Detecting Problems in Spoken Child-Computer Interaction

Dino Seppi
Matteo Gerosa
FBK–irst, Trento
Italy
{seppi,gerosa}@fbk.eu

Björn Schuller
Technische Univ. München
Germany
schuller@tum.de

Anton Batliner
Stefan Steidl
Univ. Erlangen-Nürnberg
Germany
{batliner,steidl}@informatik.uni-erlangen.de

ABSTRACT
In this paper we describe the effectiveness of some linguistic features for detecting problems in spoken child-computer interactions. To this aim, we use an Automatic Speech Recognizer for generating the spoken word chain, and a word tokenizer for obtaining the lexical and stemming information. Automatic classification of each turn is eventually achieved by exploiting the frequencies of tokens' classes. The impact of ASR and tagger accuracy on automatic detection are discussed by comparing fully automatic with manually corrected approaches.

1. INTRODUCTION
Recently, children’s speech has been gaining growing attention from the research community and in industry. In fact, children do not represent just another group of the world’s population: whereas speech and multimedia technologies are still far from perfection for adults, children’s speech poses even more compelling questions to solve. Both acoustic and linguistic characteristics of children’s speech differ from those of adults: pitch, volume, formant positions, and co-articulations vary strongly due to anatomical and physiological development [7, 5]; the linguistic structure of the children’s utterance is not too uniform, and lapses, short or not well-constructed sentences, repetitions, and disfluencies are generally frequent, mainly depending on age and socio-economic factors [7].

In spite of these difficulties, the ease of children in adopting technology [8] has stimulated and boosted interest, above all for commercial applications. The three main speech-related fields where research on children speech has acquired considerable momentum are, probably, health care (aids for diagnostic and therapy), edutainment (aids for pronunciation, understanding), and entertainment (computer games).

In this study, we investigate critical dialogue phases in spontaneous recordings of children playing with the Sony pet-robot AIBO. Although the experimental setting is fixed and the robot has to complete a pre-determined sequence of actions, the child is led to believe that the robot is responding to his or her commands (Wizard-of-Oz). Therefore, both speech and emotional content can be considered spontaneous and prototypical with respect to real-life situations.

Unlike usual approaches that mainly rely on acoustic modeling [1], here we focus on the use of linguistic information only. In doing so, we have to tackle all the aforementioned problems: children’s acoustic variability is a major issue in Automatic Speech Recognition (ASR), and linguistic cues might be affected by ASR errors. The effects on the word tokenization (both POS tagging and stemming) for extracting feature vectors, as well as the influence of the tokenizer accuracy on the final classification are also considered: fully automatic results are compared with both the manual transcriptions of the dialogues and the manual correction of the tokenizer output or the manual assignment of POS/stem classes.

In Section 2 we describe the speech data adopted for the experiments. The system and its single components are presented in Section 3, while results of the experiment appear in Section 4. A brief description of ongoing and future work closes the paper.

2. SPEECH CORPUS
The corpus used is the FAU-AIBO emotion corpus (henceforth FAU-AIBO), a German database with recordings of children communicating with Sony’s AIBO pet robot [1]. Speech was collected from 51 children, aged 10-13, from two different schools, Mont and Ohm. The total amount of speech equals to about 9 hours, after removing pauses. The data are segmented into ‘turns’ of variable length, using as criterion a pause of ≥ 1.0 seconds.

The speech is intended to be spontaneous and emotion to be natural. Children are told to talk to AIBO like they would talk to a friend, but while they are led to believe that the robot is following their commands, the AIBO is actually being controlled by a human operator, the ‘wizard’. The framework is designed to provoke the children in order to elicit emotional behaviour by finding a balanced compromise of obedient and disobedient behaviour.

Some emotional state like angry or emphatic (and any marked deviation from a neutral speaking style [2]) can be taken as possible indication of some (starting) trouble in communication. If a child gets the impression that the machine does not understand her, she tries different strategies such as repetitions, reformulations, etc. or simply the use of a pronounced, marked, emphatic speaking style.

Labeling
Each word of FAU-AIBO has been annotated by five labellers separately, choosing amongst eleven different emotion classes, including neutral as default. Final labels are assigned by majority voting: if three or more labellers agree, the label is attributed to the word. These word-based labels are then mapped onto turn labels by employing the following strategy: fragments and auxiliaries are used as stop words.
For this study, emotion classes are mapped onto two cover classes: non-negative, comprising neutral and motherese, and negative, comprising emphatic as pre-stage of anger, touchy/irritated, reprimanding, and angry; details can be found in [1]. As motivated in the previous section, the turn label negative might probably indicate the presence of some problem in the child-robot interaction.

