Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1264–1274
Marseille, 20-25 June 2022

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

Abstract

Several dialogue corpora are currently available for research purposes, but they still fall short for the growing interest in the development of dialogue systems with their own specific requirements. In order to help those requiring such a corpus, this paper surveys a range of available options, in terms of aspects like speakers, size, languages, collection, annotations, and domains. Some trends are identified and possible approaches for the creation of new corpora are also discussed.

Keywords: Dialogue Corpora, Dialogue Analysis, Dialogue Systems

1. Introduction

Driven by the growing adoption of dialogue systems and conversational agents in recent years (Dale, 2016), the need for conversational data to study specific dialogue phenomena and train the aforementioned systems has also increased. However, despite the broad range of dialogue corpora currently available, many released in the last years, this number is still far from covering all the necessities of real-world applications of dialogue systems. These include all possibly envisioned dialogue analysis tasks, domains, and their specificity, among others, such as languages. Additionally, the access to dialogue data and its usage are typically very restricted, due to the logistics involved in their collection and, especially, privacy concerns. For instance, a common application of dialogue systems is automated customer-support, e.g., through call-centers. However, real conversations of this type typically include a large amount of personal data, not always easy to anonymise, as well as sensitive information such as unsatisfied customers expressing their opinion.

Hence, in most cases, there will not be a perfect corpus for a new project involving dialogue. This means that researchers will often have to choose between: (i) going through all the work involved in the collection (and annotation) of data for creating a new corpus from scratch; or (ii) looking at publicly available corpora for selecting the most suitable for their task, possibly after some adaptations.

This survey targets both previous scenarios. It is not as extensive as Serban et al. (2018), but it gives a more practical perspective, and is more up-to-date, with a focus on corpora of human-human dialogue, available in a textual format, not exclusively in English. More precisely, we compile a list of corpora of such dialogues, and put forward a brief analysis, covering the following aspects: number and type of speakers, size, languages, annotations, data collection, and contents in general. Interested researchers may use this survey as a reference for selecting appropriate corpora for their projects, but also for deciding on suitable approaches for creating new dialogue corpora.

With this analysis, we can identify trends, which end up contributing to identify gaps. There are dialogue corpora of variable sizes, even though we could think of many more, on varied domains. The largest corpora have no annotations, whereas common annotations include dialogue acts and states, and, in some cases, also sentiment-related information. Although we target corpora available as text, some are originally spoken corpora that have been transcribed. Out of those, we focus on corpora that include linguistic annotations, not dependent on the audio. Due to the constraints of using real conversations, most of the available corpora were created in a crowdsourcing environment, where workers knew they were contributing. Moreover, the majority of the corpora are in English, suggesting that working with dialogues in other languages will require a greater effort.

Next section enumerates all the surveyed corpora and gives an overview of their analysis. After that, each section of this paper focuses on one specific aspect or a group of related aspects. Section 3 is on the speakers, corpus size and covered languages. Section 4 is on the type of linguistic annotations included in the corpora, also referring some of the tasks they can be used for. Section 5 is on the approach followed for collecting their data. Section 6 is on the contents of the corpora, namely their domain and related information. Before concluding, Section 7 builds on the previous section to discuss possible options that can be taken when a suitable corpus is not available.

2. Corpora Overview

This survey targets textual corpora of dialogue between human speakers, described in the literature, and created and used for research purposes. Alphabetically listed, the following corpora were included in this survey: AMI (Carletta et al., 2005); CamRest676 (Wen et al., 2016; Wen et al., 2017); CoQA (Reddy et al., 2019); CrossWOZ (Zhu et al., 2020); DailyDialog (Li et al., 2017); DealOrNoDeal (Lewis et al., 2017); DECODA (Bechet et al., 2012); DIME (Villaseñor and Pineda, 2001); DSTC4 (Kim et al., 2017), DSTC5 (Kim et al., 2016); DSTC6, track 2 (Hori et al., 2019); DSTC7, tracks 1, 2 and 3 (D'Haro et al., 2020); EmotionLines (Hsu et al., 2018) and its extensions MELD (Poria et al., 2019) and EMOTyDa (Saha et al., 2020); ES-Port (García-Sardiña et al., 2018); Frames (El Asri et al., 2017); KVRET (Eric et al., 2017); MapTask (Anderson et al., 1991); Mastodon (Cerisara et al., 2018); MedDialog (Zeng et al., 2020); MRDA (Shriberg et al., 2004); MultiWOZ, covering version 2.0 (Budzianowski et al., 2018), as well as further corrections and new annotations introduced in versions 2.1 (Eric et al., 2020), 2.2 (Zang et al., 2020), 2.3 (Han et al., 2020) and 2.4 (Ye et al., 2021); OpenSubtitles (Tiedemann, 2009; Lison and Tiedemann, 2016); QuAC (Choi et al., 2018); Redial (Li et al., 2018); SAMsum (Gliwa et al., 2019); Schema-Guided Dialogue (SGD) (Rastogi et al., 2020); Switchboard (Godfrey et al., 1992) and SWDA (Jurafsky and Shriberg, 1997); Taskmaster-1 (Byrne et al., 2019); Topical-Chat (Gopalakrishnan et al., 2019); Ubuntu Dialog Corpus (UDC) (Lowe et al., 2015); and Wizard of Wikipedia (WOW) (Dinan et al., 2018).

