Thesis Meeting #4

kense, for the thesis

Recap - Scheduling

- Define and Commit*

- Research & Development

- Production

- Writing

Tasks

- Explore sub-conversation via labeling

  - Conversation Archetypes

- Determine distance metric/similarity measurement (synonymous)

- Find some data

Sub-Conversation

Define a Conversation as a trace, which is in the form:

$$LE = (V, U, T)$$

Where V is a set of nodes, representing arbitrary states in the trace, partitioned into V_0 and V_1, indicating the speaker.

Where U is a set of actions, representing speech in our trace.

Where T is a set of transitions (edges), representing the speech moving the conversation to the next state. $T \subseteq V \times U \times V$

Sub-Conversation

We define each action in U as follows:

$U = \{u_1, u_2, u_3, \ldots, u_n\}$, where n is the number of total actions recorded.

Each u_i is defined as a tuple:

$u_i = (p, c, \ell)$

Where p corresponds to the speaker of the node.

Where c corresponds to the words of the utterance.

Where ℓ corresponds to the label of the utterance, $\ell \in \mathscr{L}$

Sub-Conversation

Our set of labels, $\mathscr{L}$ is preliminarily listed as:

open.question - a question that expects more than an exact singular response.

closed.question - a question that expects an exact singular response.

respond.agree - response to posed question that is in the affirmative or satisfies the question parameters.

respond.deny - response to the posed question that is in the negative, or denies answering the question.

display.reflection - an utterance that is meant to be part of an inner monologue

give.opinion - an opinion given to the other speaker on the topic, or a statement spoken by the speaker.

deflection - a response that is neither an affirmative/negative, and satisfies the parameters of the question, but the other speaker cannot confirm whether this satisfaction is valid or has occurred.

use.social.convention - speaker engages in a social convention (i.e: polite nod, greeting, gestures)

relax.atmosphere - speaker engages in a social convention with the intent to affect the mood (i.e: laughter, inside joke/statement based on rapport)

Sub-Conversation

- We look for Sub-Conversation Properties

- Some sample properties are:

    - *Clarification* (respond.agree.1 -> give.opinion.1)

    - *Active.Listening* (respond.agree.1 -> give.opinion.2)

- One of our next tasks is to come up with a larger list of sub-conversations

Similarity Measure - Conclusions

- Work at the utterance level, treat each speech as an action (speech act).

- Similarity determined by labels on actions (trace of our labels).

- Modified levenshtein distance, take into account context, position, etc.

- Uncertainty can be addressed separate of similarity.

TraceSim (Sep 2020)

- Originally in context of bug-report triages.

- Errors are presented as a stack.

- Algorithm determines first whether it is a stack overflow exception (handled the same).

- Algorithm continues to calculate weight of each line of the stack.

- Weight affected by relative distance to the top or bottom of the stack.

- Modified version of levenshtein distance (swap order irrelevant).

Context Aware Trace Clustering (2009)

- Most relevant, provides outline of similarity measures.
  - Bag of Activities
  - K-gram Model
  - Hamming Distance
  - Edit-Distance (Levenshtein)
- Add a notion of cost, penalize edits that are unnecessary.
- Add context via k-gram.
- Calculate co-occurrences, probability of occurrence of symbols, normalization, etc.

Sequence = { aabcadaba , bbadcabaa, abbadcca ,
             abadcabbacda baaacbdacddda }

Actions = { a, b, c, d }

3-grams = { aab, abc, bca , cad , ada , dab, aba , bba, bad,
            adc , dca, cab, baa , abb, dcc, cca , bac, acd
            cda , dab, baa, aaa, aac , acb, cbd, bda ,
            dac , cdd , ddd , dda }

freq = { 1, 1, 1, ... }

X_a = { ab, cd, db, bd, cb, ba, bc, aa, dc }

X_b = { ac, aa, ab, cd }

"denoted
'context'
in acts."

