Conversation Facts

Nathan Miller, Christian Kang, Jon Marks

CS 4624 - Multimedia, Hypertext & Information Access - Dr. Fox
Virginia Tech - Blacksburg 24061 - 4/25/2019

Outline

- Project Overview
- Implementation
  - Parsing
  - Preprocessing
  - Building KGs
  - Matching
- Results
- Challenges
- Deliverables
- Acknowledgements

Project Overview

Project Goal - Matching summary sentences to their origin in a conversation

Dataset - Argumentative Dialogue Summary Corpus: Version 1.

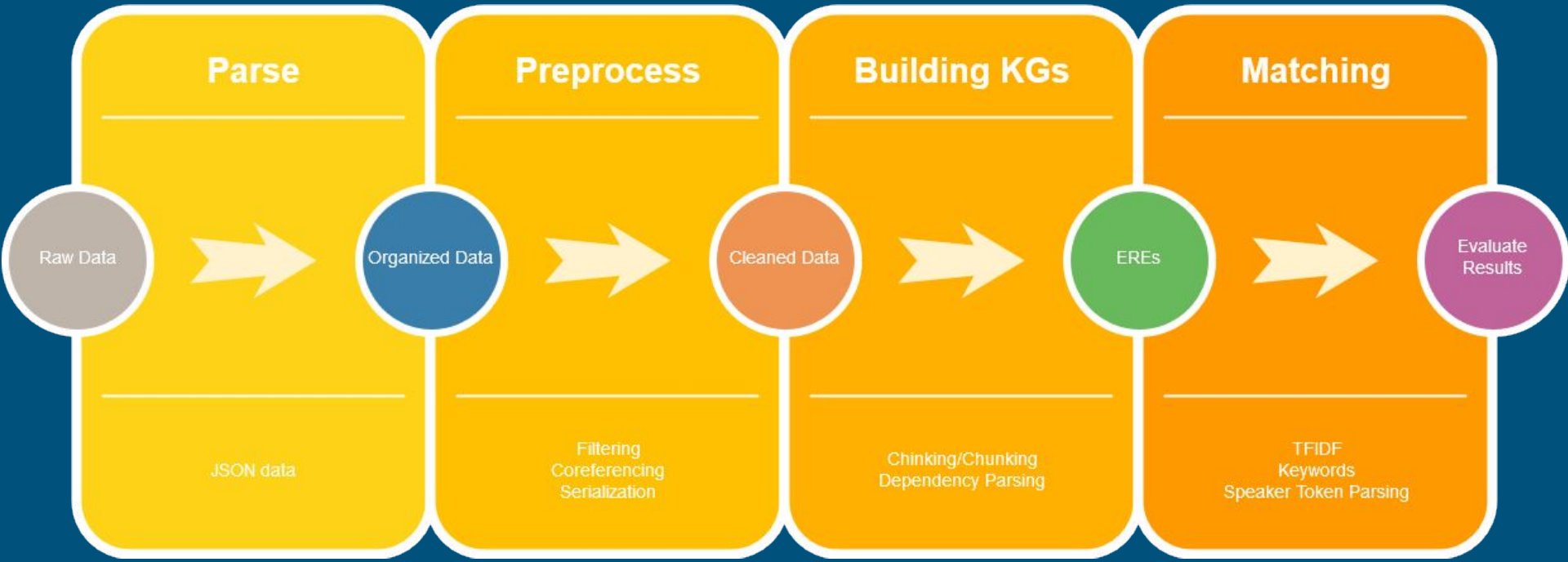
S1:1- Another problem with the study is that most (or at least more).....

S2:1- S1 you don't know how they conducted the study. You don't know....

…….

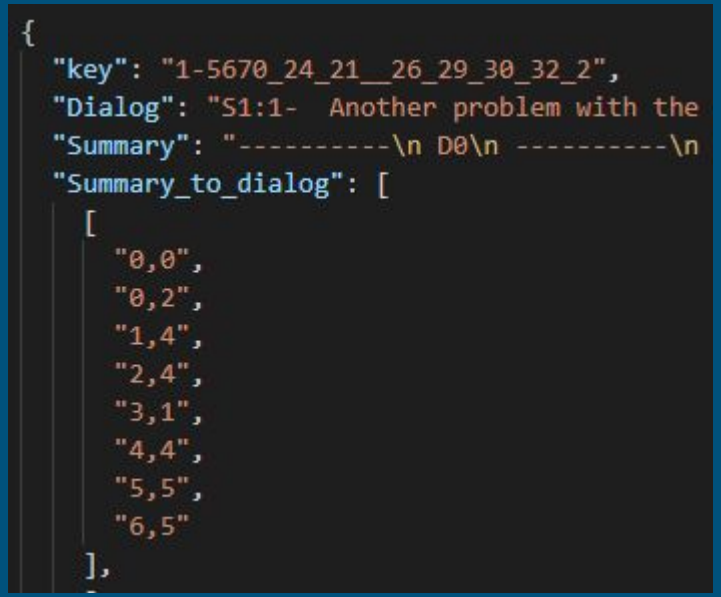
D0: S1 argues that most people in the country might identify as Chr……..

