Recognizing Affect in Dialog Systems

Nate Perkins

Problem

- Identify emotional state in human speech dialog

Why?

- Tutoring systems
- Call center systems
- Second language learning systems
- Virtual agents

What are we identifying?

Emotional state is difficult to define for humans let alone computers

Target broad categories

- Positive/negative/neutral
- Negative/non-negative
- Certain/uncertain
- Positive/negative, active/passive
  - positive-active : joy
  - negative-passive : frustration

How do we identify it and then annotate?

- Cross-validation of annotations
- Coached utterances targeting specific emotional states

What features are relevant?

Overview
- focus on ‘what’, ‘how’, and ‘when’ something is said

Acoustic prosodic
- Fundamental freq stats
- Energy/intensity
- pitch

Acoustic temporal
- Total time
- Total silence
- Speaking rate

Lexical
- Word n-grams
- Character n-grams
- Emotional salience
  - Mutual information between words and emotional state
  - derived

- Discourse
  - Acoustic barge-in
  - Question
  - Semantic barge-in
  - Rejection
  - Repeat
  - ‘local’ vs ‘global’ features
    - ‘local’ – prior two utterances’ features
    - ‘global’ – avg of all prior utterances
- Speaker
  - Gender
  - Subject
- Facial

Models

- Independent classifiers for different categories
- Aggregate classifiers via interpolation
- Try different combinations to find best result

Results

- Some instances where non-acoustic out-performed acoustic in certain experiments
- Acoustic + lexical
- Generally : mix of all feature categories performs best

Questions

What do you see as the next steps in terms of using these predictions in a dialogue system? The authors mention that this information can "enhance" their tutoring system but they don't explicitly go into how. For example, if the system knows the user is experiencing a "negative" emotion, how might it adapt to address that?

I found their classification into negative, positive and neutral groupings a little unnatural and unsatisfying. For example, "bored" is part of the negative group but it seems like one might express boredom with a lack of emotion, but "no strong expression of emotion" is how the neutral category is defined. And "frustration" and "uncertainty" are also both part of the negative category but it seems like these would be expressed with vastly different features. Thoughts?

The authors of “Predicting Emotion in Spoken Dialogue from Multiple Knowledge Sources” call contextual features, local and global, the features of the two preceding students and the average of all students features. How is this related to a ‘context’ for the emotions of a student?

Questions (cont)

The authors of “Predicting Emotion in Spoken Dialogue from Multiple Knowledge Sources” assume that implementing emotions in a automated dialog system should improve the performance of such a system. Isn’t this though contrary to the experience of people, that tend to behave differently with a machine than with a human? As the corpus for this study is on a human-human dialog corpus, the results should not be easily transferable to an automated system, or?

I’m interested in Thor’s second question—the assertion that this system may not be easily transferable to a human-machine interaction given its training on a human-human corpus. I agree with this assessment, but I also wonder: isn’t the goal of spoken dialogue systems to facilitate a conversation such as those experienced in human-human interaction? If that is the case, then training on a human-human corpus makes sense for a long-term goal. Is it feasible to expect humans’ behavior with spoken dialogue systems to change as systems improve, and should research be preparing for this purpose?

How would it extend to non-English language, and non-college level student, settings?
Is the system of annotation language independent, since it is a human (native speaker) process?
The authors mention they are exploring other emotion annotation schemes - are any of those language/ culture group agnostic (is that even a possibility)?

Could the manual features, such as barge-in or ''is question'', be automatically derived from the raw data they currently have?

Using just lexical items produced a relatively high accuracy, which differs from other studies. Is that due this specific context / domain?

Modeling Affect in Dialog

Katherine Topping

LING575 Spring 2016

Adapting to Multiple Affective States in Spoken Dialogue

Overview

- Wizard-of-Oz tutoring system
- Previous work on multiple affect systems showed no significant improvements in task success, but showed other benefits such as increased user satisfaction
- Comparing effectiveness of system recognizing only one affect (uncertainty) versus new system responding to two different user affects (uncertainty and disengagement)
    - Two most frequent user affective states that occur in system

Baseline System: UNC_ADAPT

- (Un)certainty automatically classified by logistic regression model
    - Features of speech signal (i.e. prosody)
    - Automatic transcript
    - Dialogue context
- System responds based upon answer's (in)correctness and (un)certainty
- Wizard used in present experiment
    - Inter-annotator agreement of 0.85 (correctness) and 0.62 (uncertainty) Kappa

New System: UNC-DISE_ADAPT

- Adds disengagement, characterized by signs of boredom or irritation
    - Leaden monotone, sarcasm, off-task sounds
- Inter-annotator agreement of 0.55 Kappa
- Responses divided into correct+disengaged (COR-DISE) and incorrect+disengaged (INC-DISE)
- Hypothesized that UNC_ADAPT response to incorrectness insufficient for INC-DISE turn (user already disengaged)
    - User must reengage to benefit from supplementary info
- System gives "productive interaction feedback" to INC-DISE turns, followed by fill-in-the-blank version of original question

UNC-DISE_ADAPT ITSPOKE-1: Let's begin by looking at the motion of the man and his keys while he's holding them. How does his velocity compare to that of his keys?

USER-1: vertical (INC, UNC, DISE)

UNC-DISE_ADAPT ITSPOKE-2: That doesn't sound right. Let's think about this a little more. *[Productive Interaction Feedback]* Since the man is holding his keys, they aren't moving relative to each other. So their velocities must be WHAT? *[Supplementary Question]*.

