Identification of Personal Information Shared in Chat-Oriented Dialogue

Sarah Fillwock*, **David Traum**†
*Michigan State University
428 S. Shaw Lane, East Lansing, MI 48824
fillwoc2@msu.edu

†Institute for Creative Technologies, University of Southern California
12015 Waterfront Dr, Playa Vista, CA 90094
traum@ict.usc.edu

Abstract

We present an analysis of how personal information is shared in chat-oriented dialogue. We develop an annotation scheme, including entity-types, attributes, and values, that can be used to annotate the presence and type of personal information in these dialogues. A collection of attribute types is identified from the annotation of three chat-oriented dialogue corpora and a taxonomy of personal information pertinent to chat-oriented dialogue is presented. We examine similarities and differences in the frequency of specific attributes in the three corpora and observe that there is much overlap between the attribute types which are shared between dialogue participants in these different settings. The work presented here suggests that there is a common set of attribute types that frequently occur within chat-oriented dialogue in general. This resource can be used in the development of chat-oriented dialogue systems by providing common topics that a dialogue system should be able to talk about.

Keywords: dialogue, personal information, attributes, knowledge base, chat-oriented, user model

1. Introduction

Spoken dialogue has been studied from the perspectives of many different disciplines, with a primary focus on analyzing the turn-taking procedure of dialogue, the effects of social context, and the underlying structure of sequences of dialogue (Eggins and Slade, 1997). Within this body of work, spoken dialogue is often divided into two major categories: task-oriented dialogues and chat-oriented dialogues. Although task-oriented dialogues have received the most attention historically, interest in chat-oriented dialogue has been on the rise in recent decades. One major factor in this increase of interest has been the popularity of artificial conversational dialogue agents.

As defined by Eggins and Slade (1997), chat-oriented dialogue - or casual conversation as they name it - is characterized by its topic flexibility, the informal nature of the exchange, and the fact that the participants are not trying to accomplish any particular functional task through their dialogue. Aiming to overcome the seeming open-domain generality of chat-oriented dialogues, much of the effort has been placed into identifying the sub-types of chat-oriented dialogue and characterizing the general structure of these sub-types (Eggins and Slade, 1997; Slade and Gardner, 1993; de Silva Joyce and Slade, 2000).

Although there has been much work on defining and analyzing the characteristic structure of chat-oriented dialogues, there has been less work on exploring the content of these dialogues in terms of their characteristic topics. An understanding of the topics that are commonly focused on during chat-oriented dialogues would provide both greater insight into the function of chat-oriented dialogues on an inter-personal level and also allow for interesting analyses based on the popular topics of conversation. One promising area of inquiry is the sharing of personal information between the conversational participants.It has been observed that people frequently focus on sharing personal information about themselves in chat-oriented dialogues (Mitsuda et al., 2017). Consequently, this provides evidence that focusing on topics related to personal information will cover a significant portion of topics relevant to chat-oriented dialogue.

Our goal in this work was to identify the different types of personal information that people share in chat-oriented dialogue, and to investigate how communication of personal information occurs in different dialogue activities. In order to do this, we annotated the utterances in human dialogues with the information that the speaker was sharing about themselves or about entities close to them, such as family members, friends, and organizations. We did this for three different dialogue datasets. These annotations were used to determine categories of information that a person can share, and our annotation results were compared across the three different dialogue datasets to find general patterns for all chat-oriented dialogues, as well as information-sharing variations that occurred in the different corpora.

2. Related Work

Much of the previous work in chat-oriented dialogue has focused on identifying its structure - on both the micro level in terms of grammatical patterns and speech functions and the macro level in terms of conversational stages and genres (Eggins and Slade, 1997). As Slade and Gardner (1993) note, one proposal from Suzanne Eggins is that different instances of chat-oriented dialogue differ based on qualities of the power relationships, amount of contact, and emotional attachment between the conversational participants. de Silva Joyce and Slade (2000) go on to further describe Eggins' definition of two subcategories of casual conversation: polite - where there is limited contact between the participants outside of their current conversation - and confirming - where the participants are in close contact and

have developed an emotional attachment.