Implementation



Parsing

- Unique ID key
- Dialog string split with “S#:#” tags
- Summary string split with newlines/”D#” labels
- Summary to dialog annotations
  - Group made mapping of summary sentences to their relevant dialog



Preprocessing

- Rule Based filtering
  - Remove text between parentheses
  - Remove sentences under a given length
  - Replace words to increase comprehension eg “you’re” -> “you are”
- Special Text Filtering
  - Remove special characters, extraneous punctuation, emoticons etc.
- Coreferencing
  - Resolving ambiguous pronouns (Next Slide)
- Serializing preprocessed object
  - ~20x speedup by not rerunning expensive operations eg spaCy

Preprocessing - Coreferencing

```
S1: Evidently you
    didn't read
    the links I
    gave you.
S2: I clearly did,
    I suspect you
    may have but
    didn't understand.
```

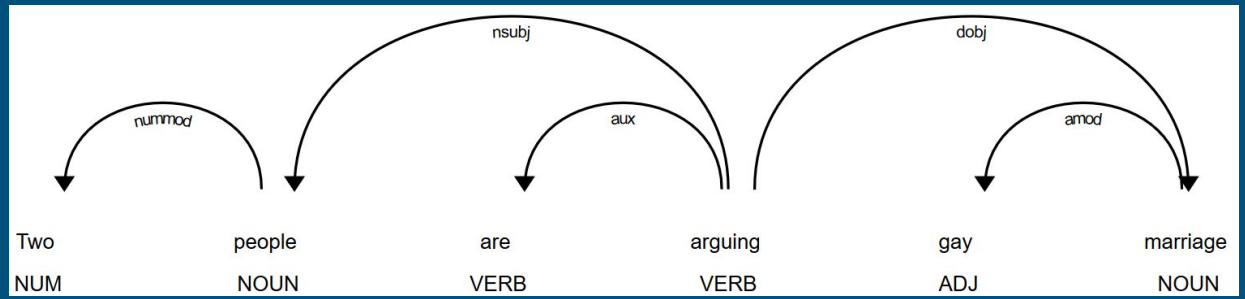
```
S1: Evidently S2
    didn't read
    the links S1
    gave S2.
S2: S2 clearly did,
    S2 suspect S1
    may have but
    didn't understand.
```

