The SpeDial Datasets: Datasets for Spoken Dialogue Systems Analytics

José Lopes[1],Arodami Chorianopoulou[2]
Elisavet Palogiannidi[2], Helena Moniz[3], Alberto Abad[3,4]
Katerina Louka [5], Elias Iosif[6,7], Alexandros Potamianos[6,7]
[1] KTH Speech, Music and Hearing, Stockholm, Sweden
[2] School of ECE, Technical University of Crete, Greece
[3] INESC-ID, Lisboa, Portugal
[4] Instituto Superior Técnico, Universidade de Lisboa, Portugal
[5] Voiceweb S.A., Greece
[6]"Athena" Research and Innovation Center, Greece
[7]School of ECE, National Technical University of Athens, Greece
jdlopes@kth.se

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 and expertise. The SpeDial consortium (www.spedial.eu) is working to create a semi-automated process for spoken dialogue service development and speech service enhancement of deployed services, where incoming speech service data are semi-automatically transcribed and analyzed (human-in-the-loop). 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; Paek and Horvitz, 2004; Schmitt et al., 2010; Swerts et al., 2000). This problem is vital in the development cycle of speech services. However, very little data is publicly available to perform research on this topic. Except for (Swerts et al., 2000) the data is not publicly available. Nevertheless, even in this case the dataset does not contain interactions with real users or annotations. The LEGO corpus (Schmitt et al., 2012) included both interactions with real users and annotations for interaction quality (Schmitt et al., 2011), emotions and other automatically extracted features. This corpus was based CMU Let's Go system from 2006, whose performance is substantially worse than 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. Therefore, we decided to look for more recent data and work on a new annotation scheme (introduced in Section 2.).

We are making two datasets publicly available. The first from Let's Go (Raux et al., 2005) collected during 2014 and a dataset from a Greek Movie Ticketing (MT) system. The datasets were used to build several detectors that could help us on the one hand to analyze the users and their behavior (such as age and task success), and on the other hand could serve as inputs to improve miscommunication detection (such anger). Therefore, both corpora will be distributed with manual transcriptions for every user turn, together with gender, task success, anger and miscommunication annotations.

So far, the results for stand-alone classifiers for anger, gender and miscommunication are very promising, proving the usefulness of the datasets that we are releasing for future research in SDS Analytics.

2. Annotation scheme

The first step prior to annotate the data was to manually transcribe the user utterances in both datasets. The system prompts in the MT dataset were also transcribed since no system logs 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. Recognition output and transcription are presented to the annotator when performing the task. The annotators had access to the audio from the utterances when they were annotating. The annotators task was to evaluate if the second system turn

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.

is problematic or not based only on these turns. 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. In the MT data annotation the annotator performed a similar task but using the whole dialogue.

Along with the miscommunication annotation, annotators had to listen to the utterance audio file and identify if anger was present. In Let's Go 1 was used when anger was detected and 0 otherwise. The labels used in 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 between corpora the values in the interval $[1-3]$ were mapped into 1 and values 4 and 5 were mapped into 0.

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.

Finally, to annotate task success, the annotators should listen to the whole dialogue and verify that 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.

3. Datasets

3.1. Let's Go

This part of the dataset is composed of 85 dialogues between real users and the Let's Go Dialogue system. Initially 105 dialogues were randomly selected from 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 corpus was annotated following the scheme described in Section 2. for Let's Go data. The dataset was enhanced with features from ASR, Audio Manager, Dialogue Manager, Spoken Language Understanding and the estimated task success extracted from system logs. Some features derived from transcription and its parsing were also included, such as Word Error Rate and Concept Error Rate.

The dataset was annotated by two expert annotators. One of them completely annotated the corpus, 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.

3.2. Movie ticketing

The movie ticketing dataset consists of 200 dialogues in Greek collected through a call center service for retrieving information about movies/showtimes and booking tickets. 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.

To verify the quality of annotations, two other annotators labeled a subset of 60 dialogues from the original dataset for anger. The agreement between annotators found was 0.58 with 0.4 Kappa value –computed as the average pairwise agreement– according to the Fleiss coefficient.