Within casual conversations, Eggins and Slade (1997) have argued that the interaction takes the form of sequences of chunks and chats. They define chunks as the more structured, monologue-like interactions in conversations where one party tends to dominate, and chats as the highly interactive segments of conversation where multiple speakers tend to be involved in the conversation and compete for turns. They further conceptualize the idea of 'chunks' by defining different categories of interaction that occur in these segments - such as gossip, anecdote, and joke-telling - and breaking down each category into a structured sequence of conversational stages.

Although these previous works examine the content of the chat-oriented dialogues in their work on understanding the dialogue structure, the focus is not on identifying and understanding what people tend to talk about in their casual conversations. Other work has taken a different approach to studying chat-oriented dialogue by focusing on the progression of topics and its effect on the conversational participants' experience in the interaction.

Previous work from Hirano et al. (2016) has investigated different personalization strategies in chat-oriented dialogue and found that the strategies related to topic elaboration and topic changing significantly increased the satisfaction of the conversational participants. A major component of these strategies is the ability to use the information known about one dialogue participant to guide the other dialogue participant in selecting an appropriate new topic. However, this work did not cover what types of partner information are used in these personalization strategies to drive the decisions made by a dialogue participant.

Mitsuda et al. (2017) studied a similar phenomenon when they developed a taxonomy of categories for the types of perceived information that a human can glean from an utterance. Their work provides evidence that chat-oriented dialogues are dominated by personal information, since 78.5% of the categories of perceived information are directly related to personal information about the speaker or about people, events, and organizations that the speaker has a relationship with. However, this work did not focus on personal information about the speaker, since it also included categories for general world knowledge and did not sufficiently discriminate types of personal information.

Work by Allwood et al. (2011) that studied the differences in topics that are discussed in monocultural and intercultural first-time encounters also uncovered common topics that are directly related to the sharing of personal information, such as age, family, and religion. It is notable that Allwood et al. (2011) observed that there was significant overlap in the topics that were discussed between all three dialogue situations that they studied: dialogues between two Chinese participants, two Swedish participants, and one Chinese and one Swedish participant. This provides strong support that personal information is a vital component of all chat-oriented dialogue, regardless of culture and participant similarity. However, similar to the work by Mitsuda et al. (2017), the topics that were identified in this work are also too broad to be useful for gaining an understanding of personal information shared in dialogue.

One source of detailed types of personal information is Schema.org, an open source resource that contains schemas for structured data, primarily for use by web developers in order to create web pages that are easily indexable by search engines. The schemas on Schema.org are collections of properties that can be used to described specific concepts. For our work, we were particularly interested in the schema definition of the concept 'Person', part of which can be found in Table 1. We used the properties of a Person as defined by Schema.org as the starting point for defining the personal attribute types that are shared in chat-oriented dialogues.

Property	**Definition**
Nationality	Nationality of the person.
Net Worth	The total financial value of the person as calculated by subtracting assets from liabilities.
Owns	Products owned by the person.
Parent	A parent of this person.
Performer In	Event that this person is a performer or participant in.
Related To	The most generic familial relation.

Table 1: Examples of the properties of a 'Person' along with their definitions, as defined by Schema.org

3. Attributes and Entities

For this work, we were interested in identifying the types of personal information that are shared by a speaker in chat-oriented dialogue. Within this context, the term *attribute instance* is used to refer to a single piece of personal information shared by a speaker, such as that they were born in Phoenix or that they are 45 years old. An *attribute instance* can be completely described in terms of its *attribute type* and its *attribute value*, such as (birthplace Phoenix) or (age 45).

In this work, an *attribute instance* is applied to a particular individual of an *entity type* (not necessarily a person). For example, it is possible to say that Phoenix is a large city, where this *attribute instance* would have an *attribute type* of 'size' and an *attribute value* of 'large'. In this case, the *entity type* of the *attribute instance* is 'Place', because the individual being described - Phoenix - is a city.

4. Corpora

In order to look at a range of chat data rather than be limited to a specific setting, we focused on three different datasets of dyadic chat. The three datasets were the Story-swapping Corpus, the Switchboard Corpus (Godfrey and Holliman, 1993), and the SpeedDate Corpus.