Figure 2: Dialogue Example Showing Adaptation for Disengaged+Incorrect Turns

UNC-DISE_ADAPT ITSPOKE-1: Let's begin by looking at the motion of the man and his keys while he's holding them. How does his velocity compare to that of his keys?

USER-1: same same same (COR, CER, DISE)

. . . . . .

UNC-DISE_ADAPT ITSPOKE-2a: Well done. Here's a quick progress update. Good effort so far!!*[Productive Interaction Feedback When Progress Report Shows Improvement]* Now let's see what happens after the man releases his keys. [...] *[Next Topic Question]*

. . . . . .

UNC-DISE_ADAPT ITSPOKE-2b: Right. Here's a quick progress update. It might help to remember we will build on the topics we're discussing now.*[Productive Interaction Feedback When Progress Report Shows Decline]* Now let's see what happens after the man releases his keys. [...] *[Next Topic Question]*

Figure 3: Dialogue Example Showing Adaptation for Disengaged+Correct Users

Experimental Procedure

- College students with no college-level physics
- Assigned to either UNC_ADAPT or UNC-DISE_ADAPT
- Users:
  - Read short physics text
  - Took pretest and pre-motivation survey
  - Worked 5 "training" problem dialogs with system
  - Took post-motivation survey and user satisfaction survey
  - Took posttest isomorphic to pretest
  - Worked a "test" problem with UNC_ADAPT

Performance

Table 2: Global Performance Metrics Across Conditions (All UNC vs. UNC-DISE Differences Yield p>=.274; All NO-ADAPT Differences Yield p≤.003)

Cond	N	LearnGain		UserSat		MotGain	
		Mn	sd	Mn	sd	Mn	sd
Unc	19	.65	.20	.69	.11	.01	.07
Unc-Dise	19	.58	.19	.66	.09	.01	.07
NoAdapt	21	.38	.20	-	-	-	-

- Small decrease in learning gain/user satisfaction means for UNC-DISE
- Previous study showed UNC had significantly higher learning gain than no-adapt system
- UNC-DISE also outperforms no-adapt consistently
    - While adding new affect adaptations may not yield additive improvements, it also doesn’t hurt performance

Performance

Table 3: Motivation Gain Differences Across Condition for High and Low DISE Users (p=.035)

Condition	Split	N	MotGain	
			Mn	sd
UNC	high DISE	9	-.01	.04
UNC-DISE	high DISE	7	.04	.07
UNC	low DISE	10	.03	.08
UNC-DISE	low DISE	12	-.01	.06

- Low-DISE users had higher motivation gain in UNC_ADAPT
- High-DISE users had higher motivation gain in UNC-DISE_ADAPT

Performance

Table 4: Differences Across Condition for Test Dialogue

Metric	Condition	Mn	sd	p
UNC → UNC	UNC	.06	.09	.05
	UNC-DISE	.01	.04	
INC+UNC+ENG →	UNC	.01	.03	.10
COR+CER+ENG	UNC-DISE	.03	.05	
INC+CER+ENG →	UNC	.00	.00	.04
INC+CER+DISE	UNC-DISE	.02	.03	

- Uncertain answers more likely to remain uncertain in UNC_ADAPT than UNC-DISE_ADAPT
- Incorrect+uncertain+engaged answers more likely to become correct and certain in UNC-DISE_ADAPT
- Incorrect+certain+engaged answers more likely to become disengaged in UNC-DISE_ADAPT

Performance

Table 6: Mean L Values for Disengagement State Transitions

Condition	Turn n	Turn n+1		
		ENG	DISE	p
UNC-DISE	ENG	.06	-.01	.04
	DISE	-.35	.06	.14
UNC	ENG	.09	-.03	.01
	DISE	-.41	.09	.06

- L = transition likelihood
- In both conditions, engaged user in turn n significantly likely to remain engaged in turn n+1
- In UNC_ADAPT, disengaged user in turn n more likely to remain disengaged in turn n+1
- In UNC-DISE_ADAPT, disengaged user equally likely to become disengaged or engaged
    - Benefit at local performance level

Critique

- Fairly low inter-annotator agreement for uncertainty and disengagement
    - Mentioned that next steps include automated UNC-DISE_ADAPT
- Binary nature of measurements across the board
- Did not increase/decrease task success
    - Argued in summary that automated system could potentially yield greater global success
- Would have liked more detail regarding motivation behind chosen response schemes

Emotion and Dialogue in the MRE Virtual Humans

Overview

- Mission rehearsal exercise with virtual humans working towards resolving a given scenario
    - Can interact with people or with other virtual humans
- Task model, dialogue model, and emotional model all working together

Task Model

- Agent’s task model represents understanding of task in general
- Agents use partial-order planning algorithm over task model to guide execution of task and handle unexpected events that require adaptive execution or re-planning
- Result of planning algorithm specifies how agent privately believes the team can collectively complete the task
    - This plan is continuously revised

