An Annotated Corpus of Film Dialogue for Learning and Characterizing Character Style

Marilyn Walker, Grace Lin, Jennifer Sawyer
{maw, glin, jsawyer}@soe.ucsc.edu

Natural Language and Dialogue Systems Lab
University of California, Santa Cruz, U.S.A.
http:\\nlds.soe.ucsc.edu\software

Introduction

Interactive Narrative often involves interacting with virtual agents. **Most character dialogue are hand-crafted** → authoring bottleneck

- *Heavy Rain* (2010): ~2000 pages
- *LA Noire* (2011): ~2200 pages
- *Star Wars: The Old Republic*: ~ 40+ novels

Basis of our work: To help authors with their creative process by using **Expressive Natural Language Generation** (ENLG). This process focuses on stylistic, social aspects of the linguistic behavior. We would like to learn character models through **film dialogue**.

Why film screenplays?

- Authored to deliberately convey the feelings, thoughts, and perceptions of the film character.
- Can examine an operationalization of archetype by looking at dialogue of heroes, villains, wise men, etc.

The Corpus

- 862 film scripts from IMSDB (Internet Movie Script Database, www.imsdb.com) as of May 19, 2010
- 7,400 film characters
- 664,000 lines of dialogue
- 9,599,000 tokens
- Ontological Features from IMDB (Internet Movie Database, www.imdb.com)

Feature	Categories
Genre	drama, thriller, crime, comedy, action, romance, adventure
Gender	male, female
Film year	year>2000, 1995>year<=2000, 1990>year<=1995, 1985>year<=1990, 1980>year<=1985, older
Film Director	WesCraven, Steven Spielberg, Stanley Kubrick, Ridley Scott, Steven Soderbergh, Alfred Hitchcock, James Cameron, Martin Scorsese,Quentin Tarantino, etc.

- Automatically Annotated Linguistic Features (examples below)

Feature (Set)	Feature Description
Basic	Number of sentences, sentences per turn, number of verbs per sentence, etc.
Polarity	Use *SentiWordNet 3.0* on all available words
Dialogue Act	Trained with *NPS Corpus* with 15 dialogue act types (e.g., Accept, Bye, Clarify)
Passive Sentence	3rd party software to detect passive sentences
LIWC Word Categories	Use *Linguistic Inquiry and Word Count* (LIWC) text analysis software to categorize words (e.g., positive emotion category words: love, sweet, nice,
Tag Question	Regular expression
Verb Strength	Averaged sentiment values of verbs

Future Augmentation

- TV series: investigate characters scripted by different authors
- More thoroughly evaluate the accuracy of our automatically generated annotations

Film Character Dialogue Stylistic Differences

Annie Hall: Sports Club

Uh … you-you wanna lift?

Turning and aiming her thumb over her shoulder
Oh, why-uh … y-y-you gotta car?

No, um … I was gonna take a cab.

Laughing
Oh, no, I have a car.

You have a car?

Annie smiles, hands folded in front of her.

So … *Clears his throat.*
I don't understand why … if you have a car, so then-then wh-why did you say "Do you have a car?" … like you wanted a lift?

Theme:
How to get a Ride

Genre:
Drama vs. Action

The Terminator 2: Cigar Biker scene

I need your clothes, your boots, and your motorcycle.

You forgot to say please.

Notice linguistic stylistic differences

- Number of dialogue turns
- Sentence structure
- Length of sentences
- Stuttering
- Pauses
- … etc.

Fight scene!

Corpus Application : Learning Character Models

Film Corpus (original character dialogue) → Extract Features → Learned Character Models → Map significant features → ENLG Engine Input Parameters → Apply story domain (e.g., *SpyFeet*) → Generated Dialogue → Evaluation (see Ref. 3)

Annie's original dialogue sample

- H'm? That's, uh … that's pretty serious stuff there. Yeah? Yeah? M'hm? M'hm. Yeah. U-huh.
- Hi. Hi, hi. Well, bye. Oh, yeah? So do you. Oh, God, whatta-whatta dumb thing to say, right?

See corpus description above
Various features

Learn character models from significant features

- Z-scores
- Classification

Annie from *Annie Hall*

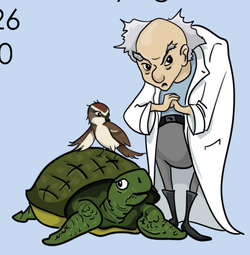
Mapping significant features
Verbosity = *f*(# of sentences per turn, words per sentence)
Content Polarity = *f*(LIWC-Posemo, LIWC-negemo)
…etc.

Methods of mapping

- Average
- Weighted average
- …etc.

PERSONAGE parameters for Annie
Verbosity=0.78
Content polarity=0.72
Repetition polarity=0.79
Concessions=0.83
Concessions Polarity=0.26
Positive content first=1.00
First Person in Claim=0.6
Claim Polarity=0.57
Claim Complexity=0.31
etc.

SpyFeet
Exercise Role-Playing Game



Generated dialogue

- Come on, I don't know, do you? People say Cartmill is strange while I don't rush to um.. judgment.
- I don't know. I think that you brought me cabbage, so I will tell something to you, alright?
- I see. I am not sure. Obviously, I respect Wolf. However, he isn't my close friend, is he?

References

1. Grace I. Lin and Marilyn A. Walker. (2011). *All the World's a Stage: Learning Character Models from Film*. Artificial Intelligence and Digital Entertainment (AIIDE) 2011.
2. Marilyn A. Walker, Ricky Grant, Jennifer Sawyer, Grace Lin, Noah Wardrip-Fruin and Michael Buell. *Perceived or Not Perceived: Film Character Models for Expressive NLG*. In Interactive Storytelling: Fourth Joint Conference on Interactive Digital Storytelling, ICIDS 2011.
3. Marilyn A. Walker, Grace Lin, Ricky Grant, Jennifer Sawyer, Noah Wardrip-Fruin and Michael Buell. *Murder in the Arboretum: Comparing Character Models to Personality Models*. In 4th International Workshop on Intelligent Narrative Technologies 2001.

Natural Language and Dialogue System Lab

Baskin Engineering UC SANTA CRUZ