4. Results

In this section, we briefly present a series of indicative experimental results for a variety of different tasks. For anger and miscommunication detection a leave-dialogue-out validation procedure was adopted. The results are reported in terms of Unweighted Average Recall (UAR) since there is often a bias to one of the classes. For the gender detection task, results are reported in terms of Accuracy for the complete set of turns from both datasets.

4.1. Anger detection in 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.

Speech-based system. Here, the aim is to capture the speaker's emotional state via the utilization of a set of low-level descriptors (LLDs) including prosody (pitch and energy), short-term spectral and voice quality features (Ververidis et al., 2004). The LLDs can be further exploited via the application of a set of functionals, in order to map the speech contours to feature vectors. OpenSmile is a widely-used toolkit that can be used for

extracting such features (Eyben et al., 2010). A detailed system description is provided in (SpeDial, 2015).

Text-based system. The goal is to estimate the emotional content of the transcribed speaker utterances. A word w can be characterized regarding its affective content 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). A detailed system description can be found in (Palogiannidi et al., 2015). For each speaker utterance, the valence, arousal, and dominance scores are computed for the respective words. The statistics of these scores (e.g., mean, median, variance, etc.) can be used as features.

Experiments and evaluations results. The goal is the detection of "angry" vs. "not angry" (i.e., 2-class classification problem) user utterances. For this purpose, the anger annotations were used. The best performance obtained by the speech– and text–based systems equals to 0.67 and 0.61 UAR, respectively for the MT data, exceeding the performance of majority–based classification regarded as naive baseline (0.5 UAR for binary problems). These performance scores were achieved by different classifiers, JRip for speech and Random Forest for text. The affective speech analysis was also applied over the Let's Go dataset for the task of anger detection achieving 0.88 UAR. The attempts to use the affective text analysis on Let's Go were in vain, since only 3 utterances in the whole corpus include lexical anger markers.

4.2. Gender detection in Let's Go and Movie Ticketing datasets

The Gender detector mainly consists of a multi-layer perceptron (MLP) trained with $\sim$70 hours of multi-lingual telephone speech data. Accuracies obtained are 91.7% and 89.7% in the Let's Go and Movie Ticketing datasets, respectively. Thus, the module seems to perform consistently in both datasets, independently of the language (notice that Greek data was not included in any of the MLP training sets). Although we consider these results quite promising, particularly considering the reduced amount of actual speech in most of the turns, we are currently developing and evaluating an alternative system based on i-vectors.

4.3. Miscommunication detection in Let's Go

We performed miscommuncation detection for this corpus using the same approach described in (Meena et al., 2015), except that instead of 10-fold cross validation we have performed leave-one-dialogue-out cross validation. Using the JRip classifier implemented in Weka (Hall et al., 2009) and the set of features that combines bag of concepts features, features derived from SLU and features derived from the utterance, both including online and off-line features, we have obtained an Unweighted Average Recall (UAR) of 0.88. The same performance was obtained in the Random Forest classifier from sklearn toolkit (Pedregosa et al., 2011). The libSVM classifier (Chang and Lin, 2011) achieved a 0.73 UAR.

5. Conclusions and Future Work

The key contribution of this work is the creation of datasets[1] in two languages that enable the investigation of various research tasks in the area of spoken dialogue analytics.

Regarding the experimental results, detectors seem to have very satisfactory performance in those tasks that humans have higher agreement such as miscommunication and gender. Also, the performance achieved for anger detection using acoustic features is quite satisfactory.

For the full version of this paper, we plan to extend the annotations, incluing more data annotated by two or more annotators. We are also currently working to integrate more detectors to those described in this abstract and we expect to report results in both datasets in the final version of the paper.

6. 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.

7. References

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`.

F. Eyben, M. Wllmer, 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).

R. Meena, J. Lopes, G. Skantze, and J. Gustafson. 2015. Automatic detection of miscommunication in spoken dialogue systems. In *Proceedings of 16th Annual*

[1]Details about the dataset access will be provided in the final version of the paper.

Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), pages 354–363, Prague, Czech Republic, sep. Association for Computational Linguistics.

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. 2015. SpeDial Project – Deliverable D2.1: Interim report on IVR analytics 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.