distributional structure. *Word*, 10(23):146–162.

Hynek Hermansky, Nelson Morgan, Aruna Bayya, and Phil Kohn. 1992. Rasta-plp speech analysis technique. In *icassp*, pages 121–124. IEEE.

Brian ED Kingsbury, Nelson Morgan, and Steven Greenberg. 1998. Robust speech recognition using the modulation spectrogram. *Speech communication*, 25(1):117–132.

C. M. Lee, S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan. 2004. Emotion recognition based on phoneme classes. In *Proc. of InterSpeech*, pages 889–892.

J. Lopes, G. Salvi, G. Skantze, A. Abad, J. Gustafson, F. Batista, R. Meena, and I. Trancoso. 2015. Detecting repetitions in spoken dialogue systems using phonetic distances. In *INTERSPEECH-2015*, pages 1805–1809, Dresden, Germany, sep.

N. Malandrakis, A. Potamianos, E. Iosif, and S. Narayanan. 2013. Distributional semantic models for affective text analysis. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(11):2379–2392.

B. Martinovsky and D. Traum. 2003. The error is the clue: breakdown in human-machine interaction. In *Proceedings of the ISCA Tutorial and Research Workshop Error Handling in Spoken Dialogue Systems*, Château d'Oex, Vaud, Switzerland, aug.

R. Meena, J. Lopes, G. Skantze, and J. Gustafson. 2015. Automatic detection of miscommunication in spoken dialogue systems. In *Proceedings of 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 354–363, Prague, Czech Republic, sep. Association for Computational Linguistics.

Hugo Meinedo and Isabel Trancoso. 2011. Age and gender detection in the i-dash project. *ACM Trans. Speech Lang. Process.*, 7(4):13:1–13:16, August.

Tim Paek and Eric Horvitz. 2004. Optimizing automated call routing by integrating spoken dialog models with queuing models. In *HLT-NAACL*, pages 41–48.

E. Palogiannidi, E. Iosif, P. Koutsakis, and A. Potamianos. 2015. Valence, arousal and dominance estimation for english, german, greek, portuguese and spanish lexica using semantic models. In *Proceedings of Interspeech*, pages 1527–1531.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Antoine Raux, Brian Langner, Dan Bohus, Alan W. Black, and Maxine Eskenazi. 2005. Let's go public! Taking a spoken dialog system to the real world. In *INTERSPEECH*, pages 885–888. ISCA.

Alexander Schmitt, Michael Scholz, Wolfgang Minker, Jackson Liscombe, and David Suendermann. 2010. Is it possible to predict task completion in automated troubleshooters?. In Takao Kobayashi, Keikichi Hirose, and Satoshi Nakamura, editors, *INTERSPEECH*, pages 94–97. ISCA.

Alexander Schmitt, Benjamin Schatz, and Wolfgang Minker. 2011. Modeling and predicting quality in spoken human-computer interaction. In *Proceedings of the SIGDIAL 2011 Conference*, pages 173–184, Portland, Oregon, USA, June. Association for Computational Linguistics.

Alexander Schmitt, Stefan Ultes, and Wolfgang Minker. 2012. A parameterized and annotated spoken dialog corpus of the cmu let's go bus information system. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

SpeDial. 2015a. SpeDial Project – Deliverable D2.1: Interim report on IVR analytics and evaluation. `https://sites.google.com/site/spedialproject/risks-1`.

SpeDial. 2015b. SpeDial Project – Deliverable D3.1: Interim report on SDS Enhancement and Evaluation. `https://sites.google.com/site/spedialproject/risks-1`.

Marc Swerts, Diane J. Litman, and Julia Hirschberg. 2000. Corrections in spoken dialogue systems. In *INTERSPEECH*, pages 615–618. ISCA.

P. Turney and M. Littman. 2002. Unsupervised learning of semantic orientation from a hundred-billion-word corpus, technical report erc-1094 (nrc 44929). Technical report, National Research Council of Canada.

D. Ververidis, K. Kotropoulos, and I. Pittas. 2004. Automatic emotional speech classification. In *Proc. of ICASSP*, pages 593–596.

Marilyn Walker, Irene Langkilde, Jerry Wright, Allen Gorin, and Diane Litman. 2000. Learning to predict problematic situations in a spoken dialogue system: Experiments with how may i help you? In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, NAACL 2000, pages 210–217, Stroudsburg, PA, USA. Association for Computational Linguistics.