Table 1 presents an overview of the results of this survey, with every corpora and their surveyed properties, namely: the number of speakers (where * stands for multiple or a variable number of speakers); the modalities they are available on (AUD for audio, TXT for text, VID for video); number of dialogues (when not available, alternative information is presented); language(s) covered (where * stands for multiple languages); linguistic annotations included; collection process; and domain.

The table shows right away that the surveyed corpora are significantly different in several aspects, from their size, to the covered domains. In the next sections we analyse some of these aspects.

3. Speakers, Size and Languages

The majority of these corpora is restricted to conversations between two human speakers, but some may include dialogues with more. This happens for AMI and MRDA, which have dialogues from meetings with several participants, or EmotionLines and OpenSubtitles, which have movie dialogues with a variable number of participants. SAMsum may also include dialogues with more than two speakers. On the other hand, UDC may include posts with no answer, and thus a single speaker.

Dialogue corpora are commonly used for training dialogue systems to respond as if they were humans, and so it makes sense that the dialogues they contain are indeed between humans. However, there are also corpora of conversations between humans and dialogue systems, typically including linguistic annotations, often made or revised by humans. These include corpora like ATIS (Price, 1990), DIHANA (Alcácer et al., 2005), LEGO (Schmitt et al., 2012), dialogues collected during the ConvAI challenges (Burtsev et al., 2018; Logacheva et al., 2020), and the corpora used in some editions of the Dialog State Tracking Challenge (DSTC) (Williams et al., 2013; Henderson et al., 2014a; Henderson et al., 2014b) or of its later rebranding as Dialog System Technology Challenge (Hori et al., 2019). The value of such corpora lies mainly in their annotations, useful for automating the process of dialogue analysis, which may later be useful for improving dialogue systems. For more on linguistic annotations in dialogue corpora, see section 4. Although we may refer to some of the previous corpora in the following sections, the main focus of this survey are corpora of dialogues exclusively between humans.

We find corpora of significantly variable sizes, ranging from just a few dozens (DIME, MapTask, MRDA, DSTC4, DSTC5) to thousands of dialogues. Of course, the size of the corpora is heavily constrained by the source of the data, domain covered and linguistic annotations made. Due to its nature, we do not have the number of dialogues in OpenSubtitles, but it is for sure the largest surveyed corpus, because it includes the subtitles for 152 thousand movies, in many languages, and each movie will have many dialogues. These dialogues are not annotated and there are many available on the web, produced by volunteers. Other large corpus are MedDialog, DSTC7-Track2 and UDC, respectively from a web forum, a social network, and chat logs, none of them with linguistic annotations. More on annotations, collection and domains can be found respectively in Sections 4, 5 and 6.

The majority of the surveyed corpora are in English, but there is a minority in other languages, namely French (DECODA), Spanish (DIME, ES-Port), or Chinese (CrossWOZ, MedDialog, DSTC5). The coverage of other languages is slightly broader if we consider human-machine dialogues, for which there are more corpora in Spanish (e.g., DIHANA) or in Japanese (e.g., DSTC6-Track3). Given its nature, OpenSubtitles covers a total of 60 languages and is available as parallel corpora. This means that researchers willing to work on different languages will probably have to find another corpus (e.g., private) or creating their own. In this case, the range of data collection approaches and domains covered by the surveyed corpora may serve as inspiration. For a discussion of available options, see section 7.

4. Linguistic Annotations

Most of the tasks in the scope of dialogue analysis benefit from having a dialogue corpus where computational models can be learned from. In some cases, annotations are not even necessary, as it happens for data-driven response generation, often framed as a problem of learning to translate user interactions to suitable re-

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	$\geq$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

Table 1: Overview of dialogue corpora, where * stands for multiple values.

sponses. It is thus no surprise that machine translation approaches have been applied for generating responses to tweets (Ritter et al., 2011) and for modelling conversations (Vinyals and Le, 2015). Corpora without annotations may also be used in unsupervised approaches, e.g., clustering utterances according to the intent and discovering dialogue flows (Ritter et al., 2010).

However, whereas response generation systems do not always require a great understanding of the dialogue, when it comes to task-oriented dialogue systems, it is important to represent the current state of the dialogue at each turn. This comes as a set of linguistic annotations, which can be useful for training dialogue systems, and may also serve as common benchmarks, thus helping to evaluate progress in the task of their automatic prediction.

4.1. Dialogue State Tracking

Some corpora are annotated with useful information for a dialog-state architecture (Williams et al., 2016), where the meaning of the user's intents is represented as slots and their values (e.g., `from:downtown`, `inform:price=cheap`), grouped in a task-related frame where some of the values might be empty. The history of the dialogue may then be represented by a sequence of such frames.

