Molweni: A Challenge Multiparty Dialogues-based Machine Reading Comprehension Dataset with Discourse Structure

Jiaqi Li, Ming Liu, Min-Yen Kan, Zihao Zheng, Zekun Wang,
Wenqiang Lei, Ting Liu, Bing Qin
Harbin Institute of Technology, National University of Singapore, Peng Cheng Laboratory

INTRODUCTION

- We present the *Molweni* dataset, a machine reading comprehension (MRC) dataset with discourse structure built over multiparty dialog. Molweni's source samples from the Ubuntu Chat Corpus, including 10,000 dialogs comprising 88,303 utterances. We annotate 30,066 questions on this corpus, including both answerable and unanswerable questions.
- Molweni also uniquely contributes discourse dependency annotations in a modified Segmented Discourse Representation Theory (SDRT)) style for all of its multiparty dialogs, contributing large-scale (78,245 annotated discourse relations) data to bear on the task of multiparty dialog discourse parsing.
- Our experiments show that Molweni is a challenging dataset for current MRC models: BERT-wwm, a current, strong SQuAD 2.0 performer, chieves only 67.7% F_1 on Molweni's questions, a 20+% significant drop as compared against its SQuAD 2.0 performance.

Dialogue 1

nbx909: how do i **find the address of a usb device** ?	U_1
likwidoxigen: try taking it out to dinner and do a little wine and dine and it shoudl tell ya	U_2
likwidoxigen : what sort of device ?	U_3
babo: ca n't i just copy over the os and leave the data files untouched ?	U_4
nbx909 : only if you do an upgrade	U_5
nuked: should i just restart x after installing	U_6
likwidoxigen: i 'd do a full restart so that **it re-loads the modules**	U_7

(a)

Q1: Why does *likwidoxigen* do a full restart?
A1: it re-loads the modules

Q2: What does *nbx909* want to do?
A2: find the address of a usb device

Q3: How to restart network?
A3: *NA.*

(b)

Figure 1. (Dialog 1) A corpus example from Molweni. There are four speakers in the dialog: *nbx909*, *likwidoxigen*, *babo*, and *nuked*. In total, the speakers make seven utterances: U_1 to U_7. Our annotators proposed three questions against the provided dialog: Q1–3, where Q1 and Q2 are answerable questions, and Q3 is unanswerable. Due to the properties of informal dialog, the instances in our corpus often have grammatical errors.

OVERVIEW OF *MOLWENI*

	Train	Dev	Test	Total
Number of Dialogs	8,771	883	100	9,754
Number of Utterances	77,374	7,823	845	86,042
Number of Questions	24,682	2,513	2,871	30,066

Table1. Overview of Molweni for MRC.

	Train	Dev	Test	Total
Number of Dialogs	9,000	500	500	10,000
Number of Utterances	79,487	4,386	4,430	88,303
Number of Relations	70,454	3,880	3,911	78,245

Table2. Overview of Molweni for DP.

Metric	Number
Average / Maximum number of speakers per dialog	3.51 / 9
Average / Maximum question length (in tokens)	5.91 / 18
Average / Maximum answer length (in tokens)	4.08 / 19
Average / Maximum dialogue length (in tokens)	104.4 / 208
Average / Maximum dialogue length (in utterances)	8.82 / 14
Vocabulary size	24,615
Answerable questions	25,779
Unanswerable questions	4,287

Table3. Detailed statistics for the Molweni corpus.

ANNOTATION FOR MACHINE READING COMPREHENSION

Dataset	Answer type	Dialogue text	Multiparty dialogue	Unanswerable questions	Discourse structure
RACE (Lai et al., 2017)	multiple-choice	✗	✗	✗	✗
NarrativeQA (Kocisky et al., 2018)	abstractive	✗	✗	✗	✗
CoQA (Choi et al., 2018)	abstractive	✗	✗	✔	✗
SQuAD 2.0 (Rajpurkar et al., 2018)	extractive	✗	✗	✔	✗
QuAC (Choi et al., 2018)	extractive	✗	✗	✔	✗
(Ma et al., 2018)	cloze	✔	✔	✗	✗
DREAM (Sun et al., 2019)	multiple-choice	✔	✔	✗	✗
FriendsQA (Yang and Choi, 2019)	extractive	✔	✔	✗	✗
Molweni (Our)	extractive	✔	✔	✔	✔

Table4. Comparison of Molweni with other MRC datasets on answer type, text type (dialogue or written text), multiparty dialogs or not, unanswerable questions, and discourse structure.