4.1. Story-swapping Corpus

In 2015, Gilani et al. (2016) conducted a study on story-swapping that investigated the impact that different virtual storytellers had on a human participant who engaged in a story-swapping dialogue with the virtual storyteller. Participants interacted with different versions of a virtual storyteller in a "get to know you" scenario where they answered predetermined ice-breaker questions. The participants were given the ice-breaker questions beforehand and initiated each exchange with the virtual storyteller by asking one of the questions. In response, the virtual storyteller produced a relevant story that answered the question, and asked the same question back to the participant, who then gave an answer in response.

Because our work aimed to identify attributes shared in natural human dialogues, only the utterances given by the human participant were included from this dataset and the utterances produced from the virtual storytellers were ignored. Although this dataset could be viewed as an inorganic source of human conversation due to the scripted nature of the topics and the artificial storyteller participant, the ice-breaker questions function as guiding topics to the conversation and the information that a human participant chose to share in response to the topics was produced naturally from the human.

4.2. Switchboard Corpus

The Switchboard: Telephone Speech Corpus for Research and Development was collected between 1990 and 1991 by Texas Instruments, with sponsorship from DARPA. It consists of two-sided telephone conversations between human participants, where the participants were connected via a robotic switchboard operator and were unfamiliar with the other person they were speaking to. Upon initiation of a telephone conversation, the operator gave the participants a specific topic to discuss with each other based on their previously indicated topics of interest. Some example topics are capital punishment, pets, cars, and recycling. In general, the course of the ensuing conversation focused on the assigned topic as speakers exchanged their opinions and relevant experiences to the topic at hand.

The Switchboard Corpus was originally released by the Linguistics Data Consortium in 1992-1993, but was released again in 1997 with some errors fixed. In total, it contains 2400 conversations between 543 speakers (Godfrey et al., 1992). Transcripts of the recorded telephone conversations were also produced, which has resulted in much work on annotating the dialogues with different linguistics features, from phonetics to syntax (Calhoun et al., 2010). Of particular importance to this work are the word-level, turn, and utterance boundary transcriptions, where the speakers were labeled as 'A' and 'B'. An extension of the 1997 "Switchboard 1 Release 2" Corpus - called the "Switchboard Dialog Act Corpus" - was used in this work. It contains word-level transcriptions of the dialogues segmented into turn-taking utterances, where each utterance is tagged with a dialog act but the dialog act was stripped from the utterance for our purposes (Stolcke et al., 2000).

4.3. SpeedDate Corpus

In 2005, Jurafsky et al. (2009) used three speed-dating sessions run at an elite private American university to collect casual dialogues. These speed-dating sessions were composed of graduate students from the university who participated in 4-minute "get to know you" sessions with each other on a one-to-one basis. Each session occurred in an open setting, and each participant wore an audio recorder during the session so that the audio could be captured. Transcribers at a professional transcription service used the recordings to create a transcript for each date between two graduate students. In total, there were 991 dates that had usable transcript data.

Although the majority of each dialogue followed the typical open-ended structure of sharing personal information in a get-to-know-you setting - as expected of a first date - the content of many of the dialogues was also directly affected by the speed-dating environment, since participants typically began their conversations with a discussion of the nature of the speed-dating activity. One common occurrence was that participants would often comment on the tasks that were required of them - such as their difficulty in filling out the surveys about each person with whom they have a date. In addition, participants would often bring up noteworthy elements of the speed-dating environment - for example, that there is a large proportion of law students taking part in the speed-dating activity.

5. Procedure for the Identification of Attribute Instances

75 dialogues from each corpus were randomly selected for annotation. Each utterance was appraised for whether there was a derivable attribute instance. If there was an attribute instance, then Schema.org was investigated to find an attribute type that would capture the information in the attribute instance. In the case that no corresponding attribute type could be found, a new type was created. This new type was then considered in each subsequent utterance as a possible attribute type - along with the properties from Schema.org - if an attribute instance was found. Once the attribute type of an utterance was identified, then the value of the attribute instance was selected as a substring from the utterance and the entity type was identified by classifying what the attribute was being assigned to. If there was no attribute instance derivable from the utterance, then it was not annotated.

Frequently, there were multiple attribute instances within a single utterance. All possible attribute instances that could be derived from a single utterance were identified and annotated for that utterance.