Dialogue Model

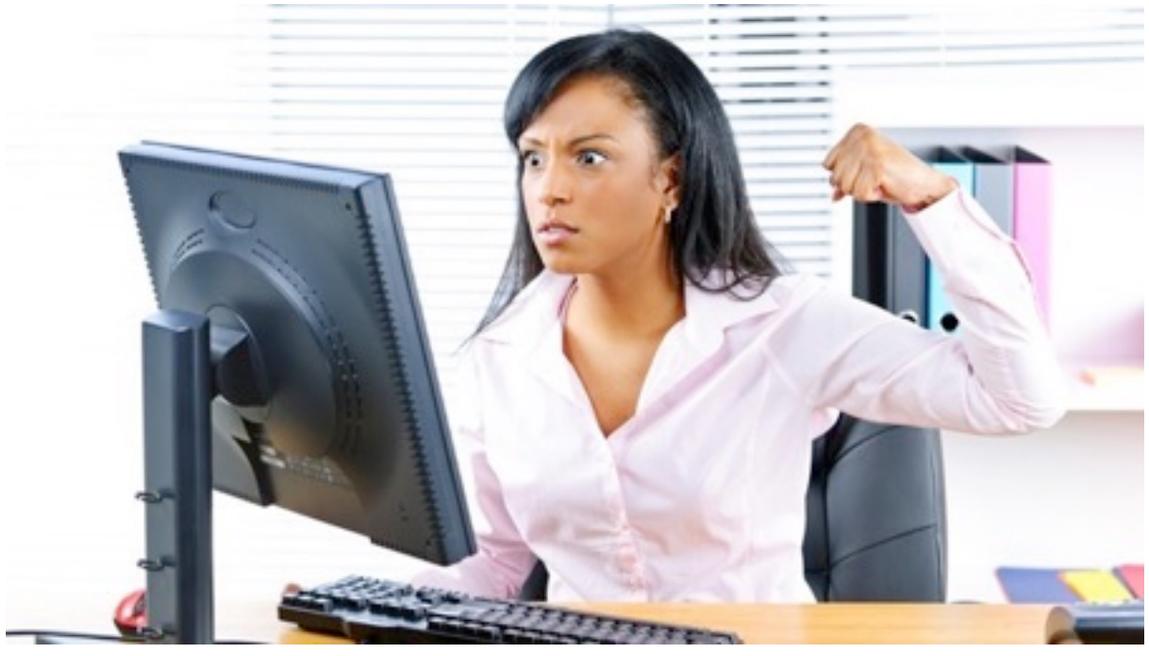
- Supports multiple simultaneous conversations with potentially overlapping groups of interlocutors
- Information state = part of context deemed relevant for dialogue modeling
    - Maintained as a snapshot of dialogue state
- Core speech acts have content which is either a state, action description, or question about one of these
    - Assert, into-request, order, request, suggest
    - Forward-looking acts and backward-looking acts

Emotion Model

- EMA (EMotion and Adaption)
- Appraisal theory
    - Events do not have significance alone, but only by virtue of their interpretation in the context of the individual's beliefs, desires, intention, and past events
- Appraisal = set of feature detectors that characterize current state of agent's mental processes
- Supports multiple appraisals of same event and multiple events simultaneously
- Coping strategies identify precursors of emotion that should be maintained or altered

Emotions in Effect

- "What happened here?"
    - Using "concerns" of agent, calculated by emotion reasoning, agent can report on the one that causes itself the strongest emotion
- Emotion module can indicate to dialogue manager that there is an important issue to discuss
    - Agent can take initiative to bring up new topic
- Coping strategy to shift blame
    - Agent can inform content realization to bias the way it phrases dialogue
    - "We collided" vs "They rammed into us"



"How was your day?" An Affective Companion ECA Prototype

Overview

- Embodied Conversational Agent (ECA)
- Not task-based; focused upon open user-initiated conversation about day at the office
- Makes empathetic and sympathetic comments, offers advice
- Can handle long user turns, generate long system turns
- User can interrupt system

System Behavior

$User_1$	Worst day ever. I arrived late because of the traffic.
$System_1$	You have my sympathy. What happened next?
$User_2$	I missed the meeting on project responsibilities; got more work dumped on me. Just what I don't need. I'm already likely to miss the deadline. [*ECA looks worried when user's negative tone is detected*]
$System_2$	Sorry to hear that. Who was at the meeting?
$User_3$	Only my team. They gave me a set of PowerPoint slides from the meeting. I spent the rest of the day reading them but my computer kept crashing because of updates.
$System_3$	Oh dear! It is understandable how you must be feeling. It's right to be unhappy about such a bad situation. Try not to worry about missing the deadline because of the technical problems. [*At this point user interrupts ECA by beginning turn* $User_4$] It's sad that your motivation could be affected by them. [*System stops speaking*]
$User_4$	No, I'm not worried about the technical problems, I'm worried about the amount of extra work on this project and the deadline.
$System_4$	What is the project about?

Table 1: Dialogue example

- Events recognized in user turn are labelled with output of Emotion Module
    - Semantic and affective info
- When system gains sufficient understanding of key event in user's day, generates complex long turn
    - Comfort, opinion, warnings, and advice
- Affective Strategy Model makes appraisal of user's situation, generates appropriate emotional strategy
- Short feedback loop and long feedback loop

A Tractable Hybrid DDN–POMDP Approach to Affective Dialogue Modeling for Probabilistic Frame-Based Dialogue Systems

Overview

- Partially Observable Markov Decision Process (POMPD) & Dynamic Decision Network (DDN)
- Two main parts of system:
    - Slot-level dialogue manager
    - Global dialogue manager
- Two new features introduced by system:
    - Ability to deal with large number of slots/slot values
    - Ability to take into account user’s affective state when deriving adaptive dialogue strategies

System Behavior

- Instead of keeping track of slot values, keeps track of probability distributions for values
- Because user's state cannot be directly observed, system uses state estimator to compute internal belief state and selects next action based upon given policy
- Slot-based part of system
    - Each slot modeled as factored POMDP
        - State set includes user's emotional states, goals, actions, etc
    - Approximated as set of DDNs
- Global part of system
    - Dialogue information state (keeps track of emotional state)
    - Action selector
- Affect focused upon detection of uncertainty and change over time

System Performance

- POMPD model ideal for small number of slots/values
- DDN-POMPD method handles larger numbers of slots/values much better
- Copes well with errors, especially speech recognition errors
- System is on-par with state-of-the-art counterparts

Discussion

GoPost Questions

- The authors state that “supplementary information can help reduce some types of disengagement for highly disengaged users.” But their disengagement status appears to be binary: engaged/ disengaged. Would it be possible and helpful to try to identify different levels of disengagement?