Annotations for the previous tasks are included in the corpora of the first five editions of the DSTC, even if some are between humans and dialogue systems, but also for each customer utterance in MultiWOZ, CamRest676, CrossWOZ, and SGD, all including conversations between humans. KVRET adopts an alternative annotation, where the knowledge of the utterances is represented in the form of key-value pairs, directly converted to triples (e.g., `<dinner, time, 8pm>`). Towards a simpler annotation, Taskmaster restricts annotations according to the type of conversation, more precisely, to the variables required to execute the transaction, which the authors refer as API arguments. The Frames corpus goes beyond dialogue tracking and includes frame tracking annotations, which the authors claim to capture more complex dialogue flows, where it might be necessary to simultaneously keep track of several frames.

4.2. Dialogue Acts

Having in mind that each utterance in a dialogue is a kind of action performed by the speaker (Austin, 1962), roughly the speaker's intention, a simpler insight on each utterance is the dialogue act (DA), also known as speech act. Another common task in dialogue analysis is thus DA classification (DAC), which consists of labelling each utterance in a dialogue according to its DA. Given that the current DA often depends on the previous, DAC can be framed as a sequence labelling problem, instead of a simple classification task. It is thus no surprise that it has been attempted with Hidden Markov Models (Stolcke et al., 2000), Conditional Random Fields (Kim et al., 2010), or Recurrent Neural Networks with a LSTM layer (Kumar et al., 2018), among others.

There are domain-independent taxonomies of DAs (including, e.g., `greeting`, `question` or `acknowledge`), but the DA may also depend on the domain and on the task at hand (e.g., `request-address`, `inform-food`). On this context, the Dialog Act Markup in Several Layers (DAMSL) (Core and Allen, 1997) organises domain-independent DAs in four main categories (communicative, information, forward-looking, backward-looking) and has been frequently adopted. SWDA is one of the most popular corpus annotated with (an augmented version of) DAMSL, but other corpora adopt DAMSL or a subset of its tags, namely DIME, EMOTyDA, Mastodon and MRDA.

The utterances of DailyDialog are annotated according to four DA classes interoperable across multiple domains (Amanova et al., 2016), roughly matching the four main categories of DAMSL: inform, question, directive, commissive. Other corpora of human dialogues are annotated with typologies of DAs adapted to their tasks, namely AMI, MapTask, QuAC. There are also corpora of human-machine dialogues with annotated DAs, namely LEGO and DIHANA. The latter has the particularity of adopting a three-level hierarchy for DAs (speech act, data used / modified, specific data).

The DA is generally also included in datasets with dialogue tracking annotations (see e.g., MultiWOZ 2.1, CamRest676, SGD), which may thus be used for the task of DAC as well. However, while the dialogue state can be performed implicitly during the dialogue (e.g., in order to fulfil a given task, one of the human agents adds and fills slots in a form), the annotation of the DA is typically done at a later stage, and resorts to human annotators (Budzianowski et al., 2018).

4.3. Sentiment and Emotion

Due to their relevance for improving client-customer interactions, or predicting user opinions, another type of annotation in dialogue corpora is related to sentiment and emotion. In Mastodon, each utterance has an assigned polarity (positive, negative, neutral), and there are corpora where each utterance has an emotion label (DailyDialog, EmotionLines), corresponding to one of the six basic Ekman's emotions (Ekman, 1999) (joy, sadness, anger, fear, disgust, surprise), plus neutral. Given its correlation with DAs, sentiment classification in dialogues has been attempted jointly with DAC (Qin et al., 2020) and emotions have also proved to be useful for the latter task (Saha et al., 2021).

4.4. Other Annotations

Some corpora of spoken conversation also include annotations based on the actual speech and acoustic features, which may be related to emotion, but also cover behaviours like filled pauses, false starts and repetitions, broken words (MapTask, ES-Port); background noise and laughter (ES-Port); head movements (AMI); or discourse referents (DIME). Still, since we are focusing on textual corpora, we did not look deep into these. Moreover, the transcriptions of the DECODA corpus have several semantic (call type) and syntactic annotations (disfluencies, parts-of-speech, chunks, named entities, syntactic dependencies).

We can say that the previous are the most common linguistic annotations we find in dialogue corpora. However, there are corpora with alternative annotations, such as summaries of the conversation (SAMsum), useful for developing abstractive summarisation tools; movies mentioned in a conversation and whether they had been suggested, seen or liked (Redial), useful, e.g., for conversational recommender systems (Jannach et al., 2021); the topic of the conversation (WOW, Topical-Chat), useful for developing systems capable of chatting on a close domain of knowledge; medical diagnosis and treatment suggestions (in some conversations of MedDialog); or the rationale for question-answering (QuAC).

We included in this survey two corpora with questions and answers in a conversation scenario (CoQA, QuAC). Each dialogue starts with a textual passage about which one of the speakers asks questions, while the other provides the answers, based on the aforementioned passage. These corpora are typically used for assessing approaches for extractive question-answering, which can be tackled by fine-tuning a neural language model like BERT (Devlin et al., 2019). However, they add the necessity of considering the conversation context for properly interpreting the questions.