Table 2 shows an example of the annotations for a dialogue from the Switchboard Corpus.

6. Attributes in Chat-Oriented Dialogues

6.1. Entity Types

We found a total of 12 different entities that people share information about in these chat-oriented dialogues. The overwhelming majority of attributes were given about a Person (96.4%). Table 3 depicts the 12 entity types that were

DIALOGUE	ATTRIBUTE VALUE	ATTRIBUTE TYPE	ENTITY TYPE
A: I used to jog somewhat.	jog	previous-activities	person
B: I had an exercise bike.	exercise bike	previously-own	person
B: I used to have one.			
B: And I finally got rid of it cause I never used it.			
B: But I do use my treadmill.	treadmill	activities	person
A: Uh-huh.			
A: Well, that's good.			
A: Yeah.			
A: My parents have a treadmill.	parents, treadmill	parent.t, owns	person, person

Table 2: Example of attribute annotations for a portion of a dialogue from the Switchboard Corpus

identified as well as their respective distributions in the dialogues.

Entity Types	Frequency
Person	4503
Place	61
Organization	32
Pet	30
Car	21
Program	7
Job	6
Course	3
Restaurant	2
Activity	2
Sports Team	1
Event	1

Table 3: Entity types found in dialogue

6.2. Attribute Types

We found a total of 166 types of attributes that people share in these chat-oriented dialogues. Most attribute types were slot-based, where a specific filler could be found for the attribute based on the information shared in the utterance - such as a person's name or the university they are currently studying at. However, there were many instances of a binary-valued attribute type, where an utterance would indicate that a particular property was true or false for the given entity under discussion. An example of this occurs in the sentence 'My daughter agrees with me' since this utterance clearly depicts that the speaker has a daughter. This can be captured through a binary attribute type, which was called 'children.daughter.t'. All attribute types that end in .t indicate a binary attribute with the value of true, whereas attribute types that end in .f indicate a binary attribute with the value of false. Both slot-based and binary attribute types were mapped to a single word in the utterance as the attribute value.

These 166 attribute types are grouped into 12 broad categories. The first 9 categories pertain to the attribute types that were used most frequently in relation to the entity type of 'Person' and can be seen in Table 4. The remaining 3 categories include attribute types that were never used in relation to the entity type of 'Person' and can be seen in Table 5. The 12 categories are described in more detail below.

6.2.1. Demographics

The category *Demographics* includes attributes for basic identifying information - such as age and name - as well as traits that distinguish different populations from one another - such as heritage and membership in different organizations. This category also contains attributes that indicate well-being.

6.2.2. Personality

Personality attributes capture the distinctive elements of a person that grant them individuality. It includes a person's likes, dislikes, fears, goals, plans, and physical traits.

6.2.3. Relationships

Relationships contains attributes that provide information about the different relationships a person has with other people - mother, father, children, neighbor, sibling, and so on. It includes whether the person has these relationships and how long they have known each other. A person's history of romantic relationships is also captured by the attributes in this category.

6.2.4. Work

The attributes in the category *Work* focus on the details of a person's current - and past - employment. It includes characteristics like job title, length of employment, location of work, and company.

6.2.5. Education

Similar to *Work*, *Education* encapsulates attributes that define a person's educational history. It captures relevant details such as where a person has attended school, what field they have studied, and the degree they pursued there.

6.2.6. Residence

Residence contains attributes for current and past living arrangements, such as the type of dwelling, the location of residency, and the length of a time a person lived there. It also includes attributes for more detailed aspects of the living arrangements, such as specific qualities of the home.

6.2.7. Possessions

Attributes in the *Possessions* category indicate the material wealth of a person. These attributes capture what objects a person owns and doesn't own, as well as any information on their financial state.

