Emotions in IVR Systems
Anger and Frustration

Emotions in Speech
November 9

Yves Scherrer

Articles

- J. Liscombe, G. Riccardi & D. Hakkani-Tür:
  Using Context to Improve Emotion Detection in Spoken Dialog Systems
- L. Devillers & L. Vidrascu:
  Real-life Emotions Detection with Lexical and Paralinguistic Cues on Human-Human Call Center Dialogs

Overview

	Liscombe et al.	**Devillers & Vidrascu**
Setting of data collection	Phone account information Automated dialog system	Medical emergencies Human-Human interactions
Motivation	Improve customer satisfaction	Study real-life speech in highly emotive situations
Studied emotions	Negative, non-negative (but 7 emotions annotated)	Anger, Fear, Relief, Sadness (but finer-grained annotation)
Corpus used for experiments	5690 dialogs 20,013 user turns	680 dialogs 2258 speaker turns
Training-test split	75% - 25%	72% - 28%
Machine Learning method used for classification	Boosting algorithm	Log-likelihood ratio (linguistic) SVM (paralinguistic)

Features

	Liscombe et al.	**Devillers & Vidrascu**
Lexical features / Linguistic cues	Trigrams of user utterances	Unigrams of user utterances, Stemming
Prosodic features / Paralinguistic cues	Loudness (energy), Pitch contour (F0), Speaking rate, Voice quality (jitter), Hesitations, Turn-final pitch contour, Pitch accent. Normalized by gender.	Loudness (energy), Pitch contour (F0), Speaking rate, Voice quality (jitter, ...), Disfluency (pauses), Non-linguistic events (mouth noise, crying, ...). Normalized by speaker.
Dialog act features	Domain-dependent dialog act tag	--
Contextual features	Differential of prosodic features Transcriptions Repetition measure Dialog acts ... of 2 previous turns	--

Liscombe et al.

- Motivation:
  - Two sources of user frustration:
    - Reason of call (complaint about bill, …)
    - Arising from interaction problems with the spoken dialog system
  - Goal:
    - Detect the problem
    - Try to repair it, or transfer to human operator
    - How could a spoken dialog system “repair” an interaction?

Corpus

- 20,013 user turns from 5,690 dialogs
- Emotional states:
  - positive/neutral, somewhat frustrated, very frustrated, somewhat angry, very angry, somewhat other negative, very other negative
  - Simplified set: positive/neutral, negative (Wise choice?)
- Inter-annotator agreement:
  - 0.32 Cohen's Kappa for full set (“fair agreement”)
  - 0.42 for simplified set (“moderate agreement”)

Automatic Classification

- Features used:
  - 1 lexical feature
  - 17 prosodic features
  - 1 dialog act feature
  - 61 contextual features
- 2000 iterations with the BoosTexter boosting algorithm
  - Each user turn must be classified as negative or non-negative given a set of 80 features

Automatic Classification

- Boosting algorithm:
  - Boosting is a general method of producing a very accurate prediction rule by combining rough and moderately inaccurate “rules of thumb.”
    http://www.cs.princeton.edu/~schapire/boost.html
  - Can a set of weak learners create a single strong learner?
    - A weak learner is a classifier which is only slightly correlated with the true classification (it can label examples better than random guessing).
    - A strong learner is a classifier that is arbitrarily well correlated with the true classification.
      http://en.wikipedia.org/wiki/Boosting

Lexical features

- Unigrams, bigrams, trigrams of transcription
- Result:
  - Words correlating with negative user state:
    - dollars, cents, call
    - person, human, speak, talking, machine
    - oh, sigh
  - What can these results tell us about emotion annotation?

Prosodic features

- Features:
  - Energy (loudness)
  - F0 (pitch contour)
  - Speaking rate
  - Turn-final pitch contour
  - Pitch accent
  - Voice quality (jitter)
- Normalization:
  - Speaker normalization not possible (data sparsity)
  - Gender normalization

Dialog act features

- 65 specific, domain-dependent dialog act tags:
  - Yes
  - Customer_Rep
  - Account_Balance
- Why should these tags work better than the words of the utterances?

Contextual features

- First-order differentials of prosodic features wrt. 2 previous utterances
- Second-order differentials of prosodic features wrt. 2 previous utterances (Why?)
- Transcriptions of 2 previous utterances
- Measure of repetition (Levenshtein edit distance)
- Dialog acts of 2 previous user turns
  - Once frustrated, always frustrated?
- Dialog acts of 2 previous system turns

Results

- No surprises…
- What do you think about these results?

	Accuracy rate
Baseline	73.1% (majority)
Lexical + prosodic features	76.1%
Lexical + prosodic + dialog act features	77.0%
Lexical + prosodic + dialog act + context	79.0%

Devillers & Vidrascu

- Motivations:
  - “The context of emergency gives a larger palette of complex and mixed emotions.”
  - Emotions in emergency situations are more extreme, and are “really felt in a natural way.”
  - Debate on acted vs. real emotions
  - Ethical concerns?

Corpus

- 688 dialogs, avg 48 turns per dialog
- Annotation:
  - Decisions of 2 annotators are combined in a soft vector:
    - Emotion mixtures
  - 8 coarse-level emotions, 21 fine-grained emotions
  - Inter-annotator agreement for client turns: 0.57 (moderate)
  - Consistency checks:
    - Self-reannotation procedure (85% similarity)
    - Perception test (no details given)

Classification

- Restrict corpus to:
  - Utterances from callers
  - Utterances annotated with one of the following non-mixed emotions:
    - Anger, Fear, Relief, Sadness
  - Justification for this choice?
- This yields 2258 utterances from 680 speakers.

Lexical cue model

- Log-likelihood ratio:
  - 4 unigram emotion models (1 for each emotion)
  - A general task-specific model
  - Interpolation coefficient to avoid data sparsity problems
    - A coefficient of 0.75 gave the best results
- Stemming:
  - Cut inflectional suffixes (more important for rich-morphology languages like French)
  - Improves overall recognition rates by 12-13 points

Paralinguistic (prosodic) cues

- 100 features, fed into an SVM classifier:
  - F0 (pitch contour) and spectral features (formants)
  - Energy (loudness)
  - Voice quality (jitter, shimmer, ...)
  - Speaking rate, silences, pauses, filled pauses
  - Mouth noise, laughter, crying, breathing
- Normalized by speaker
  - Here: ~24 client turns in each dialog
  - Liscombe et al.: 3.5 client turns in each dialog
    → data sparsity

Paralinguistic (prosodic) cues

- Voice quality
  - Jitter: varying pitch in the voice
  - Shimmer: varying loudness in the voice
  - NHR: Noise-to-harmonic ratio
  - HNR: Harmonic-to-noise ratio

Results

	Anger	Fear	Relief	Sadness	Total
Number of utterances	49	**384**	107	100	640
Lexical cues	59%	**90%**	86%	34%	78%
Prosodic cues	39%	**64%**	58%	57%	59.8%

- Relief is associated to lexical markers like *thanks* or *I agree*.
- “Sadness is more prosodic or syntactic than lexical.”
- Comments?

Results

	Liscombe et al.	Devillers & Vidrascu
Baseline	73.1% (majority)	25% (random)
Lexical/linguistic features	--	78%
Prosodic/paralinguistic features	75.2% (see thesis)	59.8%
Lexical + prosodic features	76.1%	--
Lexical + prosodic + dialog act features	77.0%	--
Lexical + prosodic + dialog act + context	79.0%	--