5. Data Collection

One of the main goals of dialogue corpora is to provide training data for dialogue systems, so they can mimic human behaviours in certain tasks and scenarios. However, due to privacy reasons, only a minority of the surveyed dialogue corpora are collected from real situations (DECODA, ES-Por, MRDA). In the majority of the corpora, if not all, the speakers knew that their conversation was being recorded, which could, of course, condition their behaviour. But this is a necessary trade-off.

Several dialogue corpora are the result of spoken conversations and are thus available in the audio form. This is the case of some of the surveyed corpora (AMI, DECODA, DIME, MapTask, MRDA, Switchboard). However, for all the previous, manual or automatic transcriptions are also provided, meaning that they are in fact multimodal. In addition to audio and text, MELD and EMOTyDA add video to EmotionLines. Despite resulting from phone calls, due to the presence of sensitive data, only the (anonymised) transcription are available for ES-Port, not the audio. The same happens for DSTC4 and DSTC5.

More than knowing that the conversation was being recorded, in many corpora, the speakers were following a script and conversing with the single purpose of creating the corpus. Several of these follow the Wizard of Oz (WOZ) (Kelley, 1984) paradigm, where a conversation occurs between two speakers with different roles. One of them will play the role of a common user, also known as the questioner, to which a task is given (e.g., to find an entity, to chat about some topic). In order to fulfil this task, the questioner must interact, using natural language, with another user, the wizard, also known as the answerer, which will have access to more information on the domain (e.g., a database or a longer collection of documents) and possibly an interface that will help them track the user requests and efficiently provide suitable answers. This paradigm has been followed in the creation of spoken corpora (DIME), also including human-machine dialogues (ATIS, DIHANA), but mostly written conversation corpora (CrossWOZ, CamRest676, MultiWOZ, Frames, KVRET, WOW, Taskmaster). Other corpora created from crowdsourcing include DealOrNoDeal, Redial, Topical-Chat, CoQA and QuAC. All of the latter rely on the Amazon Mechanical Turk[1] platform, where tasks are prepared and workers are recruited to perform them in exchange of a monetary compensation.

Alternative sources to the WOZ paradigm include the exploitation of Web sources where users engage in conversations, including chat logs (UDC), forums (MedDialog), language learning websites (DailyDialog) and, of course, social networks. Due to the common presence of short-message dialogues, as well as its flexible API, Twitter has been used as the source of conversations (Ritter et al., 2011; Hori et al., 2019). However, due to its privacy-protecting restrictions, some researchers looked for similar but more relaxed sources, specifically Mastodon.

Movie subtitles, openly available in a large number in various websites, are another source of conversations. OpenSubtitles is a large corpus of this kind. Another example is the EmotionLines corpus, which is much smaller, focused on a single TV show (Friends), but with annotated emotions.

Approaches with simpler logistics were adopted for the creation of SAMSum, where conversations in the style of messenger apps were created and written down by linguists; part of Taskmaster, where dialogues were created by a single user, writing turns for both speakers, following a conversational scenario; or SGD, where conversations follow predefined flows, translated to natural language by crowd workers.

6. Domains & Contents

Humans can converse about virtually any topic, for their daily communication, regarding a specific situation, in order to achieve a predefined goal, or just for entertainment, with no specific purpose in mind.

[1] https://www.mturk.com/

For different topics and domains, a different vocabulary will be used, and the goal to achieve will constrain the conversation significantly. It is thus important to have corpora covering as much of such situations, topics and domains as possible, enabling different kinds of analysis and the development of dialogue systems with different purposes.

Despite the growing number of domains for which dialogue corpora are available, there will always be domains or specific applications, including different languages, for which such a corpus is nonexistent. In some cases, existing corpora may be still used or adapted, but in others new corpora will have to be created.

We identified several task-oriented corpora, namely those with dialogues between a customer and an agent. Many are focused on travelling or tourism-related domains, with customers asking for information such as locations, times, or suggestions that match their requirements. This includes vacations in general (Frames), restaurants (CamRest676), or urban transports (DECODA). When considering human-machine conversations, related domains include flights (ATIS), trains (DIHANA) and buses (LEGO). Another identified domain was in-car personal assistance (KVRET).

But some corpora cover more than one domain, often including but not restricted to tourism. Examples are restaurants, hotels, attractions, taxis, trains, hospitals and police (MultiWOZ); hotels, restaurants, attractions, metros, and taxis (CrossWOZ); hotels, flights, and car rentals (DSTC4, DTSC5); ordering pizza, creating auto repair appointments, setting up ride service, ordering movie tickets, ordering coffee drinks and making restaurant reservations (Taskmaster); or a range of 20 domains, including some of the previous, and others, such as alarm, calendar, events, media, messaging, movies or weather (SGD).

