A Corpus of Word-Aligned Asked and Anticipated Questions in a Virtual Patient Dialogue System

The Ohio State University

Ajda Gokcen, Evan Jaffe, Johnsey Erdmann,
Michael White, Douglas Danforth

Motivation

- Many systems for detecting paraphrases or measuring semantic similarity rely on alignments of words and phrases
- Gold standard monolingual alignment corpora MSRP (Brockett, 2007) and Edinburgh (Cohn et al., 2008) corpora do not take paraphrase status into account when annotating word alignments
- There can be benefit to modeling word alignment and paraphrase classification as a joint process (Xu et al., 2014)

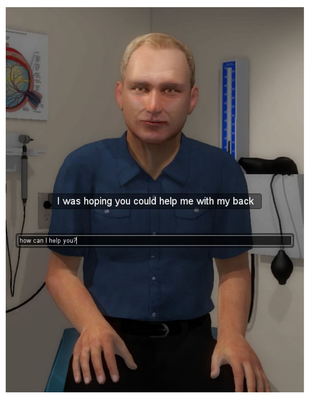
Contribution

We developed a corpus and methodology of annotation that accounts for word alignment and paraphrase status jointly. Our resources include:

1. **A corpus of virtual patient dialogues** with **gold standard alignment and paraphrase annotations**
2. **Alignment annotation guidelines** to account for both paraphrases and non-paraphrases, extending Cohn et al. (2008)
3. **GoldAlign**, the graphical alignment and paraphrasing annotation interface used to create the corpus

The Application: Virtual Patient Dialogue System

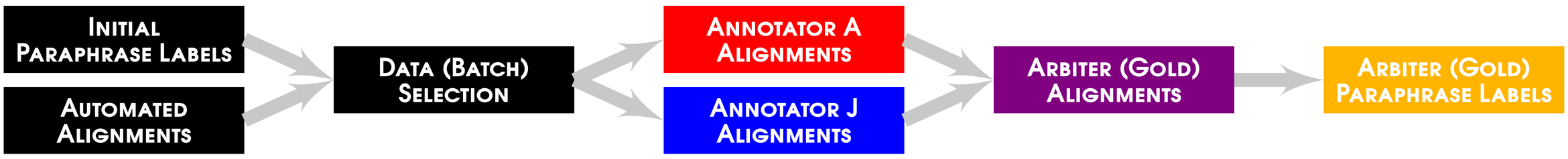
The Virtual Patient



An Example Dialogue

Input	Gold Label	ChatScript System Response
hello there, why did you come in to see us today	what brings you in today	i was hoping you could help me with my back pain, it really hurts! it has been awful. i can barely even tie my shoes
could you rate the pain	rate the pain	right now, i would call my pain a 7. when i take the ibuprofen, it goes down to about a 5 or 6
is the pain constant	is the pain constant	it is pretty constant although sometimes it is a little better or worse...
how is this affecting your life	how is this affecting you	i haven't been able to work, i am a little worried about the bills
what do you do for work	what do you do for work	i am an auto mechanic. i have a garage a couple of blocks from here...

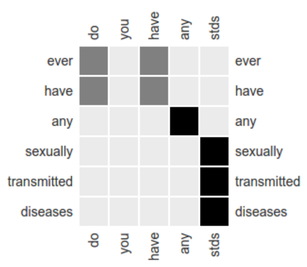
Corpus Annotation Process

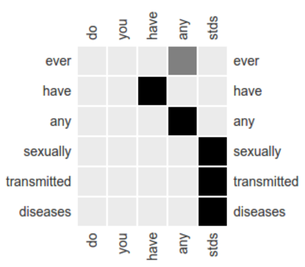


GoldAlign: Alignment Annotation Tool

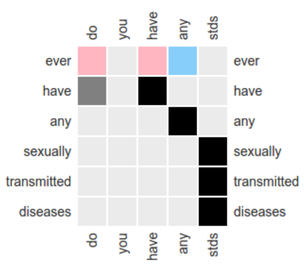
GoldAlign, a graphical interface for gold-standard corpus annotation of alignment and paraphrasing, was created to streamline the creation of our corpus (and potentially others like it).

Annotator Grids (A & J)

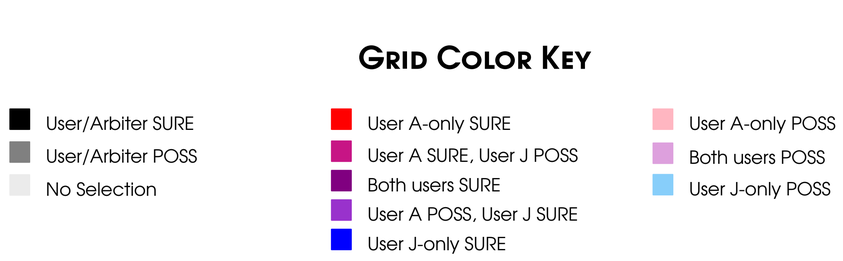




Arbiter Grid



Grid Color Key



Sentence Contexts

Paraphrase Judgments

Corpus Statistics

- **104 dialogues** containing **5437 user turns** with **290 unique labels**
- average of **52 turns per dialogue** and **19 turns per label**
- **157 annotated+arbitrated batches** with 6 sentence pairs in each:
  - selected by **maximum confusability**, i.e., the least similar paraphrases and most similar non-paraphrases
  - yielding **942 total sentence pairs**
  - of these, **441 paraphrases** and **495 non-paraphrases**

Inter-Annotator Agreement

Annotators		F1 (words)		F1 (phrases)	
		Sure	Poss	Sure	Poss
M	G	0.22	0.19	0.30	0.40
A	G	**0.72**	**0.74**	**0.81**	**0.74**
J	G	**0.63**	**0.64**	**0.69**	**0.60**
A	M	0.29	0.19	0.36	0.33
J	M	0.33	0.27	0.48	0.45
A	J	**0.47**	**0.57**	**0.60**	**0.55**

Agreement of the annotators (**A** and **J**) relative to one another and to the gold annotations (**G**) are consistently higher (shown above in bold) than agreement between the Meteor alignments (**M**) and all human annotations.

Acknowledgments

We would like to acknowledge Kellen Maicher who created the virtual environment and Bruce Wilcox who authored ChatScript and customized the software for this project. We also acknowledge the expert technical assistance of Laura Zimmerman who managed the laboratory and organized student involvement in this project. This work was made possible by a Targeted Investment in Excellence grant from the Department of Linguistics at the Ohio State University. The material is also based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1343012. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. This project was additionally funded (in part) by a National Board of Medical Examiners (NBME) Edward J. Stemmler, MD Medical Education Research Fund grant (NBME 1112-064). The project does not necessarily reflect NBME policy, and NBME support provides no official endorsement. This project was also supported by funding from the Department of Health and Human Services Health Resources and Services Administration (HRSA D56HP020687).