A Brief Survey of Textual Dialogue Corpora

Hugo Gonçalo Oliveira[1,2], Patrícia Ferreira[1,3], Daniel Martins[3], Catarina Silva[1,2], Ana Alves[1,3]

[1]CISUC, Universidade de Coimbra, Portugal
[2]DEI, FCTUC, Universidade de Coimbra, Portugal
[3]ISEC, Instituto Politécnico de Coimbra, Portugal
hroliv@dei.uc.pt, patriciaf@dei.uc.pt, daniel.martins2997@gmail.com, catarina@dei.uc.pt, ana@dei.uc.pt

1. Introduction

- Growing need of conversational data
  - train dialogue systems
  - study specific dialogue phenomena
- Broad range of available dialogue corpora
  - Still far from covering all necessities of real-world applications
  - Dialogue analysis tasks, domains, languages, etc.
  - Usage restrictions (privacy, logistics)
- Researchers have to choose between:
  - Collect (and annotate) a new corpus from scratch
  - Select the most suitable publicly available corpus
- Survey of available corpora (see table below)
  - **Human-human** conversations
  - Available as **text**
  - Different sizes, numbers of speakers, languages, collection approaches, annotations, domains.

2. Speakers, Size and Languages

- Most corpora include conversations between **two humans**, with few exceptions:
  - Meetings, movies, posts with no answer
- Variable sizes, from a **few dozens to thousands** of dialogues
- Most corpora in **English**, with few exceptions:
  - Chinese, French, Spanish

3. Linguistic Annotations

- Corpora with no annotations are still useful for **data-driven response generation**.
- **Dialog State Tracking** (*slots and their values*)
- **Dialog Acts** (*action performed by the speaker*)
- **Polarity and Emotion**
- Other Annotations
  - Summaries, movie recommendations, topic, diagnosis and treatment suggestions, rationale for question-answering

4. Data Collection

- Mostly **scripted conversations**, with few exceptions
  - including transcriptions of spoken conversations
- **Wizard of Oz**
- **Crowdsourcing**
- **Web sources**
  - Chat logs, forums, language learning websites, social networks, movie / tv subtitles
- **Written by a single person**, according to some guidelines

5. Domains & Contents

- **Task-oriented**
  - **Single-domain** (customer-agent):
    - travelling and tourism (locations, times), including restaurants and urban transports
    - in-car personal assistance
  - **Multi-domain** (customer-agent), covering some of the following:
    - restaurants, hotels, attractions, taxis, trains, hospitals, police, flights, car rentals, ordering pizza / drinks, auto-repair appointments, rider service, movie tickets, alarm, media, messaging, weather, etc.
  - Less-specific requests
    - Support: technology and telecommunications, patients and doctors
    - Cooperation: design, orientation, bargaining
- **Open conversations**:
  - Topics, debates
  - Questions and their answers
- **Open-domain conversations**:
  - Social network, messenger, daily life
  - Movie / tv subtitles

6. Wrap up

- Options when a suitable corpus is not available...
  - Creation:
    - Collect dialogues from **web sources**: movie subtitles, social networks, forums
    - From scratch: **WOZ** (crowdsourcing); **self-dialog** (written by a single person)
  - Annotation:
    - **manual** (experts, crowdsourcing); **previously-generated**; translation

Table: List of surveyed corpora and their properties.

Name	Speakers	Modality	# Dialogues	Language	Annotations	Collection	Domain
AMI	4	AUD+TXT	100 hours	EN	DAs	recorded meetings	interacting w/ technology
CamRest676	2	TXT	676	EN	state	crowd (WOZ)	restaurants
CoQA	2	TXT	8,399	EN	question rationale	crowd (questioner-answerer)	7 domains
CrossWOZ	2	TXT	6,000	ZH	state	crowd (WOZ)	5 domains
DailyDialog	2	TXT	13,118	EN	DAs, emotion	web forum	daily communication
DealOrNoDeal	2	TXT	5,808	EN	items + values	crowdsourcing (chat)	bargaining
DECODA	2	AUD+TXT	1,514	FR	call type	recorded phone calls	urban transports
DIME	2	AUD+TXT	31	ES	DAs, discourse referents	crowd (WOZ)	kitchen design
DSTC4	2	TXT	35	EN	state	crowd (phone calls)	tourist information
DSTC5	2	TXT	70	EN, ZH	state	crowd (phone calls)	tourist information
DSTC6-Track2	2	TXT	1,024	EN	–	social network	airline, car, retail, fast food chains, etc.
DSTC7-Track1	2	TXT	135,893	EN	–	chat logs	advising, technical support
DSTC7-Track2	2	TXT	3M	EN	–	social network	open
DSTC7-Track3	2	TXT	7,659	EN	–	crowdsourcing (chat)	short videos
EmotionLines	*	AUD+VID+TXT	2,000	EN	DAs, emotion	movie subtitles	open
ES-Port	2	TXT	1,170	ES	acoustic events	calls to technical support	telecommunications
Frames	2	TXT	1,369	EN	frames	crowd (WOZ)	vacations
KVRET	2	TXT	3031	EN	slots	crowd (WOZ)	in-car personal assistant
MapTask	2	AUD+TXT	128	EN	behaviours	crowd (phone calls)	map instructions
Mastodon	2	TXT	505	EN	DAs, polarity	social network	open
MedDialog	2	TXT	3.6M	EN, ZH	–	web forum	health
MRDA	*	AUD+TXT	71	EN	DAs	recorded meetings	topics, debates, issues, social dynamics
MultiWOZ	2	TXT	10,000	EN	state	crowd (WOZ)	7 domains
OpenSubtitles	*	TXT	152k movies	*	–	movie subtitles	open
QuAC	2	TXT	13,594	EN	DAs	crowd (student-teacher)	people
Redial	2	TXT	10,000	EN	suggested, seen, liked	crowdsourcing (chat)	movie recommendations
SAMSum	≥2	TXT	16,000	EN	summary	handcrafted by linguistics	messenger conversations
SGD	2	TXT	20,000	EN	DAs, state	crowd	20 domains
SWDA	2	AUD+TXT	1,155	EN	DAs	crowd (phone calls)	70 topics
Taskmaster-1	2	TXT	13,215	EN	API arguments	crowd	6 domains
Topical-Chat	2	TXT	11,000	EN	topic	crowd (chat)	8 topics
UDC	1-2	TXT	930,000	EN	–	chat logs	technical support
WOW	2	TXT	23,311	EN	topic	crowd (WOZ)	various topics

Acknowledgements

This work was financially supported by:

- the project FLOWANCE (POCI-01-0247-FEDER-047022), co-financed by the European Regional Development Fund (FEDER), through Portugal 2020 (PT2020), and by the Competitiveness and Internationalization Operational Programme (COMPETE 2020).
- national funds through FCT, within the scope of the project CISUC (UID/CEC/00326/2020) and by European Social Fund, through the Regional Operational Program Centro 2020.