Lexical and prosodic cues for emotion detection on call centers corpora

L. Devillers, L. Vidrascu

LIMSI – CNRS, TLP, France

collaboration with I. Vasilescu ENST France

HUMAINE – WP4 Workshop - Santorini

Content

- Call center data
  - Emotion annotation protocol
  - Our corpora
- On-going research on
  - Features: Prosodic, disfluences and lexical cues
  - Classifiers: Decision trees, SVM, …
  - Examples: performance detection, features analysis
- Summary – on going work
- Poster presentation

Speech recorded from real-life Call Centers

ü Cons:

- § difficult to obtain,
- § on-emotion data are sparseness,
- § impossible to share data for ethical reasons despite guaranteed confidentiality procedures,
- § telephonic data (bandwidth ~4k Hz)

ü Pros:

- § ecological emotions are rare,
- § point of reference for other material: understand limits and values of naturalistic material.

First of all… emotion annotation

ü Iterative protocol:

- § define emotion labels (limited number)/abstract dimensions:
  Task-dependent labels ←——→ Universal dimensions
  valence/activation dimensions: insuffisant for Fear/Anger classification
- § define segmental units: automatic(pauses, syntax based)/manual
- § annotate : one label/segment -> combined: major/minor labels
  contextual information
  at least 2 annotators, 3 for majority voting, …
- § validate:
  - § inter-annotator agreement (low Kappa -> complex samples)
  - § perceptual tests

LIMSI-CNRS WP4 - Santorini September, 2004

Call Center H-H Corpora

- **CORPUS 1: Stock Exchange Customer Service Center**
  Fear of losing money!
  fear/apprehension, anger/irritation, satisfaction, excuse, neutral
  2 annotators – 12 % of speaker turns with emotion - kappa 0.8
  100 dialogs, 5000 speaker turns

- **CORPUS 2: Capital Bank Service Center**
  fear/apprehension, anger/irritation, satisfaction, excuse, neutral
  2 annotators on 1K speaker turns randomly extracted – 10 % of Speaker turns with emotion - kappa 0.5
  250 dialogs, 5000 speaker turns extracted

- **CORPUS 3:** Emergency **Service Center**
  Fear of being sick, real fear/panic, call for help
  14 classes: anxiety, despair, disppointment, fear, hot anger, hurt, impatience, neutrality, panic, relief, shock, stress, surprise, worry
  annotation (on-going process, transcriber tool),
  134 speaker turns extracted from a dozen dialogs (1st annotation negative/positive)
  6h30 transcribed on 20h

Speech from Multimodal Corpora

- **CORPUS 4: Emot TV (HUMAINE)**
  51 interviews, 18 emotion labels, multimodal annotation, 2 annotators
  [Ref] S. Abrilian, L. Devillers, JC Martin, S. Buisine, SummerSchoolWP5

Results and multimodal scheme will be presented at WP5 december workshop

High subjectivity -rigorous methods

Some Problematic issues adressed in Humaine WP 3-4-5

ü Emotion definition
ü Annotation protocol
ü Assessment of annotations: quality/reliability
    ü perceptual tests
        ü audio only, video only and audio+video
    ü inter-annotator agreement measures
ü Assessment of automatic emotion recognition
    ü metrics, benchmarks
    ü dealing with confusion

Relevant features for emotion detection

~~Spectral features~~
(speech recognition)

~~Linguistic models of intonation,~~
~~(speech synthesis)~~

X

Prosodic Features
(F0, Energy, formant)
↓
Extracted dynamic features:
Max F0 slope, delta F0,…

Signal/transcription align.
(speech rate)
↓
Extracted speech disfluencies:
pauses: silent, filler « euh »…

↓
~ 50 features vector
best parameters selection -> smaller vector
Other features: lexical, dialogic, voice quality, non verbal events

Classifiers

- Decision trees (Cart (pruning) and Regression Trees), Support Vector machines (SVM), Languages Models LM, NN, HMM…

- Meta algorithms : combining models

  Bagging: t models with a training set randomly modified
  Majority vote for combining the t predictions

  Boosting: a weight is associated with each instance of the training set.

- Our results of experiments show little differences, slighty better with combining techniques.

Examples of results on call center data (Corpus 1)

POS/NEG *speaker-turn-based classification*

- Different models (prosodic/no disfluencies cues):

	C4.5	AdaBoost	ADTree	SVM
10att	73.0 (5.3)	71.5 (4.9)	**73.0** (5.8)	69.5 (5.6)

- Performances (prosodic/no disf., lexical features):

Features	RR
prosodic	73%
lexical	78%

- Higher for Neg/Pos than for Anger/Fear

Features selection

- Data-mining requires attribute selection.
- Better with only 10 best parameters: min E, max delta F0, Duration, …
- Small features vector size reduces training time

	C4.5	AdaBoost	ADTree	SVM
10att	73.0 (5.3)	71.5 (4.9)	**73.0** (5.8)	69.5 (5.6)
Allatt	69.4 (5.5)	71.7 (4.4)	71.7 (4.8)	69.7 (3.6)

C4.5: Decision Trees with pruning
AdaBoost: boosting,
AdTree: bagging

Emotion in interaction

- Emotions are speaker and conversational context dependent:
- Some results on Corpus 1:

  Speech disfluencies features:

  - higher correlation with Fear than with Anger but results are speaker dialog position dependent (client vs agent)

  F0 features:

  - different behavior between clients and agents:
    - stronger F0 cues for clients than agents
    - clients: stronger F0 cues for Fear than Anger, contrary for agents
  - different behavior between males and females
    - moderate variations for women (10 females/94 males)
  - inter-correlation between agents and clients behavior:
    - one agent shows extreme F0 manifestation/ his behavior influences the clients' attitudes

[ref] L. Devillers, I. Vasilescu, L. Vidrascu: speech prosody 2004.

Summary and on-going work

- Emotion annotation
  - New annotation scheme:

    combined labels: a major and a minor emotion (if necessary)
    valence/activation cues
    dominant appraisal ?
  - Add contextual annotation: what-for ? Position in the dialog …
  - Validation protocol is required
- Features
  - Selection procedure is useful
  - Speech disfluencies extraction (more robust detection)
- Classifiers
  - Combination of models (prosodic, lexical, …)
  - Speech model adaptation to multimodal corpora
  - Cross-corpus validation -> POSTER

Poster

- **Evaluation of robustness:**
  - **1st annotation scheme (1 label/speaker turn)**
  - **lexical and prosodic cues**

- Results for NEG/POS classification
  - based on prosodic cues for two call centers: financial/emergency
  - based on lexical cues for two call centers on financial aspects

LIMSI-CNRS WP4 - Santorini September, 2004