Though less restricted on the kind of possible requests, there are corpora roughly focused on other domains like customer-support in the technological (UDC, DSTC7-Track1) or telecommunications (ES-Port), as well as conversations between patients and doctors (MedDialog). The latter is especially interesting because health is a sensitive domain for which it is difficult to get data of this kind.

Some corpora cover tasks that involve the cooperation between two speakers. In DIME, the cooperative task is around kitchen design, and the speakers can refer to objects through a graphical user interface. In MapTask, both speakers have a map, but only one has a route, which is to be instructed to the other to follow. In DealOrNoDeal, dialogues are on a multi-issue bargaining task, i.e., several items are available and human agents have to converse towards an agreement on the distribution of these items among them. In this process, agents try to maximise their reward function, different for each agent and not visible to the other.

Though not task-oriented, there are corpora with conversations on varied topics (WOW, Topical-Chat) or focused movies and recommendations (Redial); of questions and their answers (CoQA, QuAC); as well as of debates (MRDA, AMI) on different domains.

In addition to not being task-oriented, some corpora can be considered to be open domain. These include conversations from social networks (Mastodon, DSTC6-Track2, DSTC7-Track2); conversations in the style of those exchanged in messenger applications (SAMSum); daily life conversations in a website where English learners practice (DailyDialog); or TV show or movie subtitles (OpenSubtitles, EmotionLines).

7. Discussion

Despite the broad list of identified corpora, scenarios for which none of them applies can be easily envisioned. When this happens, an alternative would be to look at the available corpora and select the closest to the task at hand, possibly considering some adaptations. Otherwise, one will have to look at available sources of conversation data, get inspiration from the creation of the available dialogue corpora, and consider the creation of a new corpus.

In projects involving companies, it is normal that there is real data, which can and should be used, but, due to privacy legislation, its public release will generally not be possible. As we have shown in section 5, alternative sources of conversation data, available in many languages, include movie subtitles or web sources where people may leave their comments and establish conversations, like forums or social networks. Still, despite being publicly available through APIs or web scrapping, privacy laws might also apply and, in some cases, prevent the public distribution of the data as a corpus, even if for research purposes.

Another option would be to create all the data, from scratch. If resources are available, one may devise a crowdsourcing task, e.g., following the WOZ paradigm. Conversations would result from the interactions between paid workers, possibly following guidelines regarding the kind of dialogue to simulate. For task-oriented systems, to be used in customer-support, a speaker performing the role of customer / questioner would ask questions to the other, which would have access to more information, e.g., in the form of a database or a collection of textual documents. Possibly not as natural, but more practical, would be to adopt self-dialog, as in the Taskmaster corpus, i.e., a single user following a given scenario and writing all the turns of a dialogue. In any case, the previous approaches will condition the conversation, not desired for spontaneous speech, but acceptable for task-oriented dialogue.

As discussed in Section 4, for most dialogue analysis tasks, data alone is not enough, and linguistic annotations have to be made. Since these annotations are generally based on human theories, and they will be used for training and evaluating computational systems, most of the times they will have to be added manually. Depending on the available resources, as well as on the difficulty of explaining the task, they can be done by linguists, domain experts, or crowd workers. An alternative is to first produce the dialogue flows first, including all necessary annotations, and use them as the guidelines for a process like WOZ, i.e., ask humans to engage in a conversation that follows such flows. In the end, one could assume that the annotations would be valid for the resulting text.

Still regarding annotations, if the only problem is the language (e.g., there is a corpus for the target task, but it is not in the desired language), an option would be to translate the corpus automatically to the target language, while keeping the annotations. However, this is rarely a good option, due to potential issues on machine translation, including some specificities of each language that would hardly be captured, resulting in less natural text, and possibly incoherent annotations.

8. Conclusion

This was a brief survey on dialogue corpora available as text, with a focus on those between human speakers, that are publicly available or, according to the authors, can be provided for research purposes. We tried to cover a broad range of distinct corpora, including different languages, tasks, domains, and creation approaches, but were not so strict, so it is likely that we left out some corpora that matched our search criteria.

This survey may be useful for researchers working on projects involving the analysis of written dialogue, including, but not limited to, the development of dialogue systems. With this analysis, we could identify corpora of variable sizes and languages, created through different approaches, and covering different tasks and domains, but we could think of many more. The largest corpora have no annotations, whereas common annotations include dialogue acts and states, but in some cases also the sentiment tags. Although we focus on corpora available as text, some are originally spoken corpora that have been transcribed. Due to privacy constrains, most of the corpora were created in a crowdsourcing environment, where speakers are guided by assigned roles and tasks to accomplish.

Moreover, the majority of the corpora are in English, meaning that working with dialogues in other languages will generally require a greater effort, both on the collection of data and on its annotation. In most cases, this does not exactly mean that corpora is not available for other languages, only that the usage of such corpora is restricted and cannot be made publicly available. Consequently, there are no common benchmarks for those languages, making it harder to compare different works and to measure progress in the area. On top of this, interested researchers will possibly have to multiply the effort involved in data annotation.

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); and by 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.