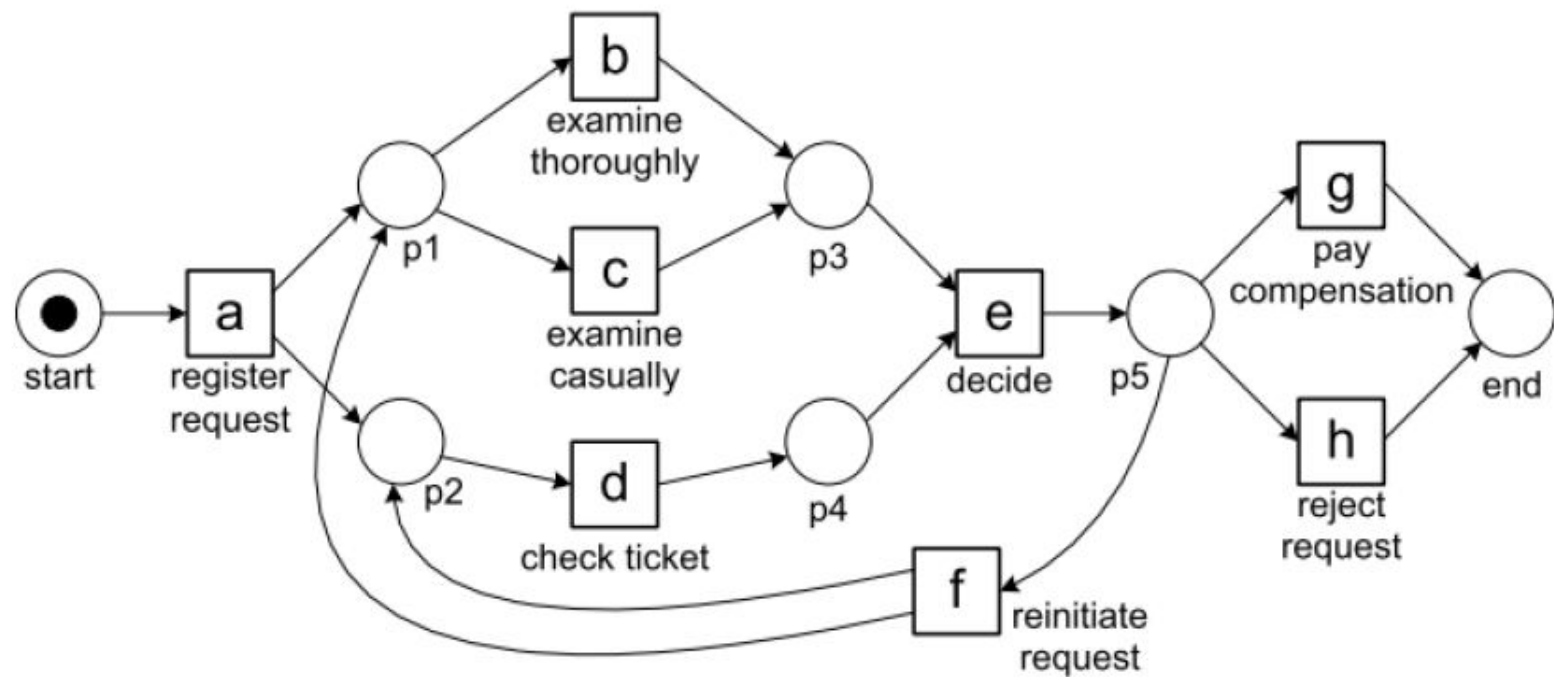$X_{a,b}$ = { ab, cd , aa }

ie : "open, question → respond, deny →
                       relax, atmosphere."

Context over sub-conversations

- Context as a part of similarity measure helps detect sub-conversations, i.e:

  "*ababc*" and "*babac*" will yield *aa* and *bb* contexts in both, and

  co-occurrence will catch *bc* and *ac* divide.

Conformance Checking Over Uncertain Event Data (2020)

- Dealing with uncertain data, i.e: where data values have a range, data values are not reliable, etc.

- Causes: Incorrectness, Coarseness, Ambiguity

- Strong and Weak Uncertainty

- Third level of uncertainty at the system level

- Behavioral Net to replay uncertain traces, obtained via dependency graph

$$\gamma_1 = \begin{array}{|c|c|c|c|c|} \hline a & d & b & e & h \\ \hline a & d & b & e & h \\ \hline \end{array} \qquad \gamma_2 = \begin{array}{|c|c|c|c|c|c|} \hline a & \gg & d & b & e & h \\ \hline a & b & d & b & e & h \\ \hline \end{array}$$

Uncertainty in speech acts

- We can identify uncertainty in speech:
  - Mishearing
  - Misunderstanding
  - Unintelligibility
- More specific uncertainty in Language Exchange:
  - Mishearing (i.e: hearing incorrect words)
  - Misunderstanding (i.e: answering question that wasn't asked, colloquial sayings)
  - Unintelligibility (i.e: grammar, pronunciation, wrong words)

Stealing Data

- Google TaskMaster data (mostly queries to an assistant)

- Assorted Kaggle datasets (movie scripts, etc)

- Current data too linear, format doesn't fit our labels (question answers, less opinions)