# Modelling Protagonist Goals and Desires in First-Person Narrative

Elahe Rahimtoroghi, Jiaqi Wu, Ruimin Wang,
Pranav Anand and Marilyn A Walker

University of California Santa Cruz

NSF IIS 1302668, Nuance SC14-74, NSF HCC 1321102

# Motivation



- Humans appear to organize and remember everyday experiences by imposing a narrative structure on them
- Thus many genres of natural language exhibit narrative structure
- Widely agreed that narrative understanding requires
  - Modelling protagonist goals
  - Tracking their outcomes
- First person social media stories full of expressions of desires and outcome descriptions

First-person blog story

# First Person Story: Live Journal

I dropped something and it was dark, he bent with his cell phone light to help me look for it. We spoke a little, but it was loud and not suited for conversation there.
**I had hoped to ask him to join me for a drink or something after the show** (if my courage would allow such a thing)
but he left before the end and I didn't see him after that. Maybe I'll try missed connections lol.

# Goal

- Identify goal and desire expressions in first-person narratives
  - E.g. "*had hoped to*"
- Infer from the surrounding text whether the desire is
  - Fulfilled
  - Unfulfilled

| Pattern |
| --- |
| wanted to |
| needed to |
| ordered |
| arranged to |
| decided to |
| hoped to |
| couldn't wait |
| wished to |
| scheduled |
| asked for |
| required |
| requested |
| demanded |
| ached to |
| aimed to |
| desired to |
| Total |

# First Person Story: Live Journal



I dropped something and it was dark, he bent with his cell phone light to help me look for it. We spoke a little, but it was loud and not suited for conversation there.
**I had hoped to ask him to join me for a drink or something after the show** (if my courage would allow such a thing)
but he left before the end and I didn't see him after that.
Maybe I'll try missed connections lol.

**Unfulfilled**

# First Person Desires



People did seem pleased to see me but all I **[wanted to]** do was talk to a particular friend.

I **[wished to]** meet new people and to get out of my own made misery and turn myself into a more sociable human being for the sake of mental health alone.

I'm off this weekend and had really **[hoped to]** get out and dance.

We **[decided to]** just go for a walk and look at all the sunflowers in the neighborhood.

I **[couldn't wait to]** get out of our cheap and somewhat charming hotel and show James a little bit of Paris.

We drove for just over an hour and **[aimed to]** get to Trinity beach to set up for the night.

She called the pastor, and he had time, too, so, we **[arranged to]** meet Saturday at 9am.

Even though my deadline wasn't until 4 p.m., I **[needed to]** write the story as quickly as possible.

# First Person Story: Live Journal

I dropped something and it was dark, he bent with his cell phone light to help me look for it. We spoke a little, but it was loud and not suited for conversation there.
**I had hoped to ask him to join me for a drink or something after the show** (if my courage would allow such a thing)
but he left before the end and I didn't see him after that.
Maybe I'll try missed connections lol.

**Unfulfilled**

# Related Work



Computational model of Lehnert's plot units (Goyal and Riloff, 2013)

- Identify and track affect states to model plot units
- **Dataset:** Aesop's fables
- **Manual annotation:** examine different types of affect expressions in the narratives
- Affect states arise from the expression of goals and their outcomes

Model desire fulfillment (Chaturvedi et al., 2016)

- **MCTest:** 660 crowd-sourced stories understandable by 7-year olds
- **SimpleWiki:** Simple English Wikipedia
- **Desire statements:** matching three verb phrases (*wanted to, hoped to*, and *wished to*)
- **Context:** five or fewer sentences following the desire expression

# Contributions

- New Corpus: DesireDB
  - 3,500 first-person informal narratives with annotations
  - Download: https://nlds.soe.ucsc.edu/DesireDB
- Modeling goals and desires
  - Classification models: Predict desire fulfillment status
  - Feature analysis
  - Study the effect of prior and post context in predicting desire fulfillment
  - Compare to previous models and datasets

# DesireDB Corpus



1. Subset of the Spinn3r corpus (Burton et al 2009, 2011)
   - First-person narratives from personal blog domains
   - First-person protagonist easily tracked throughout
2. Systematic method to identify desire and goal statements
   - Collect context before and after
3. Annotations
   - Create gold-standard labels for fulfillment status
   - Mark spans of text as evidence

# Patterns for Desires and Goals



- Many different linguistic ways to express desires
- FrameNet 1.7: Needing, Offer, Purpose, Request, ...
- Frequent and high-precision representative instances in English Gigaword
- 37 verbs ⇒ constructed past forms

| Pattern |
| --- |
| wanted to |
| needed to |
| ordered |
| arranged to |
| decided to |
| hoped to |
| couldn't wait |
| wished to |
| scheduled |
| asked for |
| required |
| requested |
| demanded |
| ached to |
| aimed to |
| desired to |
| Total |