9. Bibliographical References

Amanova, D., Petukhova, V., and Klakow, D. (2016). Creating annotated dialogue resources: Cross-domain dialogue act classification. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 111–117, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Austin, J. L. (1962). *How to do things with words: the William James lectures delivered at Harvard University in 1955*. Oxford University Press.

Burtsev, M., Logacheva, V., Malykh, V., Serban, I. V., Lowe, R., Prabhumoye, S., Black, A. W., Rudnicky, A., and Bengio, Y. (2018). The first conversational intelligence challenge. In *The NIPS'17 Competition: Building Intelligent Systems*, pages 25–46. Springer.

Core, M. G. and Allen, J. (1997). Coding dialogs with the DAMSL annotation scheme. In *AAAI fall symposium on communicative action in humans and machines*, volume 56, pages 28–35. Boston, MA.

Dale, R. (2016). The return of the chatbots. *Natural Language Engineering*, 22(5):811–817.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Procs. of 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. ACL Press.

Ekman, P. (1999). Basic emotions. *Handbook of cognition and emotion*, pages 45–60.

Henderson, M., Thomson, B., and Williams, J. D. (2014a). The second Dialog State Tracking Challenge. In *Proceedings of the 15th annual meeting of the special interest group on discourse and dialogue (SIGDIAL)*, pages 263–272.

Henderson, M., Thomson, B., and Williams, J. D. (2014b). The third Dialog State Tracking Challenge. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 324–329. IEEE.

Jannach, D., Manzoor, A., Cai, W., and Chen, L. (2021). A survey on conversational recommender systems. *ACM Comput. Surv.*, 54(5), May.

Kelley, J. F. (1984). An iterative design methodology for user-friendly natural language office information applications. *ACM Transactions on Information Systems (TOIS)*, 2(1):26–41.

Kim, S. N., Cavedon, L., and Baldwin, T. (2010). Classifying dialogue acts in one-on-one live chats. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 862–871.

Kumar, H., Agarwal, A., Dasgupta, R., and Joshi, S. (2018). Dialogue act sequence labeling using hierarchical encoder with CRF. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Logacheva, V., Malykh, V., Litinsky, A., and Burtsev, M. (2020). Convai2 dataset of non-goal-oriented human-to-bot dialogues. In *The NeurIPS'18 Competition*, pages 277–294. Springer.

Qin, L., Che, W., Li, Y., Ni, M., and Liu, T. (2020). Dcr-net: A deep co-interactive relation network for joint dialog act recognition and sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8665–8672.

Ritter, A., Cherry, C., and Dolan, W. B. (2010). Unsupervised modeling of Twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 172–180.

Ritter, A., Cherry, C., and Dolan, W. B. (2011). Data-driven response generation in social media. In *Proceedings of the conference on Empirical Methods in Natural Language Processing*, pages 583–593. Association for Computational Linguistics.

Saha, T., Gupta, D., Saha, S., and Bhattacharyya, P. (2021). Emotion aided dialogue act classification for task-independent conversations in a multi-modal framework. *Cognitive Computation*, 13(2):277–289.

Serban, I. V., Lowe, R., Henderson, P., Charlin, L., and Pineau, J. (2018). A survey of available corpora for building data-driven dialogue systems: The journal version. *Dialogue & Discourse*, 9(1):1–49.

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Ess-Dykema, C. V., and Meteer, M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

Vinyals, O. and Le, Q. V. (2015). A neural conversational model. In *Proceedings of ICML 2015 Deep Learning Workshop*, Lille, France.

Williams, J., Raux, A., Ramachandran, D., and Black, A. (2013). The Dialog State Tracking Challenge. In *Proceedings of the SIGDIAL 2013 Conference*, pages 404–413, Metz, France, August. Association for Computational Linguistics.

Williams, J. D., Raux, A., and Henderson, M. (2016). The dialog state tracking challenge series: A review. *Dialogue & Discourse*, 7(3):4–33.

10. Language Resource References

Alcácer, N., Benedı, J., Blat, F., Granell, R., Martınez, C., and Torres, F. (2005). Acquisition and labelling of a spontaneous speech dialogue corpus. In *Proceeding of 10th International Conference on Speech and Computer (SPECOM). Patras, Greece*, pages 583–586.

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., et al. (1991). The HCRC map task corpus. *Language and speech*, 34(4):351–366.

Bechet, F., Maza, B., Bigouroux, N., Bazillon, T., El-Bèze, M., De Mori, R., and Arbillot, E. (2012). DECODA: a call-centre human-human spoken conversation corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1343–1347, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Budzianowski, P., Wen, T.-H., Tseng, B.-H., Casanueva, I., Ultes, S., Ramadan, O., and Gasic, M. (2018). MultiWOZ – a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026.

Byrne, B., Krishnamoorthi, K., Sankar, C., Neelakantan, A., Goodrich, B., Duckworth, D., Yavuz, S., Dubey, A., Kim, K.-Y., and Cedilnik, A. (2019). Taskmaster-1: Toward a realistic and diverse dialog dataset. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4516–4525.

Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., et al. (2005). The AMI meeting corpus: A pre-announcement. In *International workshop on machine learning for multimodal interaction*, pages 28–39. Springer.

Cerisara, C., Jafaritazehjani, S., Oluokun, A., and Le, H. T. (2018). Multi-task dialog act and sentiment recognition on Mastodon. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 745–754.

Choi, E., He, H., Iyyer, M., Yatskar, M., Yih, W.-t., Choi, Y., Liang, P., and Zettlemoyer, L. (2018). QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium, October-November. Association for Computational Linguistics.

Dinan, E., Roller, S., Shuster, K., Fan, A., Auli, M., and Weston, J. (2018). Wizard of Wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*.

D'Haro, L. F., Yoshino, K., Hori, C., Marks, T. K., Polymenakos, L., Kummerfeld, J. K., Galley, M., and Gao, X. (2020). Overview of the seventh dialog system technology challenge: DSTC7. *Computer Speech & Language*, 62:101068.

El Asri, L., Schulz, H., Sarma, S. K., Zumer, J., Harris, J., Fine, E., Mehrotra, R., and Suleman, K. (2017). Frames: a corpus for adding memory to goal-oriented dialogue systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 207–219.

Eric, M., Krishnan, L., Charette, F., and Manning, C. D. (2017). Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 37–49.

Eric, M., Goel, R., Paul, S., Sethi, A., Agarwal, S., Gao, S., Kumar, A., Goyal, A., Ku, P., and Hakkani-Tur, D. (2020). MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 422–428, Marseille, France, May. European Language Resources Association.

García-Sardiña, L., Serras, M., and del Pozo, A. (2018). ES-port: a spontaneous spoken human-human technical support corpus for dialogue research in Spanish. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Gliwa, B., Mochol, I., Biesek, M., and Wawer, A. (2019). SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China, November. Association for Computational Linguistics.

Godfrey, J. J., Holliman, E. C., and McDaniel, J. (1992). Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, volume 1, pages 517–520. IEEE Computer Society.

Gopalakrishnan, K., Hedayatnia, B., Chen, Q., Gottardi, A., Kwatra, S., Venkatesh, A., Gabriel, R., and Hakkani-Tür, D. (2019). Topical-Chat: Towards knowledge-grounded open-domain conversations. In *Proc. Interspeech 2019*, pages 1891–1895.

Han, T., Liu, X., Takanobu, R., Lian, Y., Huang, C., Peng, W., and Huang, M. (2020). MultiWOZ 2.3: A multi-domain task-oriented dataset enhanced with annotation corrections and co-reference annotation. *arXiv preprint arXiv:2010.05594*.

Hori, C., Perez, J., Higashinaka, R., Hori, T., Boureau, Y.-L., Inaba, M., Tsunomori, Y., Takahashi, T., Yoshino, K., and Kim, S. (2019). Overview of the sixth dialog system technology challenge: DSTC6. *Computer Speech & Language*, 55:1–25.

Hsu, C.-C., Chen, S.-Y., Kuo, C.-C., Huang, T.-H., and Ku, L.-W. (2018). EmotionLines: An emotion corpus of multi-party conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Jurafsky, D. and Shriberg, E. (1997). Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual. Technical report, University of Colorado at Boulder &+ SRI International.

Kim, S., D'Haro, L. F., Banchs, R. E., Williams, J. D., Henderson, M., and Yoshino, K. (2016). The fifth Dialog State Tracking Challenge. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 511–517. IEEE.

Kim, S., D'Haro, L. F., Banchs, R. E., Williams, J. D., and Henderson, M. (2017). The fourth Dialog State Tracking Challenge. In *Dialogues with Social Robots*, pages 435–449. Springer.

Lewis, M., Yarats, D., Dauphin, Y., Parikh, D., and Batra, D. (2017). Deal or no deal? end-to-end learning of negotiation dialogues. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2443–2453.

Li, Y., Su, H., Shen, X., Li, W., Cao, Z., and Niu, S. (2017). DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995.

Li, R., Kahou, S. E., Schulz, H., Michalski, V., Charlin, L., and Pal, C. (2018). Towards deep conversational recommendations. In Samy Bengio, et al., editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 9748–9758.

Lison, P. and Tiedemann, J. (2016). OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Lowe, R., Pow, N., Serban, I. V., and Pineau, J. (2015). The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294.

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., and Mihalcea, R. (2019). MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536.

Price, P. (1990). Evaluation of spoken language systems: The ATIS domain. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*.

Rastogi, A., Zang, X., Sunkara, S., Gupta, R., and Khaitan, P. (2020). Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8689–8696.

Reddy, S., Chen, D., and Manning, C. D. (2019). CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Saha, T., Patra, A., Saha, S., and Bhattacharyya, P. (2020). Towards emotion-aided multi-modal dialogue act classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4361–4372.

