Purely Corpus-based Automatic Conversation Authoring

Guillaume Dubuisson Duplessis[1], **Vincent Letard**[2], **Anne-Laure Ligozat**[3], **Sophie Rosset**[1]
[1]LIMSI, CNRS, Université Paris-Saclay, F-91405 Orsay
[2]LIMSI, CNRS, Univ. Paris-Sud, Université Paris-Saclay, F-91405 Orsay
[3]LIMSI, CNRS, ENSIIE, Université Paris-Saclay, F-91405 Orsay
{gdubuisson, letard, annlor, rosset}@limsi.fr

Abstract

This paper presents an automatic corpus-based process to author an open-domain conversational strategy usable both in chatterbot systems and as a fallback strategy for out-of-domain human utterances. Our approach is implemented on a corpus of television drama subtitles. This system is used as a chatterbot system to collect a corpus of 41 open-domain textual dialogues with 27 human participants. The general capabilities of the system are studied through objective measures and subjective self-reports in terms of understandability, repetition and coherence of the system responses selected in reaction to human utterances. Subjective evaluations of the collected dialogues are presented with respect to amusement, engagement and enjoyability. The main factors influencing those dimensions in our chatterbot experiment are discussed.

Keywords: Example-based dialogue modelling; Open-domain dialogue system; Human-Machine dialogue corpus; Evaluation

1. Introduction

The main objective of our work is to design a conversational strategy aiming at the maintenance or increase of the human participation in dialogue, usable both in chatterbot systems (such as Eliza (Weizenbaum, 1966)), or as a fallback strategy for out-of-domain human utterances in specific dialogue systems. An important property for such a system is to enable a wide variety of possible responses, while keeping a level of control of what is said (with regard to, e.g., the desired language level, the relationship between the human and the system, the situation of interaction).

Our approach belongs to the category of example-based dialogue modelling which aims at using a database of semantically indexed dialogue examples to manage dialogue (Lee et al., 2009). However, our purpose is a complete automation of this process from the creation of the database to the conversational management process to avoid the need of a costly and time-consuming human intervention. To this end, we present a data-driven process aiming at the automatic authoring of a conversational strategy for a dialogue system interacting with a human participant. The main purpose of the conversational strategy is to select an appropriate response from a corpus given a human utterance, and adapt it by taking into account the history of dialogue (including the last utterance from the human). This process relies both on the exploitation of a large and varied corpus of Human-Human interactions and on natural language processing tools such as a named entity (NE) recogniser.

Our approach is implemented in the context of the Joker project which aims to build a generic intelligent user interface providing a multimodal dialogue system with social communication skills including humour and other social behaviours (Devillers et al., 2015). This system is primarily involved in entertaining interactions occurring in a social environment (e.g., a cafeteria). A conversational strategy could be fruitfully used to provide the system with judicious and entertaining contributions when faced with an unexpected human utterance.

Section 2. draws some links with related work. Section 3. presents an overview of our approach from the automatic creation of the database of dialogue examples to the selection process behind the conversational management process. Section 4. describes an implementation of our approach on a subtitle corpus, exploited for the creation of the database of dialogue examples. Section 5. details the conversational management process involving three main steps: the selection of candidate system-responses, the selection of the most appropriate response and its transformation according to the dialogue context. It shows examples of Human-Machine dialogues obtained in text-based interactions between a human and our system. Section 6. describes an experiment dedicated to the collection of a Human-Machine corpus of dialogues between a human and our system in a chatterbot usage. Section 7. presents the collected corpus, and discusses the capabilities of our system with regard to the self-reports filled by the human participants, as well as the chatterbot usage of the automatically authored open-domain conversational strategy. Lastly, section 8. concludes this paper.