- The authors’ prior work suggests that the noise introduced in classification errors in the fully automated system (vs. the wizard-of-oz approach) actually produces better global performance. Is this because the (uncertain or disengagement) adaptation would appear more randomly and less predictably? Why would that produce better performance?

GoPost Questions

- The paper says that the disengagement adaptation was more effective at improving task success for correct turns than incorrect turns, but that the disengagement adaptation increased user satisfaction for incorrect turns. (p.223)

- Does this imply that once the user has begun answering incorrectly, the disengagement adaptation does nothing to help them get back on track?

- It seems like a major problem that the system is ineffective at helping users get back on track. What potential solutions are there to this problem?

Sentiment and Subjectivity in Dialog

—

Micaela Tolliver

What is sentiment? Why is it useful in NLU?

- Sentiment and Subjectivity: expressing a non-objective opinion or statement
- Past research focused on online text, rather than spoken text
- Sentiment analysis can be used to extract more information and knowledge from the dialog exchange
- Useful in natural language understanding domains:
  - Meetings
  - Opinion pieces
  - Other possibilities

Annotating Subjective Content in Meetings. Proceedings of the Language Resources and Evaluation Conference, Wilson (2008)

- Purpose: How do we represent sentiment in dialog?
- Domain:
  - Multi-party conversations, primarily AMIDA corpus
  - Meeting conversations
- Problems with old schema for sentiment:
  - Didn't capture everything needed for dialog exchanges (questions)
  - Some concepts (deeply nested sentiments) less useful

Wilson: Annotations for Sentiment in Dialog

- Subjective Utterance: “a span of words where a private state is being expressed either through word choice or prosody”
  - Different types of subjective utterances, like positive or negative
- Private State: “Internal mental or emotional state, including opinions, beliefs, sentiments … among others”

Wilson: Annotations for Sentiment in Dialog

Subjective Utterances	**Subjective Questions**
Positive subjective	Positive subjective question
Negative subjective	Negative subjective questions
Positive and negative subjective	General subjective question
Uncertainty	**Objective Polar Utterances**
Other subjective	Positive objective
Subjective fragment	Negative objective

Wilson: Annotations for Sentiment in Dialog

- Example:
  - Um it's very easy to use. Um but unfortunately it does lack the advanced functions which I I quite like having on the controls.
  - Um <POS_SUBJ it's very easy to use>. Um <NEG-SUBJ but unfortunately it does lack the advanced functions><POST-SUBJ which I I quite like having on the controls>.

Multimodal Subjectivity Analysis of Multiparty Conversations, Raaijmakers et al (2008)

- Purpose and Domain:
  - Recognize subjectivity in Multi-Party Meeting Dialogs
- Method and Data:
  - Use transcribed and annotated meeting recordings from the AMI Meeting Corpus with AMIDA annotations
  - Utilize linguistic features and machine learning to classify subjectivity
  - Understand which features and combinations improve the output

Raaijmakers et al: Tasks

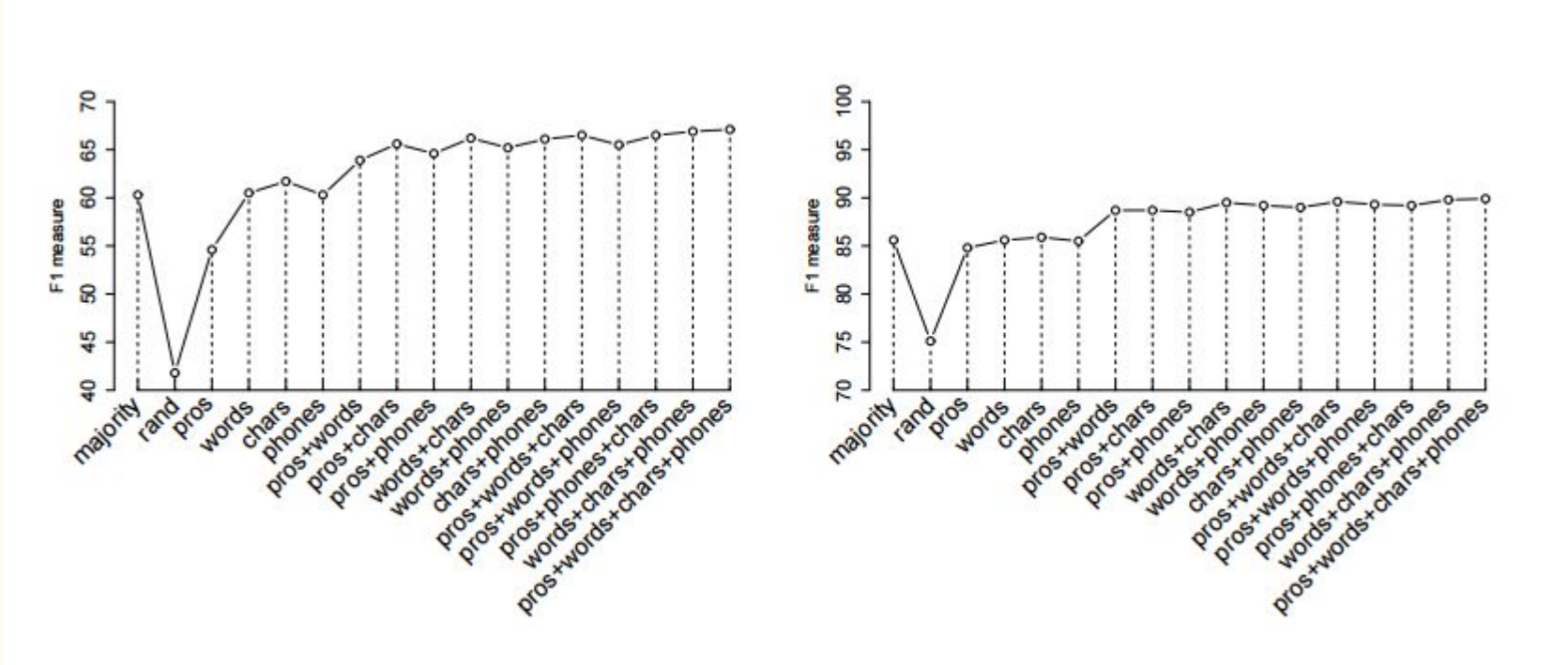
- Two main tasks:
- Recognize subjective utterances
- Discriminate between positive and negative utterances