Schmitt, A., Ultes, S., and Minker, W. (2012). A parameterized and annotated spoken dialog corpus of the CMU Let's Go bus information system. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3369–3373, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Shriberg, E., Dhillon, R., Bhagat, S., Ang, J., and Carvey, H. (2004). The ICSI meeting recorder dialog act (MRDA) corpus. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 97–100.

Tiedemann, J. (2009). News from OPUS-a collection of multilingual parallel corpora with tools and interfaces. In *Recent advances in natural language processing*, volume 5, pages 237–248.

Villaseñor, Luis, A. M. and Pineda, L. A. (2001). The DIME corpus. In *Encuentro Internacional de Ciencias de la Computación, vol. 2, 1–10*.

Wen, T.-H., Gasic, M., Mrkšić, N., Barahona, L. M. R., Su, P.-H., Ultes, S., Vandyke, D., and Young, S. (2016). Conditional generation and snapshot learning in neural dialogue systems. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2153–2162.

Wen, T.-H., Vandyke, D., Mrkšić, N., Gašić, M., Rojas-Barahona, L. M., Su, P.-H., Ultes, S., and Young, S. (2017). A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449, Valencia, Spain, April. Association for Computational Linguistics.

Ye, F., Manotumruksa, J., and Yilmaz, E. (2021). Multiwoz 2.4: A multi-domain task-oriented dialogue dataset with essential annotation corrections to improve state tracking evaluation. *arXiv preprint arXiv:2104.00773*.

Zang, X., Rastogi, A., Sunkara, S., Gupta, R., Zhang, J., and Chen, J. (2020). MultiWOZ 2.2: A dialogue dataset with additional annotation corrections and state tracking baselines. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 109–117, Online, July. Association for Computational Linguistics.

Zeng, G., Yang, W., Ju, Z., Yang, Y., Wang, S., Zhang, R., Zhou, M., Zeng, J., Dong, X., Zhang, R., et al. (2020). MedDialog: A large-scale medical dialogue dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250.

Zhu, Q., Huang, K., Zhang, Z., Zhu, X., and Huang, M. (2020). CrossWOZ: A large-scale Chinese cross-domain task-oriented dialogue dataset. *Transactions of the Association for Computational Linguistics*, 8:281–295.

A. Corpora URLs

Table 2 compiles the URLs where the surveyed corpora can be downloaded from, as of December 2021. Out of them, DECODA is not available because, according to its authors, includes much personal data, hard to anonymise. Moreover, DSTC4 and DSTC5 have restricted availability, exclusive to the participants in the challenges where they were used. For ES-Port, we found two download links, but none was working in April 2022.

Name	**URL**
AMI	https://groups.inf.ed.ac.uk/ami/download/
CamRest676	https://github.com/WING-NUS/sequicity/tree/master/data/CamRest676
CoQA	https://stanfordnlp.github.io/coqa/
CrossWOZ	https://github.com/thu-coai/CrossWOZ
DealOrNoDeal	https://github.com/facebookresearch/end-to-end-negotiator
DailyDialog	http://yanran.li/dailydialog.html
DECODA	N/A
DIME	https://data.stanford.edu/dime
DSTC4	https://colips.org/workshop/dstc4/ (restricted to participants in the challenge)
DSTC5	https://github.com/seokhwankim/dstc5 (restricted to participants in the challenge)
DSTC6	https://github.com/dialogtekgeek/DSTC6-End-to-End-Conversation-Modeling
DSTC7	https://github.com/mgalley/DSTC7-End-to-End-Conversation-Modeling
EmotionLines	http://doraemon.iis.sinica.edu.tw/emotionlines/index.html
ES-Port	https://aholab.ehu.eus/metashare/repository/browse/ es-port-the-spanish-technical-support-corpus/ b5643034100f11e8b066f01faff11afaa7fc3195f8a64c929bb313140a170cdb/ (link not working)
Frames	https://www.microsoft.com/en-us/research/project/frames-dataset/
KVRET	https://metatext.io/datasets/a-multi-turn, -multi-domain-dialogue-dataset-(kvret)
Mastodon	https://github.com/cerisara/DialogSentimentMastodon
MedDialog	https://github.com/UCSD-AI4H/Medical-Dialogue-System
MRDA	https://github.com/NathanDuran/MRDA-Corpus
MapTask	https://groups.inf.ed.ac.uk/maptask/
MultiWOZ	https://github.com/budzianowski/multiwoz
OpenSubtitles	https://opus.nlpl.eu/OpenSubtitles-v2018.php
QuAC	http://quac.ai/
Redial	https://redialdata.github.io/website/
SAMsum	https://arxiv.org/abs/1911.12237
SGD	https://github.com/google-research-datasets/ dstc8-schema-guided-dialogue
SWDA	http://compprag.christopherpotts.net/swda.html
Taskmaster-1	https://research.google/tools/datasets/taskmaster-1/
Topical-Chat	https://github.com/alexa/Topical-Chat
UDC	https://github.com/rkadlec/ubuntu-ranking-dataset-creator
WOW	https://parl.ai/projects/wizard_of_wikipedia/

Table 2: URLs of the dialogue datasets, as of December 2021.