2. Related Work

Several approaches have been previously undertaken to automatically author a conversational strategy based on movie scripts (Banchs and Li, 2012; Nio et al., 2014) and on movie subtitles (Ameixa et al., 2014). (Banchs and Li, 2012) present the IRIS chat-oriented dialogue system. It implements a dual-strategy to select a system-response from a corpus of movie scripts that takes into account the history of dialogue and the current user input. (Nio et al., 2014) describe a chat-oriented dialogue system based on a corpus of drama television dialogues. This system retrieves the shortest system-response from the corpus based on semantic and syntactic similarity with the human-utterance. They outline the benefit of taking into account NEs to generalise the system-response. (Ameixa et al., 2014) design a conversational strategy to deal with out-of-domain human utterances in Human-Agent interaction based on a corpus of movie subtitles. This approach distinguishes the retrieval of candidate system-responses and the selection of the most appropriate response. (Gandhe and Traum, 2007; Gandhe and Traum, 2013) present several surface text-based models

for virtual agents to select an appropriate system response from an in-domain Human-Human dialogue corpus. These models are evaluated in the domain of simulation training involving an army captain and a doctor.
In our work, we intend to go further than these previous approaches on several aspects. First, we aim at the automatic design of an open-domain conversational strategy based on a corpus of Human-Human interactions for both chatterbot systems and fallback strategies. Then, our goal is to design a conversational strategy that makes it possible to maintain some control over the selected response of the system. To this end, our approach explicitly takes into account three main steps in the conversational management process by discerning: (i) the selection of candidate system-responses, (ii) the selection of the most appropriate response, and, last but not least, (iii) the transformation of the system-response by taking into account the local context of the corpus, the human-utterance, and the history of dialogue.

3. Overview of the Approach

Our approach involves three main steps. First, it consists in the selection of a corpus of Human-Human interactions which is going to be exploited on the overall process.
Next, an automatic pre-processing is applied on the selected corpus. It aims at building a database of initiative/response pairs enriched with additional information including, e.g., NEs. Notably, extracted pairs are more than just surface text pairs. Indeed, a pair represents an initiative/response context from which additional information can be used to fruitfully adapt the response to a given initiative. This property is used during the conversational management process. Several operations can be applied to the corpus including lexical normalisation, segmentation of the corpus in order to extract initiative/response pairs, filtering of inappropriate pairs and enrichment of utterances forming pairs via their automatic annotation.
Then, the conversational management process takes advantage of the built database of initiative/response pairs. It consists in selecting and adapting a system-utterance from the database of pairs, given the last human-utterance and the history of dialogue, in order to produce a system-response. This process can be broken down into three main parts:

1. Selection of initiative/response pairs relevant to the human-utterance.
2. Selection of the most adequate response from the previously picked pairs that forms the system-utterance.
3. Transformation of the system-utterance by taking into account the pair, the human-utterance and the history of dialogue to produce the system-response.

4. Creation of the Database from a Subtitle Corpus

4.1. Description of the Corpus

Our corpus is made from subtitles of television dramas[1]. Given the context of the Joker project, we selected a variety of genres including comedy and sci-fi series [2]. Alternative recent approaches use movie scripts instead (Banchs and Li, 2012; Hu et al., 2013). The goal is to make the robot provoking laughter, notably with the incongruity of its answers, while maintaining a coherence with the ongoing dialogue.
The corpus contains 432 subtitle files, for a total of 274,659 utterances (635.8 utterances per file). Statistics about the utterances in the corpus are given in Table 1. Due to the spoken source for subtitles, most of the utterances are rather short. The ratio of unique utterances over the total number is 0.81, meaning that one sentence out of 5 from the total corpus is repeated several times – short usual expressions, mainly.

Files	432
Total utterances	274,659
Utterances per subtitle file	635.8
Ratio of unique utterances	81.1%
Tokens per utterance (mean / std)	6.23 / 3.56
Unique tokens	33,995

Table 1: Statistics about the subtitle corpus

4.2. Limitations

Using a subtitle corpus for the example base of a chatterbot comes with some difficulties. First, interactions in the TV series are not exclusively between two characters, and someone may interrupt the character speaking. Used as is, the risk is to introduce incoherent sequences of utterances when read as a dialogue between two persons. Preventing this problem would require the costly annotation of the speaker in every episode of the corpus. To the best of our knowledge, there is no automatic approach for this task which does not need multimodality. However, we chose not to go further into this problematic for the following reasons: the system is intended to be fully unsupervised, which rules out annotation; also, it is a first version and an advanced combination of development choices would make it complex to evaluate.
Another limitation that follows from this corpus is that scenes are not encoded in the subtitles. This prevents the direct detection of incoherent utterance pairing at the edge of scenes. We tried to use the time gap between utterances to automatically determine scene limits, but it turned out not to be a relevant factor.