6.2.8. Behavior

Behavior attributes encapsulate the different recurring activities that a person does, as well as the different experiences they have had in their lifetime. There are also at-

Demographics	Personality	Relationships	Work	Education	Residence	Behavior
age	dislikes	affiliatedwith	company.type	currenteducation.field	home.details	activities
alive.f	familiarwith	children.daughter.t	previouswork.company.name	currenteducation.graduation	home.length	activities-no
birthdate	favorites	children.f	previouswork.company.type	currenteducation.length	home.location	activities.length
birthmonth	fears	children.number	previouswork.description	currenteducation.location	home.situation	experiences-no
birthplace	goals	children.son.t	previouswork.length	currenteducation.t	home.type	experiences-yes
birthyear	goals-maybe	children.t	previouswork.location	currenteducation.type	previoushome.details	previousactivities
childhood	goals-no	cousin.t	previouswork.status	currenteducation.year	previoushome.length	travelled.location
deathdate	interestedin	frequencyofvisits	previouswork.t	degree	previoushome.location	travelled.not
deathyear	isa	friend.t	previouswork.title	instructor	previoushome.type	usesservice
gender	isnota	grandchild.f	previouswork.type	previouseducation.completed.f	usesservice-no	
heritage	languages	grandchild.granddaughter.t	vacationtime	previouseducation.completed.t		
memberof	likes	grandparent.grandfather.t	work.company.name	previouseducation.field		
mentalstate	misses-no	grandparent.grandmother.t	work.company.type	previouseducation.graduation		
middlename	misses-yes	grandparent.t	work.f	previouseducation.length		
name	notfamiliarwith	inlaw.brother.t	work.length	previouseducation.location		
nickname	notinterestedin	inlaw.mother.t	work.status	previouseducation.t	**Possessions**	**Distinguishments**
physicalstate	plans	inlaw.t	work.t	previouseducation.type	financialstate	accomplishments
previousmemberof	plans-maybe	knows	work.title	previouseducation.type.f	owns	artistof
	plans-no	neighbor.t	work.type		owns-no	producerof
	previousfavorites	nibling.t			owns.length	
	previousgoals	parent.father.t			previousfinancialstate	
	previousinterests	parent.mother.t			previouslyown	
	previouslikes	parent.t				
	religion	previousromantic.length				
	trait	previousromantic.t				
	traits-no	previousromantic.type				
		romantic.f				
		romantic.length				
		romantic.t				
		romantic.type				
		sibling.brother.number				
		sibling.brother.t				
		sibling.sister.t				
		sibling.t				
		stepparent.father.t				
		uncle.t				

Table 4: Attribute types found in dialogue, attributable to a Person

Pet	Vehicle	Location
breed	model	commonactivity
	numberofdoors	containedinplace
	vehicleengine	crime
	vehiclemodeldate	demographics
	price	largerthan
		proximity
		previousnames
		servescuisine
		similarto
		size
		type
		rules
		clients.type

Table 5: Attribute types found in dialogue, not attributable to a Person

tributes that indicate certain activities or experiences a person does not participate in, or specific locations that a person has never travelled to.

6.2.9. Distinguishments

The attributes in the *Distinguishments* category list a person's accomplishments.

6.2.10. Pet

Pet attributes include only those attributes which were used to describe pets (namely, their breed). Other attributes from the *Personality* category were also associated with pets in the dialogues.

6.2.11. Vehicle

Vehicle attributes focus on the model and price of a Vehicle entity, as well as other features like the number of doors.

6.2.12. Location

Location attributes are used in the descriptions of places, such as cities and businesses. These cover a variety of characteristics, including demographic distributions, proximity to other places, and size.

7. Attribute and Entity Type Evaluation

We performed a small inter-annotator agreement study on 15 dialogues, which was composed of 5 randomly selected dialogues from each corpus. We calculated inter-annotator agreement on both attribute type and entity type, but not for attribute values. Because the possible attribute values for any attribute type are open-ended paraphrases of a part of the utterance, it is impossible to give annotators a finite list of possible attribute values. For this reason, it is difficult to operationalize the similarity between the attribute values chosen by different annotators. In addition, we believe that the inter-annotator agreement on the attribute types provides a good approximation for the agreement that would be found on the attribute values, since these two concepts are directly related.