# Data Collection

- Extracted 600K stories containing the verbal patterns of desire
  - Desire-Expression Sentence
  - Prior-Context (Labov & Waletzky, 67)
  - Post-Context

**Prior-Context**: (1) I ran the Nike+ human Race 10K new York in under 57 minutes! (2) Then at the all-American rejects concert, I somehow ended up right next to this really cute guy and he seemed interested in me. (3) Was I imagining things? He was really nice; (4) I dropped something and it was dark, he bent with his cell phone light to help me look for it. (5) We spoke a little, but it was loud and not suited for conversation there.

**Desire-Expression-Sentence**: I **[had hoped to]** ask him to join me for a drink or something after the show (if my courage would allow such a thing) but he left before the end and I didn't see him after that.

**Post-Context**: (1) Maybe I'll try missed connections lol. (2) I didn't want to tell him I think he's cute or make any gay references during the show because if I was wrong that would make standing there the whole rest of the concert too awkward... (3) Afterward, I wandered through the city making stops at several bars and clubs, met some new people, some old people (4) As in people I knew - I actually didn't met any old people, unless you count the tourist family whose dad asked me about my t-shirt. (5) And when I thought the night was over (and the doorman of the club did insist it was over) I met this great guy going into the subway.

# Annotations



- Sample 3,680 instances for annotation
  - 16 verbal patterns
  - Sample skewed as per distribution in the original dataset

- Mechanical Turk
  - Specified the desire expression verbal pattern using square brackets in the data
  - 3 Prequalified workers per instance
  - Label the desire expression sentence based on the prior and post context: **Fulfilled, Unfulfilled, and Unknown** from the context
  - Mark a **span of text as the evidence** for the label they had chosen

# Creating Gold-Standard Data



- Total agreement rate
  - Fulfilled → 75%
  - Unfulfilled → 67%
  - Unknown from the context → 41%
- Marking evidence
  - 79% of the data all three annotators marked overlapping spans

| Pattern | Count | Ful | Unf | Unk | None |
|---|---|---|---|---|---|
| wanted to | 2,510 | 49% | 35% | 14% | 2% |
| needed to | 202 | 65% | 16% | 16% | 3% |
| ordered | 201 | 71% | 21% | 6% | 2% |
| arranged to | 199 | 68% | 13% | 16% | 3% |
| decided to | 68 | 87% | 9% | 4% | 0% |
| hoped to | 68 | 19% | 68% | 12% | 1% |
| couldn't wait | 68 | 79% | 3% | 15% | 3% |
| wished to | 66 | 27% | 35% | 30% | 8% |
| scheduled | 60 | 43% | 25% | 27% | 5% |
| asked for | 60 | 53% | 27% | 15% | 5% |
| required | 58 | 69% | 16% | 15% | 0% |
| requested | 30 | 60% | 20% | 20% | 0% |
| demanded | 30 | 60% | 23% | 17% | 0% |
| ached to | 20 | 50% | 40% | 10% | 0% |
| aimed to | 20 | 55% | 30% | 15% | 0% |
| desired to | 20 | 50% | 25% | 25% | 0% |
| Total | 3,680 | 53% | 31% | 14% | 2% |

# DesireDB Data Instance



- Testbed for modeling desires in personal narrative and predicting their fulfillment
  - Open domain first-person narratives
  - Prior and post context
  - Reliable annotations
  - Download: https://nlds.soe.ucsc.edu/DesireDB

**Data-Instance:**
Prior-Context: ConnectiCon!!! Ya baby, we did go this year as planned! Though this year we weren't in the artist colony, so I didn't see much point in posting about it before hand.
Desire-Expression-Sentence: This year we [wanted to] be part of the main crowd.
Post-Context: We wanted to get in on all the events and panels that you cant attend when watching over a table. And this year we wanted to cosplay! My hubby and I decided to dress up like aperture Science test subjects from the PC game portal. It was a good and original choice, as we both ended up being the only portal related people in the con (unless there were others who came late in the evening we didn't see) It was loads of fun and we got a surprising amount of attention.