4.3. Pre-Processing of the Corpus

The following operations are applied to the corpus in order to automatically generate a usable database.
First, a cleaning process is necessary in order to remove any text that is not part of the actual dialogue such as subber annotations, font shaping or speaker naming. Fancy or misread characters are also corrected or removed at this stage.

[1] Downloaded from tvsubtitles.net

[2] The exhaustive list is: “Real Humans”, “Awkward”, “Game Of Thrones”, “Malcolm In The Middle”, “The Americans”, “The Big Bang Theory” and “The Good Wife”.

Then, the NEs are identified with the help of the Stanford NER (Finkel et al., 2005). After being tagged by the parser, the NEs are memorised and replaced by their type so that they stay neutral for the further similarity calculations of the lookup phase.
Finally, the utterances are lemmatised using the English version of the lemma dictionary from (Courtois, 1990).
The resulting base contains 274,227 pairs of which 25% contain at least one tagged NE[3]. 11% of the latter contain NEs both in the initiative and the response, and a quarter of those cases (that is 1626 pairs in total) have a followed entity. An entity is followed when it is present both in the initiative and the response.

5. Conversational Management

This section details the three steps of the response production process (cf. figure 1).

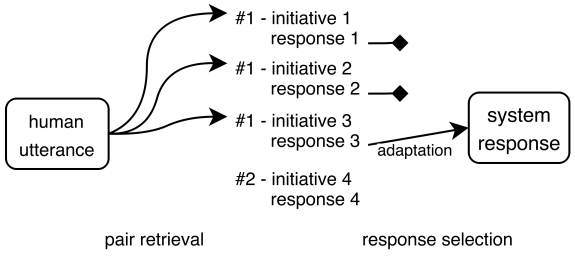

Figure 1: Conversational Management Flow

5.1. Pair Retrieval

Currently, the last human-utterance is used to retrieve appropriate pairs from the base. After being applied the same NER tagging and lemmatisation as described in previous section, the utterance is used to rank all initiative utterances from the base by their similarity. Our similarity measure is based on tf-idf: the similarity between the human-utterance and any initiative utterance from the base is the mean of the tf-idf for each token.

$$\sigma(u_{user}, u_{base}) = \frac{1}{|(u_{user})|} \sum_{w \in u_{user}} tfidf(w, u_{base}, base)$$

In this particular use of tf-idf, utterances are considered as the documents. $base$ is the set of all the documents. This choice of similarity measure permits to retrieve pairs in which the initiative is lexically close to the human-utterance. In future work, we intend to more generally take into account the dialogue history in this process.

5.2. Response Selection

An eligible response is the response in the base to an initiative that is similar to the human-utterance. Those responses can be multiple for many reasons:

- several initiatives can have the same best similarity with the human-utterance
- the exact same initiative can appear more than once in the base
- a threshold can be set in order to choose more than one pair

The default strategy can be to randomly select among those, however we observed that often, several of these possible responses were close if not identical. The random strategy will select more likely (more often) the responses that are more represented, but the final selection is still out of precise control. Another strategy that can be derived from this observation is to choose the most representative of the responses. This can be done by computing the mean word vector of all the possible responses and using it as the ideal mean of the responses. We then rank the responses by the cosine similarity of their word vectors with the mean word vector. After this last step, still having responses that are have the same rank can mean either that they are very similar, or that they are so different that the mean word vector is equidistant to each of them. In both cases, the random selection among the best ranked is harmless.
However, the choice of a selection strategy between pure random or a more elaborate one like our proposal is not easy. Indeed, the elaborate strategy is (nearly) deterministic and will provide the same output each time whereas the random one is not, by definition. If a similar utterance is told twice by the human, the order in which possible responses should be given is certainly not random and strongly depends on the dialogue history. In that perspective, the mean word vector strategy at least guarantees a certain coherence in track of the system's responses.
Finally, we can remark that the case of repetition of utterances by the speaker is rather unlikely. It is mainly an issue for short utterances which are theoretically less numerous.