Building ERE triples KGs

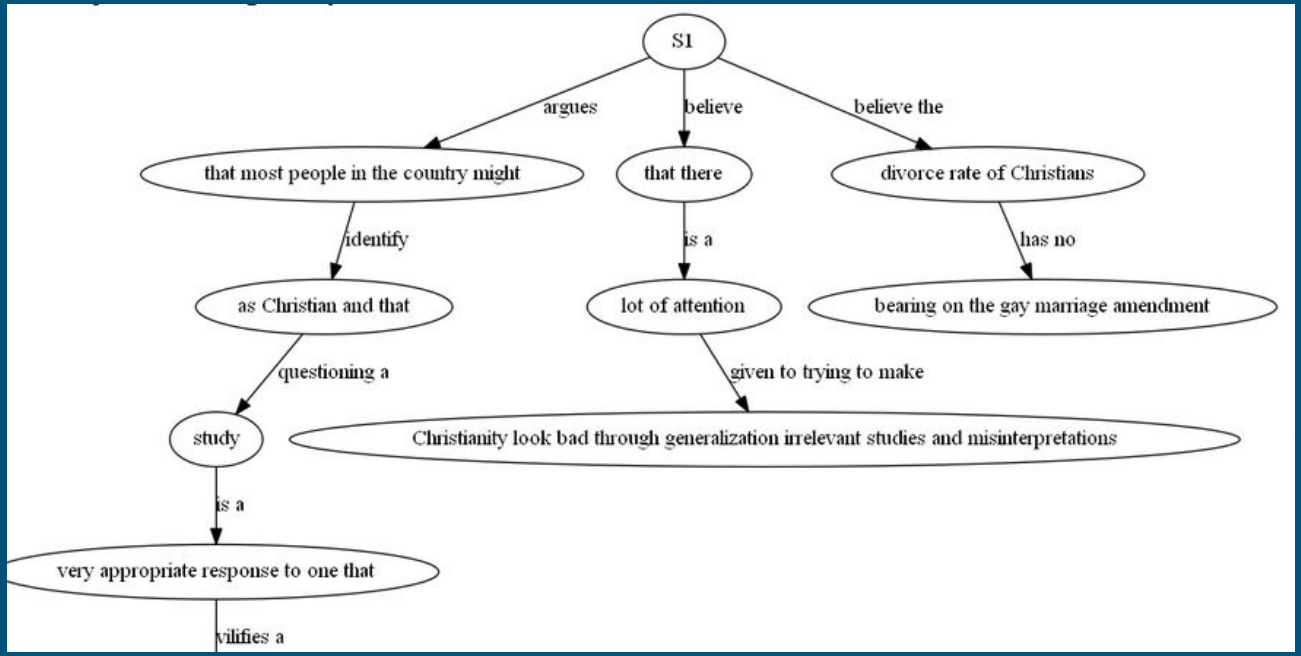
- Chinking/Chunking

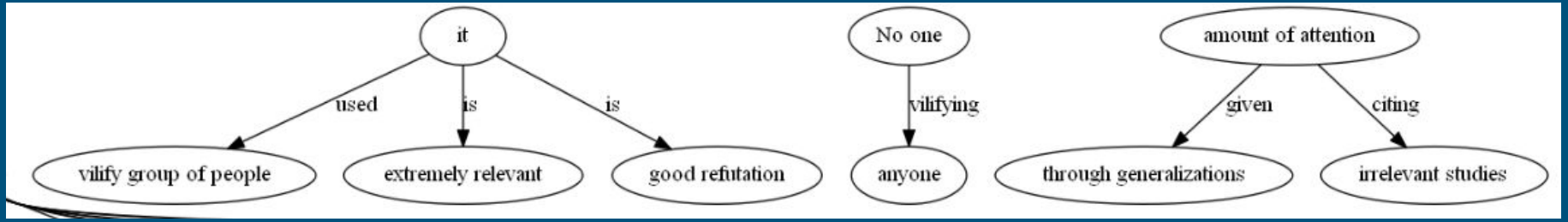
```
tag_pattern = """NP: {<.*>+}
                    }(<DT|RB.*|TO>?<VB.*>+<DT|RB.*|TO>?)+{
                VP: {(<DT|RB.*|TO>?<VB.*>+<DT|RB.*|TO>?)+}
               """
```

- Dependency Parsing
  - spaCy



Example KG


S1
argues
believe
believe the
that most people in the country might
that there
divorce rate of Christians
identify
is a
has no
as Christian and that
lot of attention
bearing on the gay marriage amendment
questioning a
given to trying to make
study
Christianity look bad through generalization irrelevant studies and misinterpretations
is a
very appropriate response to one that
vilifies a


it
used
is
is
vilify group of people
extremely relevant
good refutation
No one
vilifying
anyone
amount of attention
given
citing
through generalizations
irrelevant studies

Matching Summaries and Dialogs

- Cosine Similarity
  - Use TFIDF to compare summary EREs and dialog
- Key/Rarewords
  - Check summary EREs for important specific words in dialog
- S1/S2 token parsing
  - Use order and count of S1/S2 strings in summaries
  - “S1 argues that…..”

Results

- Check correctness against manually annotated dataset

	Random	spaCy	Our Method
% Correct	16.96%	24.87%	33.18%
% Relative change over random	0%	46.63%	95.63%

Challenges and Lessons Learned

```
"I'm talking ab out teh major earthquakes taht"
"( as Kronus so eloquently put it :) )"
"p.s. Did yo daddy plug da hole today???"
"S2:3-  Here you go :  "
```

1. **Dataset**
   a. Short conversations are more difficult to match.
   b. Conversations lack starting posts/links/quotes and posters handles.
   c. Grammar, spelling, formatting, and general relevance are not priorities to forum posters.
2. Off the shelf libraries don't necessarily work for all use cases
   a. Neural coref, AllenNLP
   b. SNAP.py
   c. spaCy dependency parsing, similarities

Deliverables

- Final Report
  - Document detailing project
- Code
  - Github Repo (link in final report)
- Docker
  - Publicly accessible container (link in final report)

Acknowledgements

Client - Saurabh Chakravarty saurabc@vt.edu

Resources -

- Stanford NLP Dependency Manual https://nlp.stanford.edu/software/dependencies_manual.pdf
- Christopher Manning Lectures https://www.youtube.com/watch?v=0QQ-W_63UgQ&list=PL3FW7Lu3i5Jsnh1rnUwq_TcylNr7EkRe6
- Amita Misra, Pranav Anand, Jean E. Fox Tree, Marilyn Walker. "Using Summarization to Discover Argument Facets in Online Idealogical Dialog", In The *North American Chapter of the Association for Computational Linguistics*