Question	Example		Proportion(%)
How	How to do an upgrade?	How can I use this machine?	9.9
Why	Why is it not mounted?	Why does *jimcoonact* meet the error?	4.3
Who	Who is chart's service customers?	Who is using ubuntu?	4.7
When	When does *rhodry* have the error?	When is *SuperMiguel* back?	1.7
Where	Where did *earthen* write in?	Where is the device?	5.7
What	What does *elnomade* choose?	What does *noone* need?	71.7
Others	Does *elnomade* choose the print?	Which version does *xxiao* find?	1.9

Table5. Examples of questions in Molweni.

ANNOTATION FOR DISCOURSE PARSING

Dialogue 2

toma-: but its well worth the wait	U_1
woodgrain: i have a decently fast p4 should i still be waiting ?	U_2
toma-: have you run updatedb before ?	U_3
woodgrain: no never before -- but it worked and now i have all the files i need .	U_4
woodgrain: i do n't have a path to the jre -- do i need to add it ?	U_5
toma-: a path ? ? you compiling somehting ?	U_6
woodgrain: do n't need jdk as witnessed by eclipse irc	U_7
woodgrain: no i 'm installing this newer ver from the eclipse site .	U_8

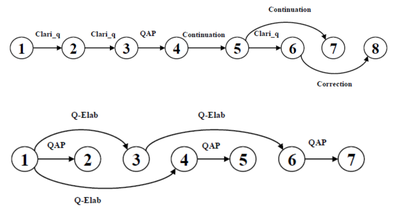

Figure 2. Dialog 2 is a two-party dialog example with eight utterances—U_1 to U_8 —proposed by two speakers: toma- and woodgrain. The discourse dependency structure and relations for Dialog 2 (Top, two-party) and Dialogue 1 (Bottom, multiparty). Clari_q, QAP, and Q-Elab are respectively short for Clarification question, Question-answer pair, and Question-Elaboration. The label on the link represents the discourse dependency relations between two utterances.

DATA QUALITY

- Manually Check.
- Programmatic Check.
- The FLESS Kappa: 0.91 for discourse dependency links, and 0.56 for both links and discourse relations.

EXPERIMENTAL RESULTS

Method	EM		F1	
	Squad 2.0	Our	Squad 2.0	Our
BERT-base	73.1	45.3	76.2	58.0
BERT-large	80.0	51.8	83.1	65.5
BERT-wwm	86.7	54.7	89.1	67.7
Human performance	86.8	64.3	89.4	80.2
Human-machine gap	0.1	9.6	0.3	12.5

Table6. Results of machine reading comprehension for multiparty dialogs.

Method	Link		Link & Relation	
	STAC	Our	STAC	Our
Deep sequential	73.2	78.1	55.7	54.8
Deep sequential(C)	78.0	77.0	54.7	54.3

Table 7. Results of discourse parsing on multiparty dialogs (F_1-score). Deep sequential (C) means combine the training set of STAC and Molweni as the training set and test the model respectively.

CASE STUDY

Dialogue 3

nuked: ok likwidoxigen ill reboot and let you know how it goes	U_1
likwidoxigen: who makes the printers ? and they woked before yets ?	U_2
nuke: yes they worked excellently on dapper . they are two hp deskjets	U_3
nbx909: does n't give me the address	U_4
likwidoxigen: and they just dont ' print properly ?	U_5
likwidoxigen: ok let me keep poking	U_6
nbx909: i know but it 's a ups (battery backup) device would it be under sda ?	U_7
nuked: i used kde 's add printer wizard , and only samba printers are allowed	U_8
likwidoxigen: i 'd assume so , it still has to access the device	U_9
likwidoxigen: damn do any usb device work ?	U_{10}

(a)

Q1: Who does ask for the address?
Gold answer: *nbx909*
BERT-wwm answer: *likwidoxigen*

Q2: how are printers working?
Gold answer: *NA.*
BERT-wwm answer: they worked excellently on dapper.

(b)

Figure 3. Dialogue3. (a) A real example from Molweni dataset with three speakers and ten utterances. (b) Two questions for Dialog 3 and the pridected answers of BERT-wwm model.