**Annotations:**
Fulfillment-Label: Fulfilled
Fulfillment-Agreement-Score: 3
Evidence: Though this year we weren't in the artist colony. We wanted to get in on all the events and panels that you cant attend when watching over a table.
Evidence-Overlap-Score: 3

# Modeling Desire Fulfillment



- Define feature sets motivated by narrative structure
  - Some features motivated by prior work
- Classification experiments
  - LSTM models to generate sentence embeddings
  - Three-layer RNN classifier
  - Feature analysis
  - Explore using different parts of context
  - Comparison to previous work (both models and datasets)

# Desire Expression Features

- Properties of the desire expression
- Desire verb pattern
- Focal words and their synonym/antonym mentions in the context
- Desire subject and its mentions
- First-person subject

Eventually, I just **decided to speak**, and I can't even remember what I **said**, but people were very happy and proud of me for **saying** what I wanted to **say**.

# Discourse Features

Discourse relation markers in the Penn Discourse Treebank as:

- **Violated-Expectation** (31): e.g. *although, rather, yet, but*
- **Meeting-Expectation** (15): *accordingly, so, ultimately, finally*
- **Neutral : none of these appear**

I wanted to regroup and prepare for battle **so** I laid him down while I pseudo relaxed.

I wanted to do visual editing and management very much, **but** one of the core courses is such that it requires a prereq.

# Features

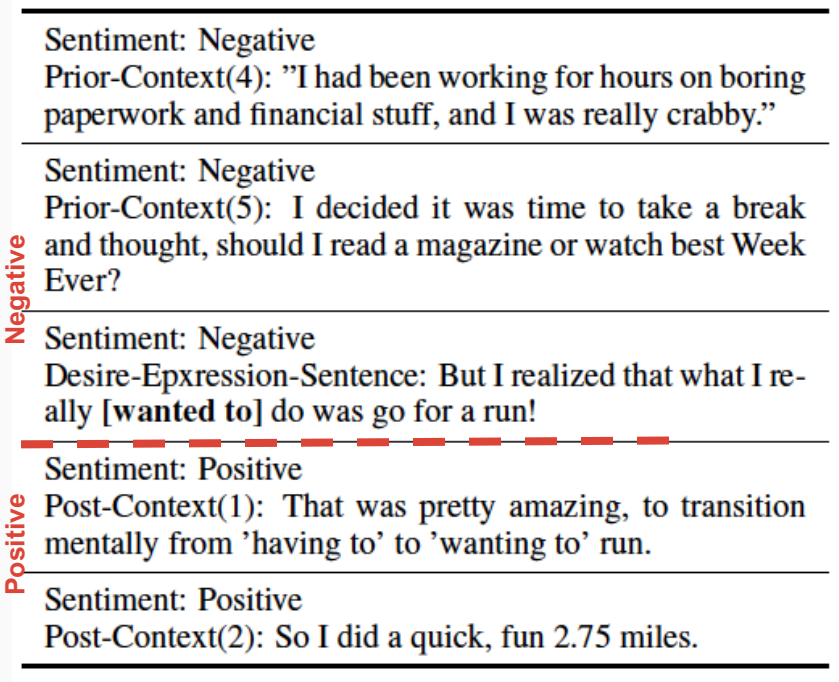


- Connotation Features
  - Connotation Lexicon (Feng et al., 2013)
  - Polarity agreement of context words with focal words
    - Connotation-Agree
    - Connotation-Disagree

So, we decided to **go** together and **play** backgammon in between loads.I usually **win** at the **game** and today was no exception. We had a **fine** time.

# Sentiment Flow Features

- Detect a change in sentiment in the surrounding context
- Could be the mention of a thwarted effort or a victory
  - Sentiment-Agree
  - Sentiment-Disagree



