Creating Knowledge Bases from Text

Alon Halevy

Joint work with Dan Iter, Wang-Chiew Tan + RIT Team

February 9, 2018 -- KBCOM

R.I.T. and Messy Data

Big Gorilla (https://www.biggorilla.org/)

Open-source data integration and preparation in Python

- FrameIt: the topic of this talk
- Koko, an information extraction language: query regex + dependency tree structure

Find entities that are described by the word "delicious" and have an associated action of eat/eating/eaten/ate/drink…

*I drank some **coffee**, which was delicious, and also ate some pie*

Definition: Knowledge Base Construction

Input: a corpus of many short texts, such as:

- Descriptions of happy moments
- Product/hotel reviews
- Open-ended survey answers, transcripts of conversations

Output:

- A set of frames that model some fraction of the texts
- An extractor (i.e., SRL) for every frame

Problem Setting

Apriori, while the user understands the domain, but she doesn’t know what’s in the corpus.

→ System must support exploration and iteration.

There is no attempt to cover *all* the corpus. Just enough to extract enough of the value.

→ best effort scenario. User decides when to stop.

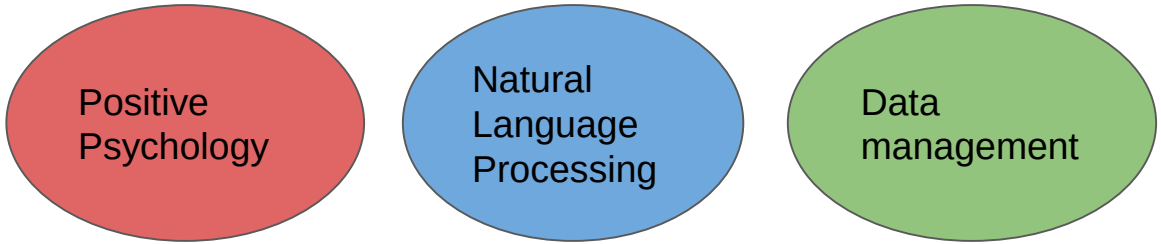
Outline

- Background: technology for happiness
- HappyDB
- The FrameIt System

Technology for happiness

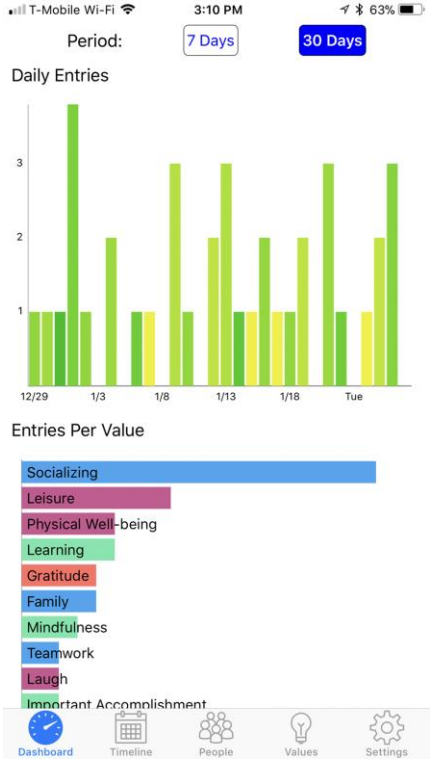
Key question:

