



Automated Parsing of Dialogue Games in Open Ended Dialogue

Matthew Marchesani

Advisor – Professor Nick Webb

Abstract

Dialogue Games (DG) are patterns that are formed in dialogues. They make up the structure of natural spoken language. DGs are built up from Dialogue Acts (DA). This happens the similarly to the way that sentences are built up from words. My research uses the of the following low level structural abstractions of natural language, including dialogues, turns and utterances. Dialogue are made up from turns by speakers. And turns are made up from utterances by a speaker.

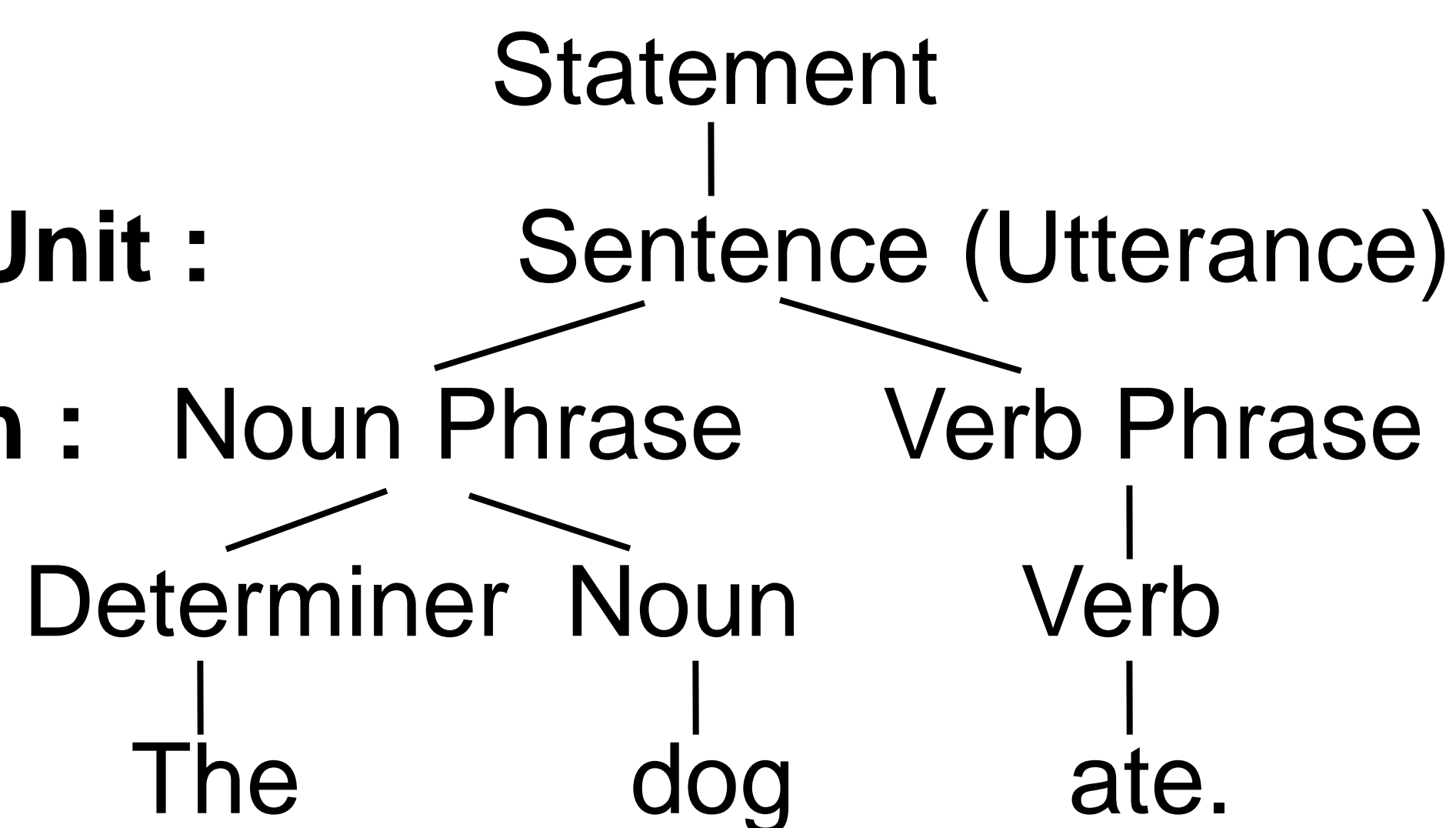
Corpus

The Switch Board (SWBD) corpus is in XML format. It is a set of 1155 dialogues between two participants recorded via telephone call. The conversations are open ended (ie. non-task based). A subset of 35 of the Dialogue Act Markup in Several Layers (DAMSL) DA tags tag utterances in the dialogues.

Dialogue Act :

Grammatical Unit :

Part of Speech :



My Work

I used Python, Natural Language Tool Kit (NLTK) and Wordnet to automate parsing of the DGs with respect to DA tags and topic words. The NLTK part of speech tagger parsed utterances, allowing nouns to be located. Wordnet measured noun similarity for DG parsing.

Statistics

DGs with No Topic

DGs parsed: 7262
 Percentage of questions parsed: 80.79%
 Percentage of total utterances parsed: 26.94%
 DGs immediately followed by DGs: 26.94%
 Non-parsed question followed by DG: 25.06%

DGs with Topic

DGs parsed: 6479
 Percentage of questions parsed: 72.01%
 Percentage of total utterances parsed: 20.24%
 DGs immediately followed by DGs: 15.41%
 Non-parsed question followed by DG: 19.62%

Greeting DG

→ A: Hi.
 → B: Hey.

Question/Answer DG

→ A: What's your address?
 → B: 807 Union St.

Nested Question/Answer DG

→ A: What's your address?
 → B: Why do ask?
 → A: I have a package.
 → B: 807 Union St.

Future Work

Design a parsing algorithm for nested DGs. Then, divide dialogues into sections according to topics. Inside of the topic sections, parse the games according to topic more accurately than now. Increase the length of DGs by including more turns, they will contain nouns with the same topic as the proceeding DG. Once 90% or more of the utterances in the corpus are parts of a DGs, represent the corpus with DG only in XML.