# Classification Models



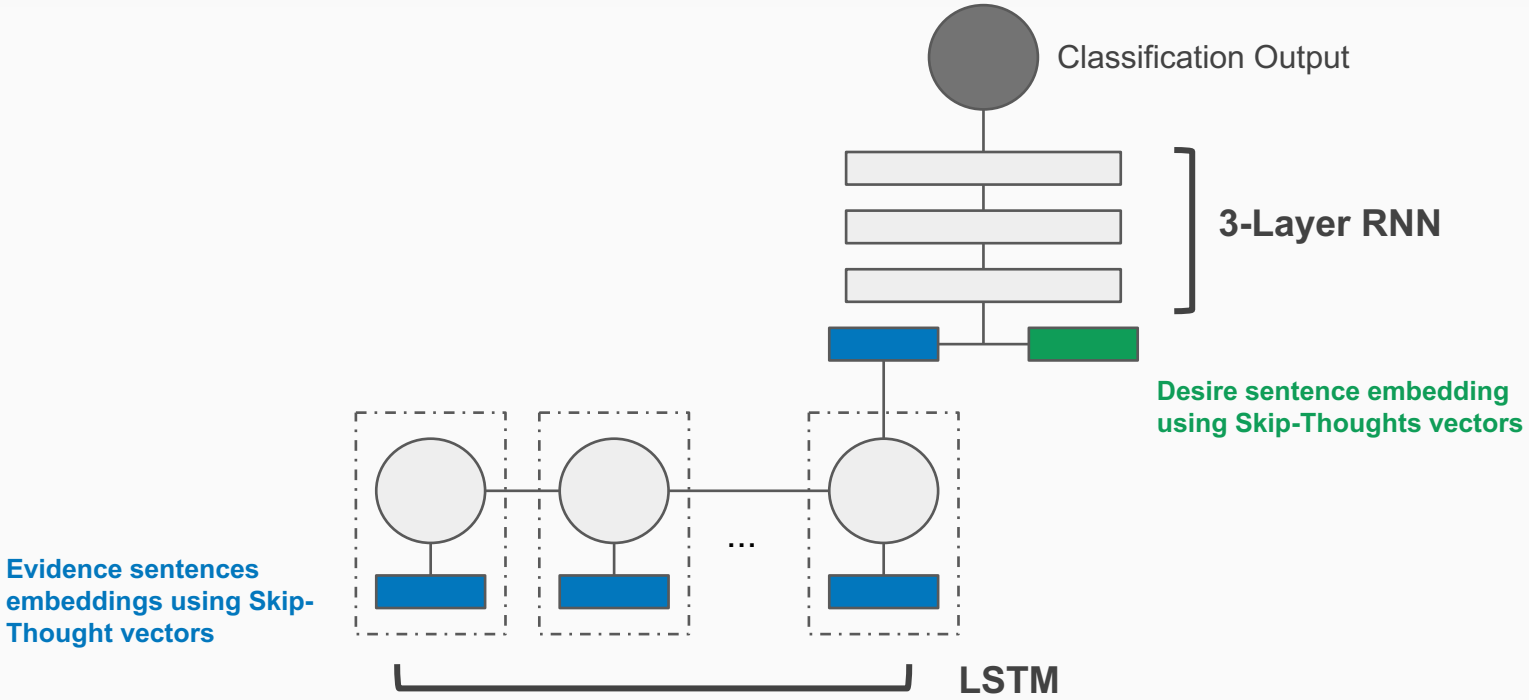
- Four types of features
  - Motivated by narrative characteristics
  - Ablation experiments
- Method
  - LSTM for sentence embeddings
  - Three-layer RNN for classification
    - Suitable for sequence learning
    - Encode the order of the sentences to distinguish between prior and post context

# Generating Sentence Embeddings



- Two approaches
  - **Skip-Thought:** Using pre-trained skip-thought model (Kiros et al., 2015) as the embeddings
    - Concatenates features, if any, with embeddings
    - Uses LSTM to generate a single representation
  - **CNN-RNN:** Using 1-dimensional convolution with max-over-time pooling introduced (Kim, 2014) to generate the sentence embedding
    - Uses LSTM to generate a single representation
    - Used Google News Vectors (Mikolov et al., 2013) for word embedding

# LSTM with Skip-Thoughts Embeddings & RNN Classifier



# Experiments



- A subset of DesireDB: **Simple-DesireDB**
  - In order to compare more directly to previous work
  - Five verbal patterns
    - wanted to, hoped to, wished to, couldn't wait to, decided to
  - Fulfilled: 1,366
  - Unfulfilled: 953
  - Train (1,656), Dev (327), and Test (336) sets

# Classification Experiments: Study the Robustness of Features

- BOW: Bag of Words features
- ALL: All four sets of features

| Method | Features | Ful-P | Ful-R | Ful-F1 | Unf-P | Unf-R | Unf-F1 | Precision | Recall | F1 |
|---|---|---|---|---|---|---|---|---|---|---|
| Skip-Thought | BOW | 0.75 | 0.70 | 0.72 | 0.54 | 0.61 | 0.57 | 0.65 | 0.65 | 0.65 |
| | **ALL** | **0.80** | **0.71** | **0.75** | **0.59** | **0.70** | **0.64** | **0.70** | **0.70** | **0.70** |
| CNN-RNN | BOW | 0.75 | 0.73 | 0.74 | 0.57 | 0.60 | 0.58 | 0.66 | 0.66 | 0.66 |
| | ALL | 0.75 | 0.79 | 0.77 | 0.61 | 0.56 | 0.59 | 0.68 | 0.68 | 0.68 |