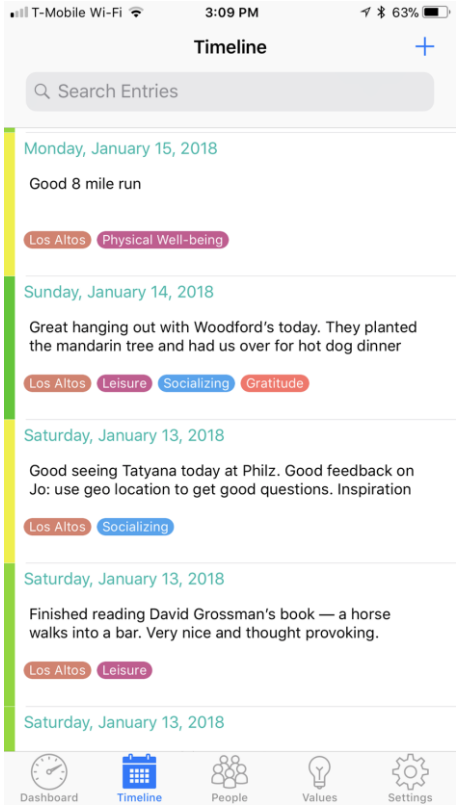
Can we develop technology that makes us happier?



Jo(urnaling) App

- Users journal their significant daily moments.
- App maps them to user's *values* (socializing, learning, family, etc)
- User can reflect.
- Ultimately: Jo should provide insights and advice.





Jo: the Smart Journal (iOS)

HappyDB

HappyDB: A Corpus of 100,000 Crowdsourced Happy Moments. LREC '18.

- Collected 100,000 happy moments by crowdsourcing.
  - Publicly available at **https://rebrand.ly/happydb**

"I went to the park with the kids. The weather was perfect!"

"I quit smoking cigarette since the tax increase of this year here in California. I am hoping to keep it up and improve my health."

"I had dinner with my mom"

"A few weeks ago I received a letter from the President of my University letting me know that I've received tenure and promotion to Associate Professor."

Why Create Structure?

We want to know what's in HappyDB.

Create an "ontology" of happy moments that will enable Jo to provide:

- Intelligent replies
- Activity suggestions
- Better insights into people's activities

Structure = Frames

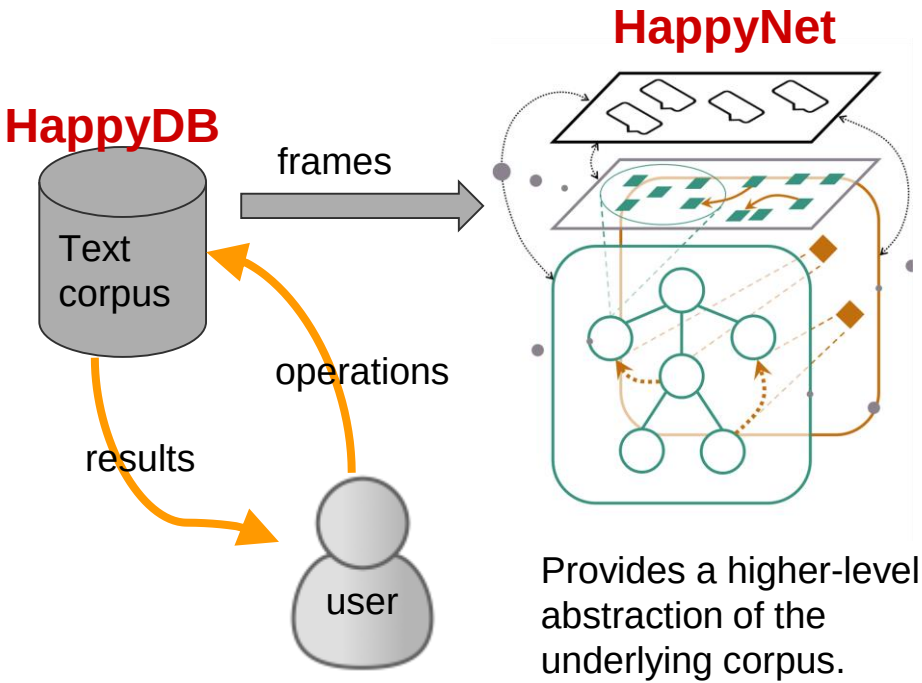
- A *frame* has a name and a set of attributes

Example frame:
Had_meal
Participant:
Meal type:
With whom:

Instance of a frame:
Had_meal
Participant: *I*
Meal type: *breakfast*
With whom: *my mom*

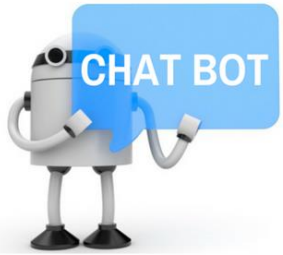
- Frames are user-defined. FrameNet and PropBank are “frame snippets” for our purposes.
- Different from information extraction: looking for a set of triples.
- No need to fill every slot.

