Collecting task-oriented dialogues

Linguistic pragmatics concerns how context and world knowledge influence language production and interpretation. In order to study such phenomena effectively, we need large amounts of data involving people interacting in controlled settings addressing specific issues. Historically, such corpora have been expensive and time-consuming to generate, and thus there are relatively few of them. We argue that crowdsourcing can help fill this gap. We report on a project to collect task-oriented transcripts using Mechanical Turk as a recruiting tool. Our basic scenario is a two-person collaborative search game. The game-world consists of a maze-like environment in which a deck of cards has been randomly distributed. Each player can always see his own location, but the location of the cards and the other player are limited by distance and line of sight. Players interact via IM style chat box to decide upon and collect card sequences. We were initially unsure that a crowdsourcing effort would succeed with this game. It involves lots of typing, extensive interaction with strangers, and some waiting around for a partner to connect. All these factors are thought to be problematic for crowdsourcing. However, the effort proved extremely successful. Over a two-week period, we collected 439 good transcripts (12,280 utterances, 64,900 words), out of 473 transcripts in all. In the talk, we will describe the data collection effort and the lessons we learned from it, and we will highlight some of the important phenomena we observe in the corpus we obtained.