Subjective Utterances	**Subjective Questions**
Positive subjective	Positive subjective question
Negative subjective	Negative subjective questions
Positive and negative subjective	General subjective question
Uncertainty	**Objective Polar Utterances**
Other subjective	Positive objective
Subjective fragment	Negative objective

Raaijmakers et al: Method and Feature Structure

- Utilize the BoosTexter machine learning algorithm to train multiple classifiers, and investigate combinations of the following features:
  - Word n-grams
  - Prosody (PROS) feature
    - Features based on pitch, intensity, and distribution
  - Phoneme n-grams
  - Character n-grams
    - “This cat” -> {“#Th”,”Thi”,”his”,”is#”,”s#c”,”#ca”,”cat”,”ta#}
    - Captures stemming and other information

Raaijmakers et al: Results



Other Approaches to Sentiment Analysis:

Can prosody inform sentiment analysis? Experiments on short spoken reviews. Mairesse et al, 2012

- Utilized short spoken reviews and online text to classify subjectivity
- Data sparsity problems
- Showed that, in the absence of annotated data, prosody can help with noise from ASR errors

Other Approaches to Sentiment Analysis:

Sentiment analysis of online spoken reviews, Perez-Roasa and Mihalcea, 2013

- Utilized short reviews collated from online sources
- Showed ASR had an impact on the quality of the score
- Concluded spoken and written reviews different

Other Approaches to Sentiment Analysis:

A cross-corpus study of subjectivity identification using unsupervised learning, Wang and Liu, 2011

- Unsupervised learning method (Calibrated EM) vs Supervised Learning Method (Naive Bayes)
- Three different domains (movie data, news, meeting dialog)
- Compared unsupervised to supervised methods by genre
- Gained improvements on genres differently
  - Movies had improvements over supervised methods
  - News had improvements, but less dramatic than movies
  - Meeting dialogs had no improvements over supervised methods

Sentiment and Subjectivity Conclusions

- Linguistic features can be utilized to classify subjectivity relatively well in spoken dialog exchanges
- Character n-grams can be useful features in NLU tasks
- Prosody isn't as informative about subjectivity as I anticipated
  - However, prosody can help alleviate ASR errors
- Written subjectivity is expressed differently than spoken subjectivity
- Genre can have a large effect on system performance

DEEP LEARNING FOR DIALOG SYSTEMS

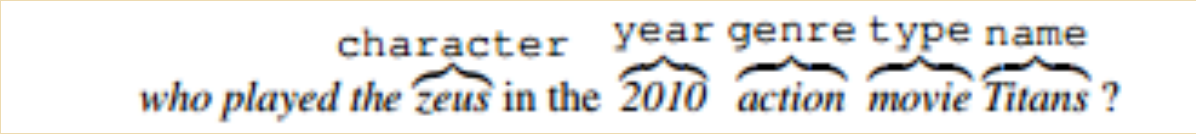
-Lopez G G

UNIVERSITY *of* WASHINGTON

Goal

Paper: Enriching Word Embeddings Using Knowledge Graph for Semantic Tagging in Conversational Dialog Systems.