The FrameIt System



Provides a higher-level of abstraction of the underlying corpus.

What types of meals do friends tend to have together?

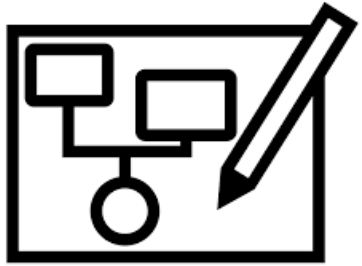


You: “I had lunch with my mom.”
Bot: “What did you have for lunch?

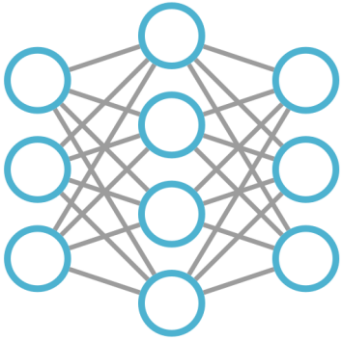
FrameIt Workflow

Explore the corpus

Define frames



Train SRL



The Tasks Supported while Exploring the Corpus

1. Discover topics for frames (meals, promotions, …)
    - Show random sentence
    - Show most frequently occurring words/lemmas
    - Show all sentences containing a word/lemma
2. Decide on level of granularity:
    - Eating a meal, or also preparing a meal?
    - Getting a promotion, or being congratulated for it?

Find Nearest Moments

I had dinner with my mom.

Candidate frame:
Have meal
  Participant:
  Meal type:
  With whom:
  When:

“I had dinner with my mom.”
“I had breakfast with my daughter.”
“I had breakfast with my girlfriend.”
“I had breakfast with my kids.”
“I had dinner with my Dad.”
“I had lunch with my husband.”
“I ate my breakfast with my wife today.”
“I had dinner with my mother.”
“I made dinner with my mom.”
“I had dinner with my aunt.”
“I had dinner with my grandmother.”
“I had dinner with my mom and girlfriend.”
“ I had lunch with my family.”
“I had dinner with my wife.”
“I went to lunch with my mom.”
“I had lunch with my cousin.”
“I had lunch with my wife today.”
“I had a nice lunch with my grandmother.”
“I had a really nice dinner with my mom.”

Leverage FrameNet (using Semaphor)

What are the most common FrameNet frames triggered by a set of sentences?

Food	3,933
Social_even	2,867
Calendric_unit	2,800
Personal_relationship	2,684
Possession	2,327
Ingestion	2,182

Desirability	1,324
Chemical-sense_description	1,130
Measure_duration	839
Getting	816
Apply_heat	748

A Closer Look: The Words Triggering the Frames

Possession

[(had, 1659), (have, 234), (having, 120)]

Ingestions

[(ate, 889), (eat, 381), (eating, 172)]

Causation

[(made, 1052), (Making, 36), (Made, 30)]

Apply_heat

[(cooked, 405), (cook, 111), (baked, 65)]

Leverage FrameNet (using Semaphor)

What are the most common FrameNet frames triggered by a set of sentences?