We calculated the AC1 measure of inter-annotator agreement for both attribute and entity types. As defined by Gwet (2002), AC1 aims to overcome the issue in other inter-annotator agreement calculations - namely, Cohen's Kappa and Scott's π-statistic - that causes them to produce unexpectedly low agreement measures when given data with large differences in the trait frequencies, which

is an accurate description of our data. It accomplishes this through a revised chance-agreement measure, which is explained in detail in Gwet (2002). We calculated AC1 for each attribute type by treating them as a binary labels for each utterance, such that an annotator either indicated the particular attribute type was 'present' or 'not present' in a particular utterance. AC1 is calculated as:

$$AC1 = \frac{p - e(\gamma)}{1 - e(\gamma)} \quad (1)$$

where p is the base agreement measure between the annotators and $e(\gamma)$ is the revised chance-agreement probability. We only calculated AC1 for those attribute and entity types that were annotated by at least one annotator, which was 94 attribute types (out of 166 total) and 2 entity types (out of 12 total). The final AC1 value is calculated as the average of the AC1 values calculated for each attribute type and entity type and is shown in Table 6.

Annotators were a mix of experts and non-experts. Half of the annotation material came from the authors' work on identifying the attribute and entity types. The other half came from novice annotations, since they were conducted by an outsider to the work who was given an annotation manual written by the authors. Each annotator was given a finite list of possible attribute types and entity types that an utterance could be classified with.

Annotation Element	AC1
Attribute Type	99.76%
Entity Type	96.28%

Table 6: Inter-Annotator Agreement Statistics

The inter-annotator agreement is so high because each utterance contains at most a few attributes or entities, and annotators agree on the many attribute and entity types that are not present in each utterance.

Since we are mainly interested in annotator agreement on the presence of the different labels, we corrected for this artifact of the inter-annotator agreement measure by also calculating precision, recall, and F1-score for the attribute and entity type annotations. We calculated these measures for each individual attribute type and entity type. The final measures shown in Table 7 reflect the average of these individual measures.

Annotation Element	Precision	Recall	F1-Score
Attribute Type	36.4%	34.6%	35.5%
Entity Type	59.4%	60.7%	60.0%

Table 7: Precision, Recall, and F1-Score for Annotations

The resulting precision, recall, and F1-score presents a more pessimistic perspective on the inter-annotator agreement for attribute and entity types in dialogue utterances, especially for the attribute types. The fine-grained distinction between the different attribute types seems to be a limiting factor in the reliability of the attribute type labels. Frequently, the annotators agreed that an utterance should be labeled with a specific type category (such as 'currenteducation' or 'previouswork'), but they then chose different sub-categories (such as 'previouswork.type' versus 'previouswork.description'). In addition, it can be argued that there is a degree of semantic overlap between many of the attribute types, such as 'goals' and 'interests', which is also evidenced through the annotations made by each of the annotators. As such, these attribute types would be better captured through a higher-order type that encapsulates both.

With the limitations of the current annotation scheme identified, it is important to note that this work does not purport to have identified the single true taxonomy of personal information present in chat-oriented dialogues. Instead, we aimed to get a sense of the types of information that tend to occur in chat-oriented dialogues and offer reasonably plausible categories of topics that can be found.

	Attribute	Instances		Utterances		
Corpora	Total	Range	Average	Total	Range	Average
Story-swapping	1921	6-61	25.6	4509	21-173	60.1
Switchboard	1495	2-70	19.9	13982	64-444	186.4
SpeedDate	1251	2-46	16.7	8673	66-189	115.6

Table 8: Distribution of the number of attribute instances and utterances between the three corpora

8. Observations and Analysis of the Attribute Distributions

The total number of utterances and attribute instances that occurred over the 75 dialogues of each corpus - as well as their ranges and averages within a single dialogue in each corpus - can be found in Table 8. Although the total number of utterances is much lower in the Story-swapping Corpus than the other two, this can be attributed to the fact that only the half of each dialogue that was spoken by the human participant was considered. In total - when taking into account the other half of the dialogues - the total number of utterances over the 75 dialogues for the Story-swapping Corpus is approximately 9000, which is similar to the total number of utterances of the other two corpora.

Tables 9, 10, and 11 show the twenty most frequently occurring attributes for the Story-swapping Corpus, the Switchboard Corpus, and the SpeedDate Corpus, respectively. Table 12 shows the twenty most frequently occurring attribute types over all of the corpora and their distribution across the corpora. Out of the twenty most-frequent attribute types for each corpus, eight attribute types appeared for all corpora: likes, activities, dislikes, birthplace, home.location, travelled.location, goals, and previouseducation.type.