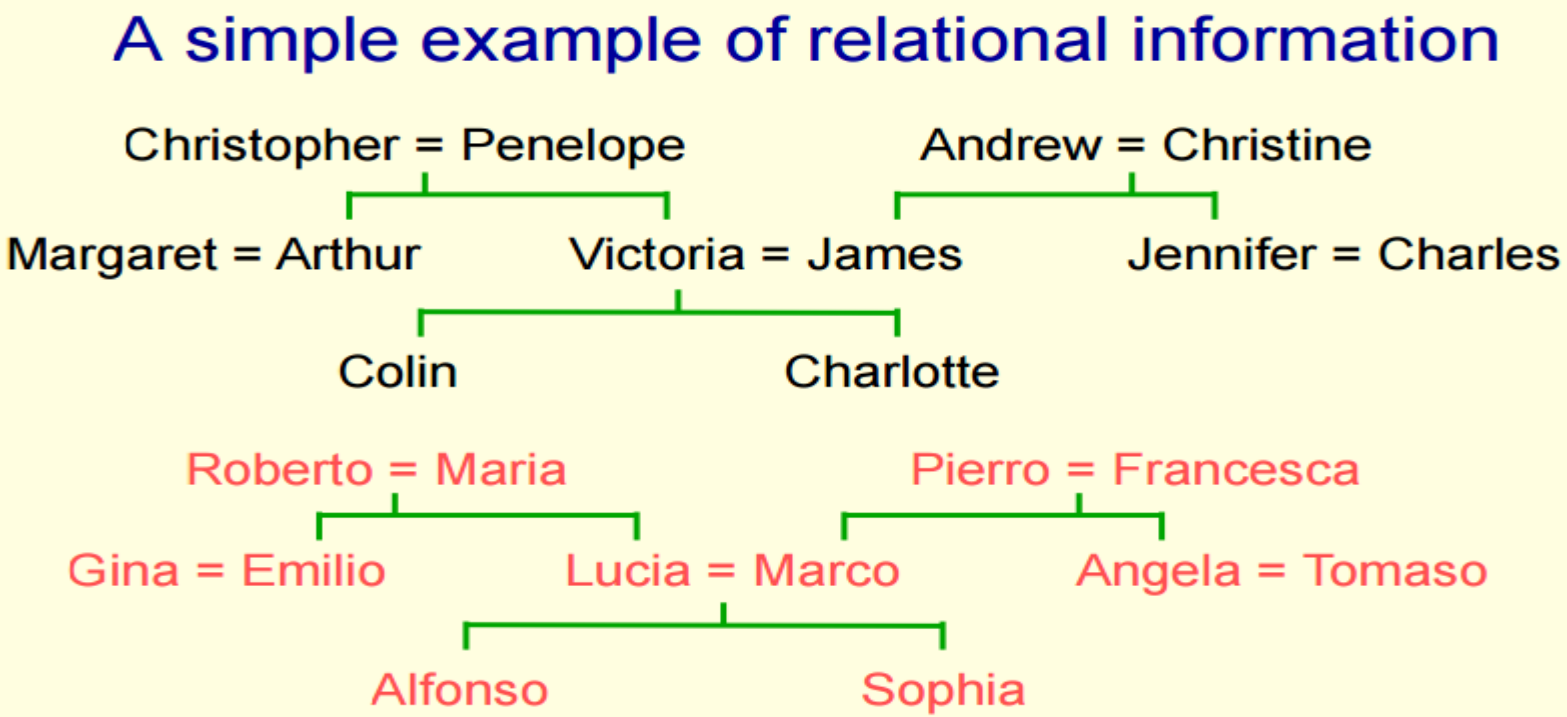
- **Understand a deep learning technique for semantic tagging**
- **Semantic Tagging:**



Neural Net : An overview of 2 types

- **RTM: Relational Learning Task**
- **CBOW : Probabilistic language model (context based)**