Data partitioning

We partition the whole corpus into two main parts. The turns selected for automatic classification are those characterized by the following property: the annotators agreement on turn labeling must be equal or higher than 60%. This means that automatic classification is performed on that part of the data that is most reliably annotated. This criterion leads to a sub-group of turns, called EVAL (3990 turns, i.e. 1775 non-negative and 2215 negative). The rest of the data, called TRAIN consists of 9652 turns and is used for ASR training only. More details about the characteristics of the datasets are illustrated in Table 1 and in [10].

3. SYSTEM DESCRIPTION

The system chain is composed by an ASR engine, a word tokenizer, and a statistical classifier.

Automatic Speech Recognition

The parametric representation of speech signals is obtained as follows. Each speech frame is parameterized into 13 Mel Frequency Cepstral Coefficients (MFCCs). Frame energy is represented as the first MFCC. These coefficients, plus Frequency Cepstral Coefficients (MFCCs). Frame energy, as well as the log of the sum of spectral magnitudes of the speech frame are used as features. Frame energy is represented as the first MFCC. These coefficients, plus the log of spectral magnitudes of the speech frame, are used as features. Cepstral mean subtraction is represented as the first MFCC. These coefficients, plus the log of spectral magnitudes of the speech frame, are used as features. Cepstral mean subtraction is performed on static features on an utterance-by-utterance basis. Acoustic models are state-tied, cross-validated, and tied to form 39-dimensional observation vectors. Cepstral mean subtraction is performed on static features on an utterance-by-utterance basis. Acoustic models are state-tied, cross-validated, and tied to form 39-dimensional observation vectors. Cepstral mean subtraction is performed on static features on an utterance-by-utterance basis.

Word Tokenization & Feature Extraction

Features are extracted from the transcribed turns by a two-step procedure: first, words are tokenized by lexical or morphological rules, then, the normalized frequencies of the tokens are used to generate feature vectors of the size of the tokens’ alphabet. The automatic extraction of POS and stems is performed using TreeTagger [9], using models trained on out-of-domain data. To avoid too large feature spaces, we resort to relatively coarse taxonomies:

POS: the Part-of-Speech is the most compact approach of lexical tokenization adopted in this paper. More specifically, we use two sets of POS classes (a compact set with 6 tags, and a more detailed one with ≈ 40 tags) and two extraction methods (fully automatic or manual). Manual methods are COVER and FULL: in the former, the ASR lexicon containing all word forms of FAU-AIBO is annotated manually with six POS cover classes such as noun, adjective, particle, etc.; ambiguities are solved either by clustering or by using heuristic rules. In the latter method, POS classes are extracted automatically and subsequently, the assignment is corrected manually. Fully automatic methods are FULL-AUTO and COVER-AUTO; they differ in the number of POS eventually obtained: COVER-AUTO is obtained by mapping FULL-AUTO on 6 POS tags.

STEM: the second tokenization adopted is stemming. Stemming stands for clustering of morphological variants of lexemes. This clustering reduces the number of entries in the vocabulary, i.e. the feature space, by mapping each word to its root (thereby the name; with Bag of Words, BOW, we identify stem-based features without any structured, either sequential or hierarchical infor-

Both acoustic and (bigram, closed vocabulary) language models (LMs) are trained on data of one school (within TRAIN ∪ EVAL) and tested on data of the other school. We distinguish two cases: (1) testing on the data of one school within EVAL only, or (2) testing on all the data of one school (TRAIN ∪ EVAL). In this 2-fold cross validation framework, out of vocabulary words of one fold are added as unigrams with flat probabilities to the LM estimated on the other fold, so to be in a closed vocabulary condition. Recognition performance, in terms of Word Error Rate (WER), on the data from the two schools together (2) equals to 22.6%. WER on the EVAL set only (1) is 24.0%. Both figures are obtained by 2-fold (school based) cross validation. Speech recognition results (2) are shown in Table 2: the difference in performance between the two schools is mainly explained by the fact that more data are available when training on OHM.