Furthermore, out of the 166 total different attribute types that were found, 62 of them occurred in all three corpora. Since only 129 different attribute types appeared in the Switchboard Corpus, 110 in the Storyswapping Corpus, and 91 in the SpeedDate Corpus, the fact that 62 attribute types occurred in all three corpora shows that there is a great degree of overlap between the different corpora, since, at minimum, half of all of the attribute types found in each corpus were also found in both of the others. A major source of overlap between the three corpora is their abundance of personality, demographics, and behavior attribute

Attribute Types	Frequency
likes	423
trait	124
experiences-yes	114
experiences-no	82
activities	70
dislikes	69
previouseducation.location	63
favorites	63
birthplace	62
home.location	52
travelled.location	52
previousactivities	46
age	44
goals	40
traits-no	38
activities-no	37
work.type	32
previouseducation.type	31
work.title	30
work.status	24

Table 9: Top twenty attributes in the Story-swapping Corpus

Attribute Types	Frequency
likes	141
activities	96
home.location	88
romantic.type	72
owns	58
trait	56
work.t	52
children.t	49
age	47
dislikes	35
work.status	30
goals	30
travelled.location	26
birthplace	24
home.type	24
parent.mother.t	24
previoushome.location	23
previouseducation.type	21
friend.t	21
childhood	20

Table 10: Top twenty attributes in the Switchboard Corpus

Attribute Types	Frequency
currenteducation.field	157
name	122
likes	103
currenteducation.year	82
birthplace	80
currenteducation.type	76
goals	46
activities	41
previouseducation.location	39
travelled.location	38
previouseducation.type	35
currenteducation.t	29
friend.t	26
previoushome.location	26
plans	24
dislikes	17
home.location	16
currenteducation.length	13
size	12
sibling.sister.t	12

Table 11: Top twenty attributes in the SpeedDate Corpus

Attribute Types	Frequency	Top-Twenty Appearances
likes	667	3
activities	207	3
trait	191	2
currenteducation.field	170	1
birthplace	166	3
home.location	156	3
name	137	1
experiences-yes	135	1
dislikes	121	3
travelled.location	116	3
goals	116	3
previouseducation.location	106	2
age	100	2
experiences-no	92	1
romantic.type	89	1
previouseducation.type	87	3
currenteducation.year	86	1
currenteducation.type	84	1
owns	72	1
favorites	69	1

Table 12: Twenty most frequent attribute types found overall in the three corpora, and the number of corpora in which each appeared in the individual top-twenty lists

types, which can most clearly be seen when comparing their respective most-frequent attribute types, although the specific attribute types in these categories varied between the corpora.

Upon closer examination of the distributions for each corpus, there were telling differences between the three corpora that relate to their variation in domain and dialogue participants. The most apparent discrepancy occurs in the distribution of work-related and education-related attribute types. It is seen in the Story-swapping Corpus and Switchboard Corpus that both had more frequent work-related attribute types than the SpeedDate Corpus, which had more frequent education-related attributes. This makes sense in light of the fact that all of the conversation participants in the SpeedDate Corpus were graduate students, and thus their education - and not their employment - was most salient.

It was also observed that relationship-related attribute types occurred most frequently in the Switchboard Corpus, as compared to the other corpora. The focus of the dialogues in the Switchboard Corpus was less on the person speaking - unlike in the ice-breaker paradigms of the other two corpora - and instead, the conversational participants were discussing any relevant experience in relation to specified topics - such as childcare and the justice system. Often, these speakers drew on the experiences of those closely related to them - either as romantic partners, family members, or friends - which can explain why much information about a person's relationships was observed in these dialogues.

Overall, there was much overlap in the distribution of attribute types between the three different chat-oriented corpora. In general, it was observed that many of the most frequently occurring attribute types between all three corpora often belong to the same personal information categories of personality, behavior, education, and relationships. These distributions of common attribute types observed in this work begin to indicate what personal information is focused on in chat-oriented dialogue, regardless of the paradigm in which the dialogue was conducted, and provides further evidence that the act of sharing personal information is similar across different dialogue paradigms. In addition, the differences that are observed between the attribute types of the three corpora aid in delineating different types of corpora, by indicating dominating topics for the participants involved.