RTM: Relational Learning Task

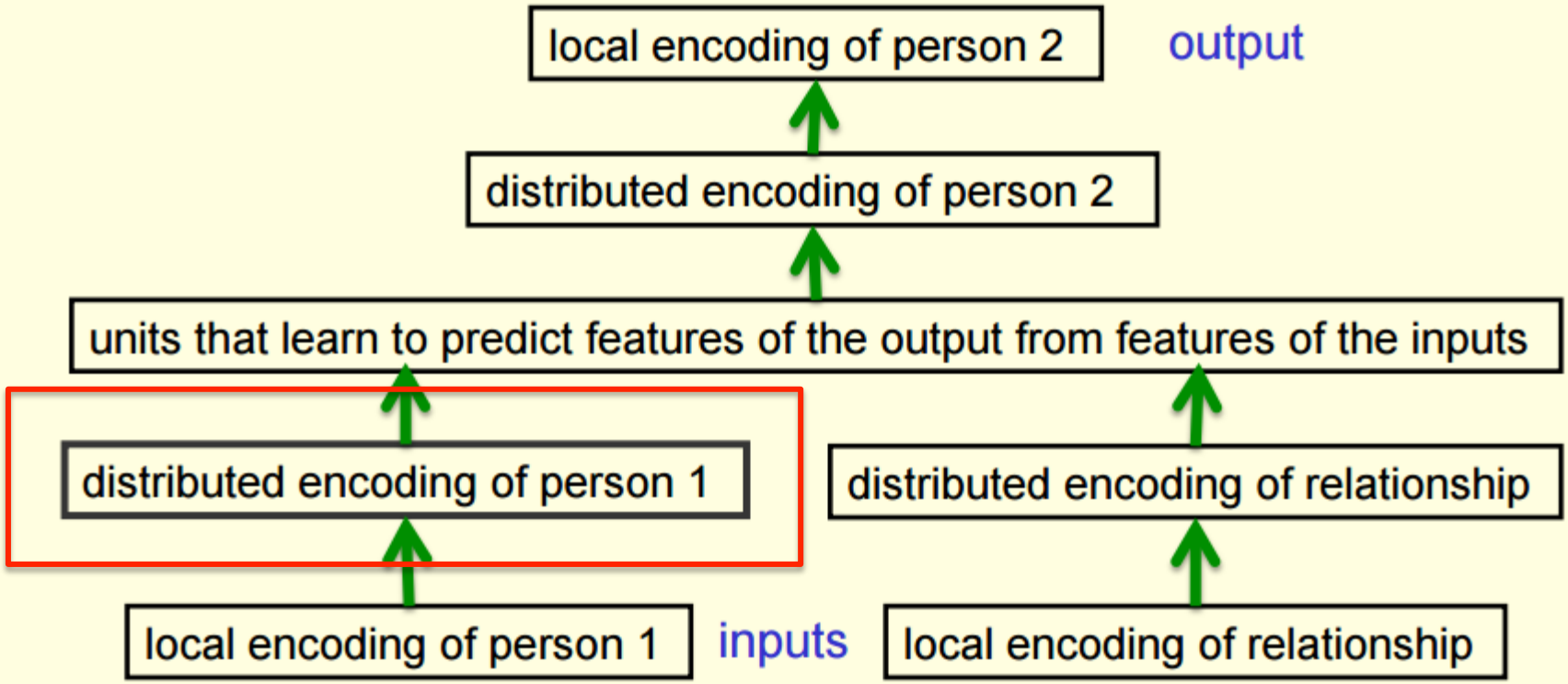


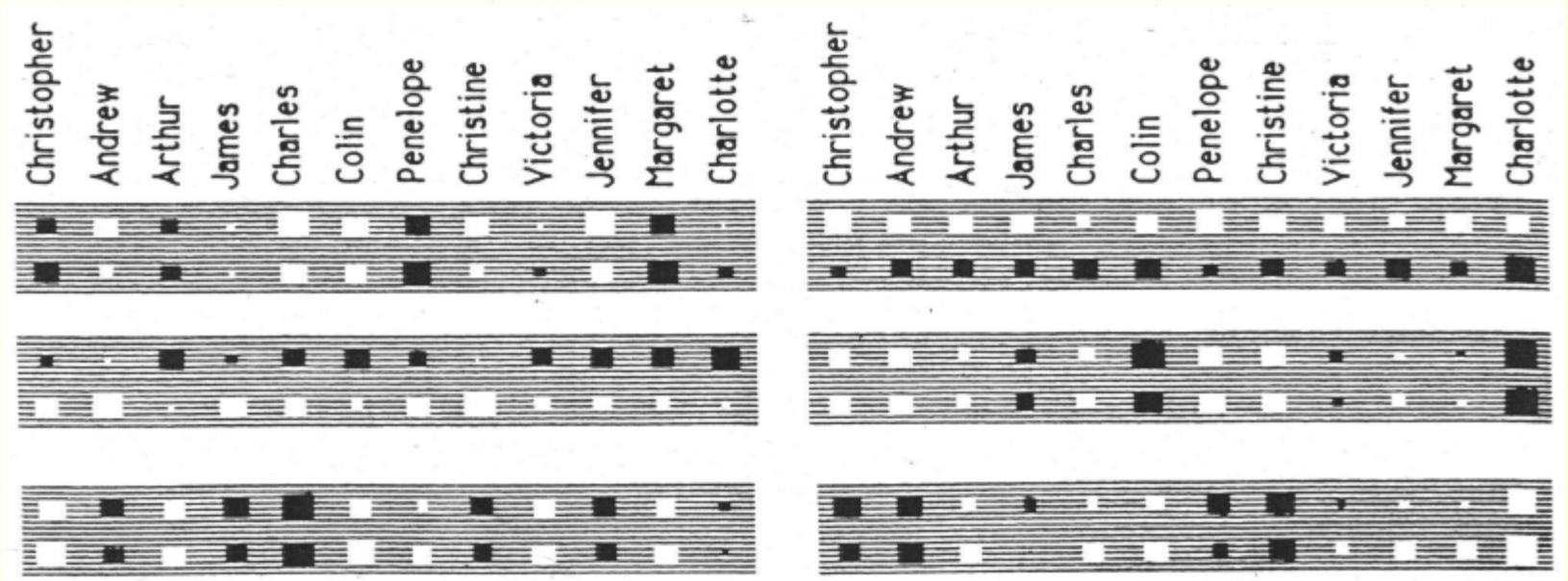
Hinton's slide (https://d396qusza40orc.cloudfront.net/neuralnets/lecture_slides/lec4.pdf)

Another way to express the same information

- Make a set of propositions using the 12 relationships:
  - son, daughter, nephew, niece, father, mother, uncle, aunt
  - brother, sister, husband, wife
- (colin has-father james)
- (colin has-mother victoria)
- (james has-wife victoria) this follows from the two above
- (charlotte has-brother colin)
- (victoria has-brother arthur)
- (charlotte has-uncle arthur) this follows from the above