Table 1: Partitioning of FAU-AIBO speech corpus into TRAINing and EVALuation sets. TRAIN and EVAL are divided into 2 groups (schools) for ASR cross-validation. ASR results reported in Table 2 are obtained training on MONT data, and testing on OHM (third and last column), and vice-versa

<table>
<thead>
<tr>
<th>school</th>
<th>TRAIN</th>
<th>EVAL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MONT</td>
<td>OHM</td>
</tr>
<tr>
<td></td>
<td>MONT</td>
<td>OHM</td>
</tr>
<tr>
<td># speakers</td>
<td>25</td>
<td>26</td>
</tr>
<tr>
<td># words</td>
<td>15385</td>
<td>15405</td>
</tr>
<tr>
<td># turns</td>
<td>4915</td>
<td>4737</td>
</tr>
</tbody>
</table>
Table 3: Results of problem detection in child-computer interaction: above: POS-based features, below: STEM-based features; 51-fold CV, speaker independent. F-measures [%]

<table>
<thead>
<tr>
<th>id</th>
<th>transcription</th>
<th>MANUAL</th>
<th>ASR</th>
<th>CROSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>COVER-AUTO</td>
<td>68.4</td>
<td>61.6</td>
<td>61.3</td>
</tr>
<tr>
<td>6</td>
<td>COVER</td>
<td>68.1</td>
<td>63.1</td>
<td>62.6</td>
</tr>
<tr>
<td>40</td>
<td>FULL-AUTO</td>
<td>71.6</td>
<td>66.3</td>
<td>59.4</td>
</tr>
<tr>
<td>36</td>
<td>FULL</td>
<td>72.9</td>
<td>67.2</td>
<td>66.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>id</th>
<th>transcription</th>
<th>MANUAL</th>
<th>ASR</th>
<th>CROSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>250</td>
<td>COVER-AUTO</td>
<td>69.5</td>
<td>68.4</td>
<td>71.5</td>
</tr>
<tr>
<td>303</td>
<td>COVER</td>
<td>69.4</td>
<td>68.4</td>
<td>71.9</td>
</tr>
<tr>
<td>407</td>
<td>FULL-AUTO</td>
<td>76.1</td>
<td>74.1</td>
<td>71.4</td>
</tr>
<tr>
<td>436</td>
<td>FULL</td>
<td>76.1</td>
<td>74.4</td>
<td>71.9</td>
</tr>
</tbody>
</table>

Table 4: Results of problem detection in child-computer interaction: combining best POS- and STEM-based features, 51-fold CV, speaker independent. F-measures [%]

<table>
<thead>
<tr>
<th>id</th>
<th>transcription</th>
<th>MANUAL</th>
<th>ASR</th>
<th>CROSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>447</td>
<td>FULL-AUTO</td>
<td>76.0</td>
<td>73.7</td>
<td>71.5</td>
</tr>
<tr>
<td>472</td>
<td>FULL</td>
<td>75.6</td>
<td>73.8</td>
<td>72.6</td>
</tr>
</tbody>
</table>

conditions, where the SVMs are trained on human transcriptions and tested on automatic ones.

Most important results are: 1) The differences in classification performance using manual rather than human corrected tokenization is almost negligible, except for POS features, but only if ASR is involved. 2) ASR errors do affect classification: figures in column ASR are always lower than in column MANUAL; furthermore, in CROSS configuration (when the two transcriptions are used in parallel), we meet the worst behaviour. 3) Classification performance grows with the granularity of the feature set, even for ASR transcriptions. This is not entirely obvious, since POS features might have smoothed ASR errors, and could have been more robust in the presence of noise. Almost the same behaviour is observed when merging POS and STEM features (Table 4), where classification performance does not improve over the best row of Table 3. This means that POS and STEM features are highly correlated, also when ASR introduces inaccuracies in the transcriptions.

5. CONCLUSIONS AND FUTURE WORK

These outcomes aim at shedding light on the influence of ASR and tokenizer accuracy. In respect of this aim, an improvement of linguistic features for classification seems to be obtainable by improving ASR. In this direction, acoustic adaptation should probably be the first, mandatory step; first experiments using cluster-based speaker normalization alone (CMLLR) allow to gain a 10% (relative) improvement in ASR performance.

Unpublished experiments will extend this work by adopting more elaborate approaches, such as word and character n-grams (in the model proposed in this paper the information on the order of the words is lost), higher semantics, and string kernels. We also plan to systematically compare and integrate the linguistic feature sets described in this paper to acoustic feature vectors. Finally, significance tests will be used to measure the differences among results.

6. ACKNOWLEDGMENTS

The initiative to co-operate was taken within the EU Network of Excellence HUMAINE under the name CEICES [1]. Combining Efforts for Improving automatic Classification of Emotional user States. The research leading to these results has received funding from the European Community’s Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 211486 (SEMAINE), and the projects PFFSTAR under grant IST-2001-37599, and HUMAINE under grant IST-2002-50742. The responsibility lies with the authors.
7. REFERENCES