Although the three corpora used in this work differed in the paradigms by which they were collected, it could be argued that all three belong to the same overarching category of polite casual conversation as defined by Eggins (de Silva Joyce and Slade, 2000). It may be the case that the common attribute types observed in this work would not extend to the confirming subcategory of casual conversation, where the participants are well-acquainted and in frequent contact with one another, or to other types of casual conversation. Further exploration of the similarities and differences between the common attribute types of different categories of chat-oriented dialogue remains as an area of future work.

9. Conclusions and Future Work

We have created two useful resources for chat-oriented dialogue - first, an enhanced collection of personal attributes that can be used as topics for dialogue interaction and, second, a corpus of dialogues annotated for attribute types, attribute values, and entity types. These resources were created through examination of three different kinds of chat dialogue corpora. Comparative analysis shows that while there are significant differences in the types and frequencies of attributes in the different corpora (which could be used to distinguish different genres of chat), there are also significant similarities that occur in two or all three.

We plan to use these resources for several purposes in our future work. First, these annotations can be used as training and testing data for classifiers that recognize when people mention these attributes in their utterances. This capability will enable dialogue agents to develop a user model of the other conversational participants and to personalize the dialogues to the topics most interesting to them. In addition, the commonly discussed personal attributes indicated by this work can be incorporated into the backstory development of dialogue agents so that they can engage in more human-like chat.

10. Acknowledgements

The first author was supported by the National Science Foundation under grant CNS-1560426, "REU Site: Research in Interactive Virtual Experiences" (PI: Evan Suma). The effort described here has been partly supported by the U.S. Army. Any opinions, content or information presented does not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

11. Bibliographical References

Allwood, J., Lindström, N. B., and Lu, J. (2011). Intercultural dynamics of fist acquaintance: comparative study of swedish, chinese and swedish-chinese first time encounters. In *International Conference on Universal Access in Human-Computer Interaction*, pages 12–21. Springer.

Calhoun, S., Carletta, J., Brenier, J. M., Mayo, N., Jurafsky, D., Steedman, M., and Beaver, D. (2010). The nxt-format switchboard corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language Resources and Evaluation*, 44:387–419.

de Silva Joyce, H. and Slade, D. (2000). The nature of casual conversation: Implications for teaching. *Teachers' Voices*, 6:viii–xv.

Eggins, S. and Slade, D. (1997). *Analysing Casual Conversation*. Continuum.

Gilani, S. N., Sheetz, K., Lucas, G., and Traum, D. (2016). What kind of stories should a virtual human swap? In *The Sixteenth International Conference on Intelligent Virtual Agents*, pages 128–140, Los Angeles, CA, USA, September. Springer.

Godfrey, J., Holliman, E., and McDaniel, J. (1992). Switchboard: telephone speech corpus for research and development. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520. IEEE.

Gwet, K. (2002). Kappa statistic is not satisfactory for assessing the extent of agreement between raters. *Statistical methods for inter-rater reliability assessment*, 1(6):1–6.

Hirano, T., Higashinaka, R., and Matsuo, Y. (2016). Analyzing post-dialogue comments by speakers – how do humans personalize their utterances in dialogue? –. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 157–165, Los Angeles, September. Association for Computational Linguistics.

Jurafsky, D., Ranganath, R., and McFarland, D. (2009). Extracting social meaning: Identifying interactional style in spoken conversation. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 638–646.

Mitsuda, K., Higashinaka, R., and Matsuo, Y. (2017). What information should a dialogue system understand?: Collection and analysis of perceived information in chat-oriented dialogue. In *International Workshop on Spoken Dialogue Systems Technology (IWSDS)*.

Slade, D. and Gardner, R., (1993). *Teaching Casual Conversation: The Issue of Simplification*. ERIC.

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).

12. Language Resource References

John J. Godfrey and Edward Holliman. (1993). *Switchboard-1 Release 2*. Linguistic Data Consortium, 1.0, ISLRN 988-076-156-109-5.