The structure of the neural net





Christopher = Penelope

Andrew = Christine

Margaret = Arthur

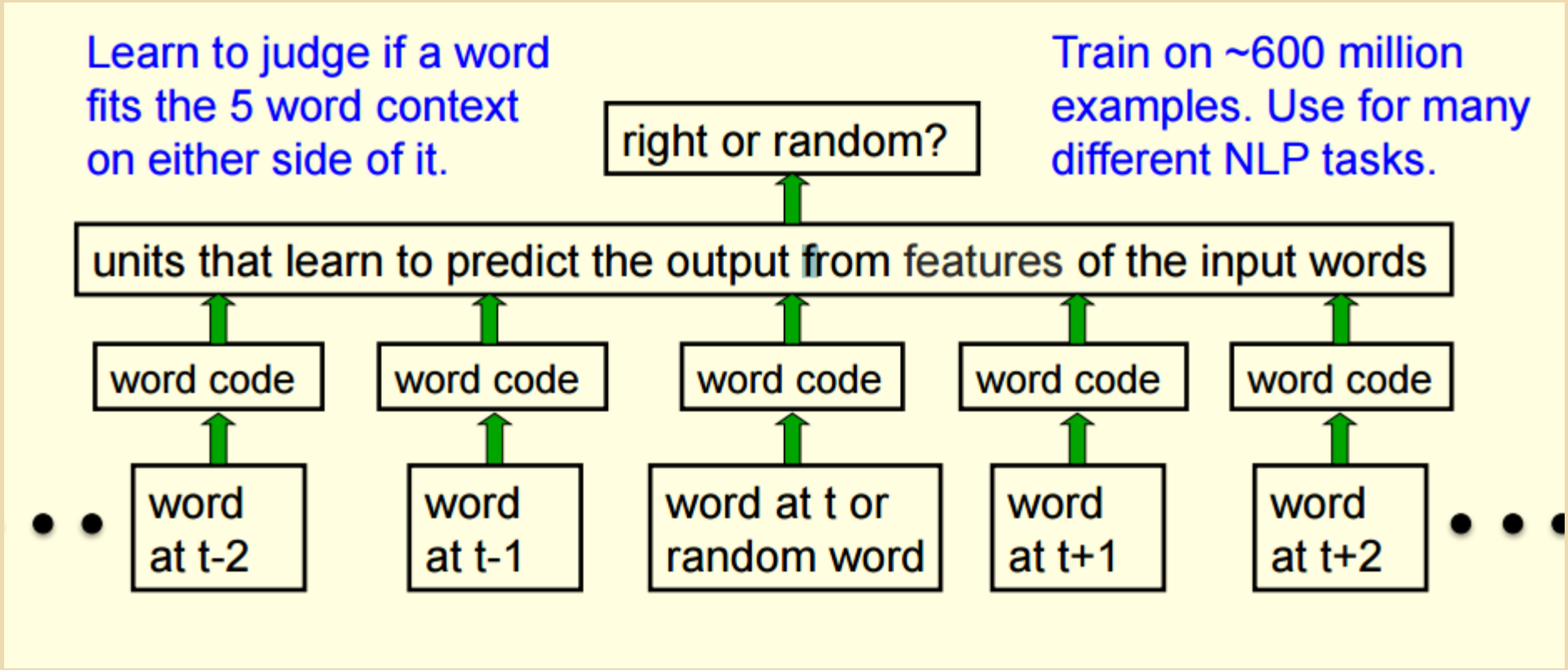
Victoria = James

Jennifer = Charles

Colin

Charlotte

CBOW: Probabilistic language model : Mostly (Context based)



CBOW mod:

CBOW

$$\sum_{t=1}^{T} \log p(w_t | w_{t-c}^{t+c})$$

CBOW mod

$$\sum_{t=1}^{T} \log p(w_t | w_{t-c}^{t+c}, e)$$

Current Paper:

- **Word Embedding = Arg Max (CBOW mod + (Some_Regularization * RTM))**
- **CBOW mod = CBOW with conditional dependency on an entity**

Current Paper Overview

- **Obtain word embedding vectors based on the model just described**
- **Convert them to feature classes based on K-means clustering**
- **Use CRF on these feature classes to tag**
- **Claim 2% improvement in F-score**

Advantages of Word Embedding

- **Dense encoding of words unlike one hot encoding**
- **More robust and resilient to noise or incorrect training data**
- **Captures semantic and syntactic features**

Advantages of CRF

Based on “Is it time to switch to Word Embedding and Recurrent Neural Networks for Spoken Language Understanding?”

- **Demonstrated**
  - **Word embeddings are better than ordinary features**
  - **CRF with normal features is better than embedding with RNN**
- **Did not know to convert word embeddings to features for CRF which current paper does.**

Shortcomings of the current paper

- **Need additional information on the clustering and feature creation**
- **High level overview : sparing in details**

Feature creation :

Based: Bootstrapping Dialog Systems with Word Embedding

- **Feature creation: Provides alternate way to creating features from word vectors.**
- **Combines word count and uses a special Extrema function to create vectors from words in a sentence**

The End !! ☺

ON-BRAND STATEMENT

FOR GENERAL USE

> **What defines the students and faculty of the University of Washington? Above all, it's our belief in possibility and our unshakable optimism. It's a connection to others, both near and far. It's a hunger that pushes us to tackle challenges and pursue progress. It's the conviction that together we can create a world of good. And it's our determination to Be Boundless. Join the journey at uw.edu.**

THIS POWERPOINT THEME

- **A UW color palette is built into this theme.**
- **There are three layout styles and three designs in this theme: Purple, Gold and White**

- **The graphic elements, like the bar and the logos are in the Master Sheets. To edit them go to view > master > slide master.**

Joint Model (Yu M and Dredze ,2014)

> **Joint Model = CBOW + (Some_Regularization * RTM)**

$$\max \frac{1}{T}\sum_{t=1}^{T}\log p(w_t|w_{t-c}^{t+c}) + \frac{C}{N}\sum_{i=1}^{N}\sum_{w\in R_{w_i}}\log p(w|w_i) \quad (4)$$