(Food, 3933),

(Social_event, 2867),

(Calendric_unit, 2800),

(Personal_relationship, 2684),

(Possession, 2327),

(Ingestion, 2182),

The Tasks Supported while Exploring the Corpus (2/2)

3. Which slots to define for the frame?
    - Scope of slots (including holiday meals? coffee?)
4. Create lists or dictionaries for slot values (or at least give the system good hints about them)

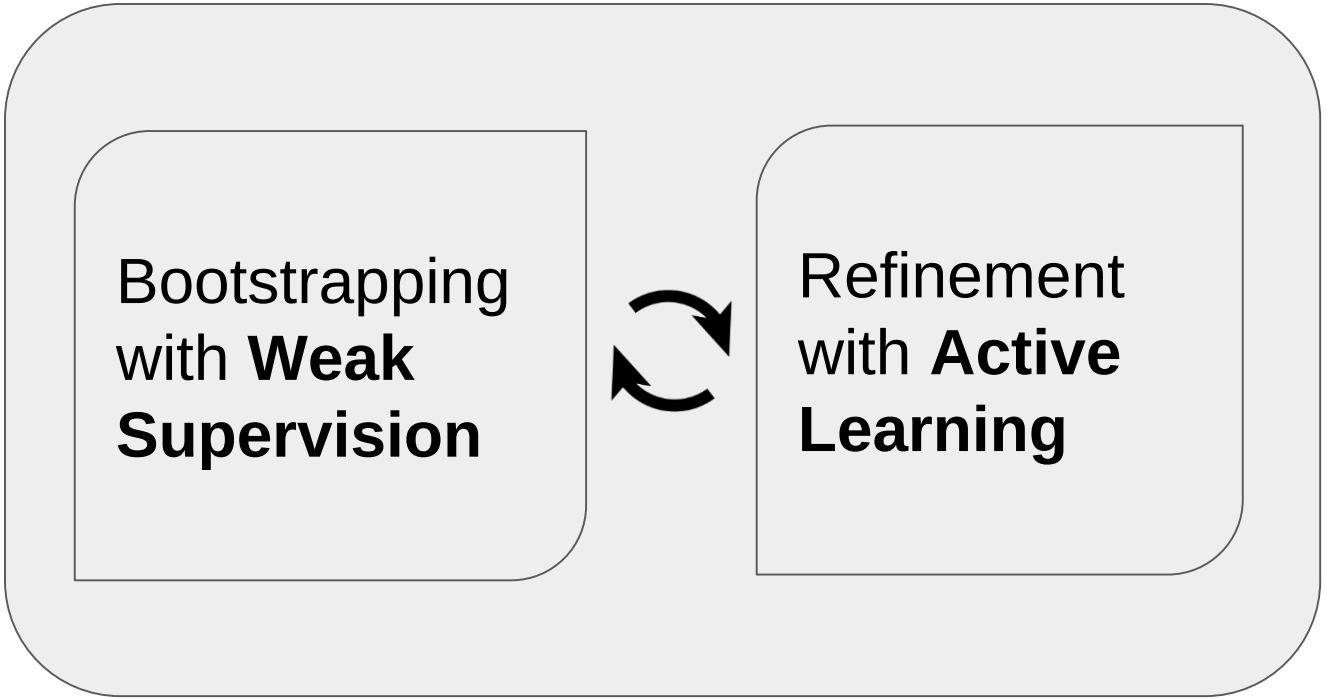
Use WordNet to expand terms: (NELL is also an option)

Dinner → lunch, brunch, breakfast, Seder, …

Mom → dad, sister, brother, cousin, fiance, …

Training SRL Models

Training needs to fit in with FrameIt workflow.



Evaluation of FrameIt

Mostly work in progress.

Corpora: HappyDB, TripAdvisor hotel reviews, ANES 2008 presidential election survey.

- We’re able to quickly create frames that cover a large fraction of the corpus
- The SRL models have high accuracy after relatively short active learning feedback.

We are hiring! (And New Name Coming Soon!)

- Software Engineers.
- Research Scientists.
- Internships (Summer and Spring)
- http://www.recruit.ai

Natural Language Processing, Data Management and Integration, Data Mining, Machine Learning, Knowledge Representation, Bot interfaces, Crowdsourcing, and Visualization.

Extra Slides