A Large-Scale Corpus for Conversation Disentanglement

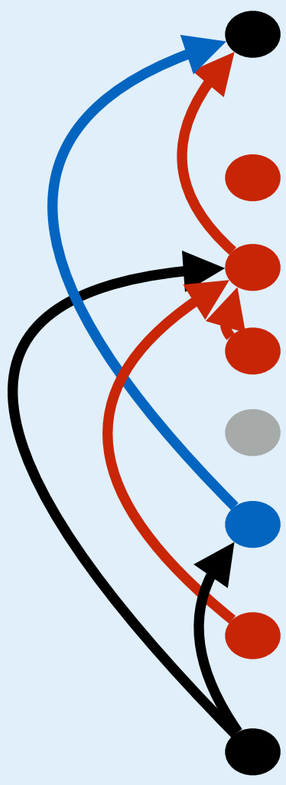
Jonathan K. Kummerfeld, Sai R. Gouravajhala, Joseph Peper, Vignesh Athreya, Chulaka Gunasekara, Jatin Ganhotra, Siva Sankalp Patel, Lazaros Polymenakos, Walter S. Lasecki

Overview

When a group of people communicate in a common channel there are often multiple conversations occurring concurrently. We created a new dataset of English messages manually annotated with reply-structure graphs that both disentangle conversations and define internal conversation structure.

- **77,563** annotated messages (16x all prior public data)
- From **173 points** in time over 14 years
- First with **adjudicated dev and test sets**
- Automatically extracted **496,469** conversations
- Basis of DSTC 8, Task 2: http://bit.ly/dstc8-task2

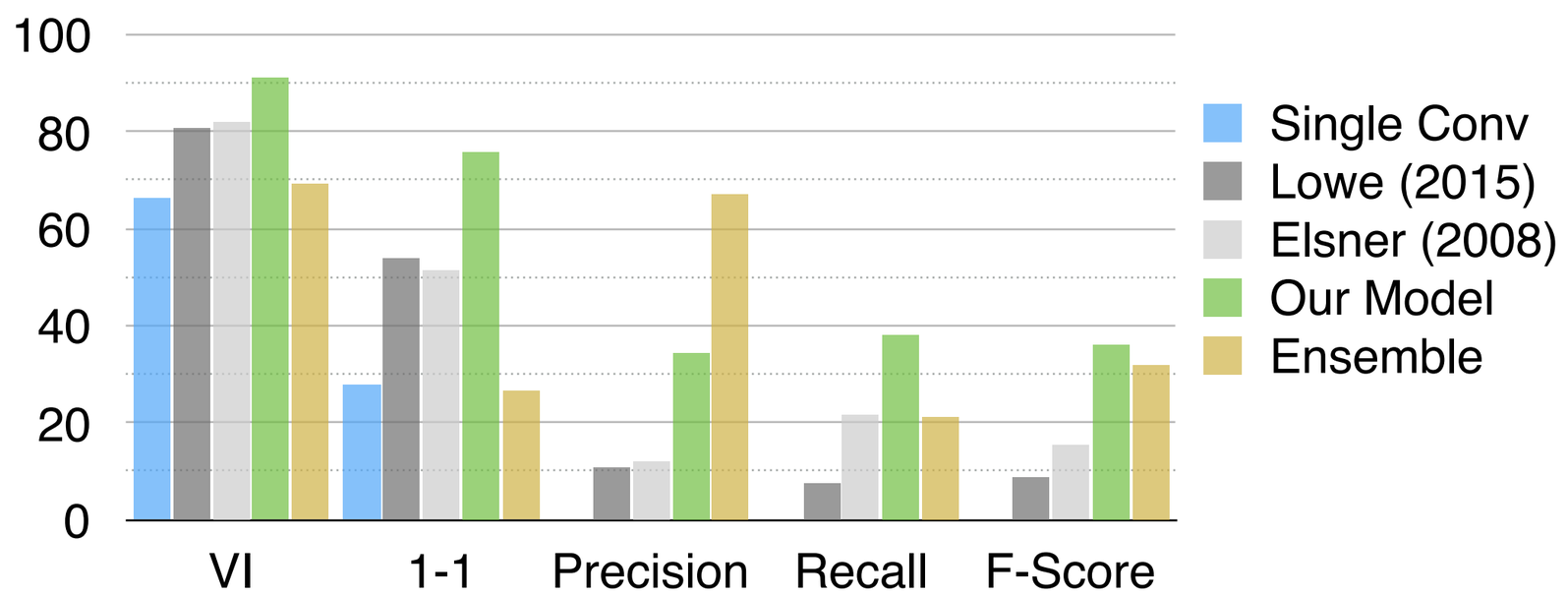
Code and data:
https://jkk.name/irc-disentanglement/



```
[03:06] <BurgerMann> does anyone know a consoleprog that scales jpegs
fast and efficient?.. this digital camera age kills me when I have to
scale photos :s
[03:06] <Seveas> delire, yes
[03:06] <Seveas> BurgerMann, convert
[03:06] <Seveas> part of imagemagick
=== E-bola [...@...] has left \#ubuntu []
[03:06] <delire> BurgerMann: ImageMagick
[03:06] <Seveas> BurgerMann, i used that to convert 100's of photos in
one command
[03:06] <BurgerMann> Oh... I'll have a look.. thx =)
```

Results

- Annotator agreement is 0.71-0.74 κ on graph structure.
- Our model performs 9 - 22 points better than prior work.
- 67% of output conversations from our ensemble are perfect, 14% more are prefixes of a conversation.



First Evaluation of the Ubuntu Dialogue Corpus (Lowe et al., 2015)

- 10% of their conversations are perfect, 10% are prefixes.
- The heuristic links together messages far apart in time.
- Re-evaluating dialogue models on our data leads to comparable conclusions.