# Effects of Prior and Post Context



- Using best-performing model
  - Skip-Thoughts embeddings
  - ALL features
- Adding features from prior context alone improves the results

| Data | Ful-P | Ful-R | Ful-F1 | Unf-P | Unf-R | Unf-F1 | Precision | Recall | F1 |
|---|---|---|---|---|---|---|---|---|---|
| Desire | 0.74 | 0.75 | 0.75 | 0.57 | 0.56 | 0.57 | 0.66 | 0.66 | 0.66 |
| Desire+Prior | 0.78 | 0.73 | 0.75 | 0.58 | 0.65 | 0.61 | 0.68 | 0.69 | 0.68 |
| Desire+Post | 0.76 | 0.70 | 0.73 | 0.55 | 0.62 | 0.59 | 0.66 | 0.66 | 0.66 |
| **Desire+Context** | **0.80** | **0.71** | **0.75** | **0.59** | **0.70** | **0.64** | **0.70** | **0.70** | **0.70** |

# Experiments on DesireDB



- Comparing BOW, ALL, and Discourse features (best among 4 sets)
- Baselines: Logistic Regression (best-performing on Dev set), Naive Bayes, and SVM

| Method | Features | Ful-P | Ful-R | Ful-F1 | Unf-P | Unf-R | Unf-F1 | Precision | Recall | F1 |
|---|---|---|---|---|---|---|---|---|---|---|
| Skip-Thought | BOW | 0.78 | 0.78 | 0.78 | 0.57 | 0.56 | 0.57 | 0.67 | 0.67 | 0.67 |
| | All | 0.78 | 0.79 | 0.79 | 0.58 | 0.56 | 0.57 | 0.68 | 0.68 | 0.68 |
| | **Discourse** | **0.80** | **0.79** | **0.80** | **0.60** | **0.60** | **0.60** | **0.70** | **0.70** | **0.70** |
| Logistic Regression | BOW | 0.69 | 0.65 | 0.67 | 0.53 | 0.57 | 0.55 | 0.61 | 0.61 | 0.61 |
| | All | 0.79 | 0.70 | 0.74 | 0.52 | 0.64 | 0.58 | 0.66 | 0.67 | 0.66 |
| | Discourse | 0.75 | 0.84 | 0.80 | 0.60 | 0.45 | 0.52 | 0.67 | 0.65 | 0.66 |

# Identifying Unfulfillment is Harder



- Similar features and methods achieve better results for the **Fulfilled** class as compared to Unfulfilled
- Same in annotations
  - Data labeled Fulfilled by two annotators
    - 64% labeled Unknown from the context
    - only 36% labeled Unfulfilled
  - Data labeled Unfulfilled
    - 49% labeled Unknown from the context
    - 51% labeled Fulfilled

# Comparison to Previous Work Datasets



- Previous work methods (Chaturvedi et al., 2016)
  - BOW
  - Logistic Regression
  - Structured Model: LSNM (best-performing)
- Our methods:
  - LR with Discourse features
  - Skip-thought embeddings
    - BOW features
    - ALL features

| Dataset | Method | Precision | Recall | F1 |
|---|---|---|---|---|
| MCTest | BOW | 0.41 | 0.50 | 0.45 |
| | Unstruct-LR | 0.71 | 0.63 | 0.67 |
| | LSNM | 0.70 | 0.84 | 0.74 |
| | **Discourse-LR** | **0.63** | **0.83** | **0.71** |
| | **SkipTh-BOW** | **0.72** | **0.68** | **0.70** |
| | **SkipTh-ALL** | **0.70** | **0.84** | **0.76** |
| Simple Wiki | BOW | 0.28 | 0.20 | 0.23 |
| | Unstruct-LR | 0.50 | 0.09 | 0.15 |
| | LSNM | 0.38 | 0.21 | 0.27 |
| | **Discourse-LR** | **0.32** | **0.82** | **0.46** |
| | **SkipTh-BOW** | **0.71** | **0.26** | **0.38** |
| | **SkipTh-ALL** | **0.33** | **0.16** | **0.22** |

**Results on Fulfilled Class**

# Conclusions



- DesireDB corpus: to study goals and their fulfillment in narrative discourse
- Modeling goals fulfillment
  - Features motivated by narrative structure are effective
  - Both prior and post context are useful
- Future work
  - Explore richer features
  - Apply other sequential classification models
  - Explore using evidence data
  - Detecting hypothetical goals (e.g., ‘If I had wanted to buy a book’)

# Thank you!

# Creating Gold-Standard Data



- Three annotators assigned to each data instance
- Majority vote to create gold-standard label
  - Cases with no agreement labeled as ‘None’
- Average Kappa between each annotator and the majority vote: 0.88
- 66% of the data labeled with total agreement
- 32% of data was labeled by two agreements and one disagreement