5.3. Response Transformation

The final step is to adapt the single selected response to the human-utterance. In this work, the adaptation consists in replacing or deleting the original NEs. For this purpose, we distinguish several situations regarding the selected pair (i, r) from the base:

1. every NE found in r is also present in i:
   (a) every NE from i has an equivalent of the same type in the human-utterance
   (b) some NEs from i cannot be linked to any in the human-utterance
2. some NEs from r are not present in i

When NEs from r can be linked to i and to the human-utterance, the default strategy is to replace it accordingly. For every other unlinked entity, it can be deleted if it is not inside a sentence, or left identical. We plan to implement a generalisation process to replace unlinked entities by more neutral expressions such as "someone" or "that place". Also, a POS tagging will allow to accurately modify sentences to allow entity deletion without meaning loss.

[3]We observe a difference of 432 with the total utterances in Table 1. This is because the last utterance of each file is not paired with the first of the following one.

5.4. Dialogue Examples

Our system targets the selection of an appropriate response to an open-domain human utterance. We present some success and failure cases of our system collected in text-based interactions with a human in a chatterbot usage (described in section 6.).

Our system is able to select coherent and sometimes relevant responses allowing fluid interactions spanning several exchanges, e.g. ("H" is the human, "S" is the system):

H: hey
S: hey
H: what's up
S: nothing
H: seriously?
S: yeah
H: what do you do all day?
S: mostly read comics

Some surprising and sometimes amusing interactions have emerged between the human and the system, e.g.:

H: are you a machine?
S: i am who i was yesterday
H: who were you yesterday?
S: anybody
H: anybody is not an answer
S: it was

or such as:

H: would you like some wine?
S: no
H: beer may be?
S: it's hard to get drunk on beer

However, these results should be mitigated (see section 7. for a discussion of the capabilities of the system). Indeed, we have observed some failure cases. It includes the selection of incomprehensible responses, e.g. incomplete ones:

H: i am tired
S: well that's

or incomprehensible ones:

H: ok. i don't know what can i say to you.
S: hejsan ooh

Sometimes, the system selects responses that are clearly invalid given the previous utterance:

H: what are you?
S: hello
H: could you describe yourself?
S: thanks

The system also selects responses that are pragmatically valid but nonetheless inappropriate, e.g.:

H: 2 times 2 equals ?
S: 1

6. Preliminary Study: Chatterbot Usage

6.1. Description of the Experiment

We conducted an experiment dedicated to the collection of a Human-Machine corpus of dialogues between a human and our system in a chatterbot usage. Dialogues consist in free open-domain conversations with the system through a text-based web interface. Dialogues are carried out in English. Our chatterbot system is a reactive-only dialogue system. Hence, dialogues follow a common structure which is a sequence of human participant-system utterance pairs. Dialogues are initiated by the human participant.

6.2. Description of the Questionnaires

Participants were encouraged to immediately assess each conversation they had with the system via two questionnaires.

The first questionnaire is dedicated to the evaluation of individual system utterances produced in response to human ones. For each system utterance, the participant was asked to answer by one of the three possibilities "yes", "no", "N/A" to the following questions: (i) "Is the utterance understandable?", (ii) "Is the utterance polite?", (iii) "Is the utterance coherent?", and (iv) "Is the utterance relevant?". Understandability of an utterance is defined as whether or not the response of the system is understandable taken alone. For instance, the utterance "Hello !" can be said to be understandable whereas the utterance "I..." cannot. Coherence of an utterance assesses whether a system response is acceptable given the directly previous human utterance. In other words, an exchange of two utterances is said to be coherent if the presence of the second utterance is easily explained by taking into account the first one. For example, the answer "Hi!" can be said to be coherent in response to the utterance "Hello!", whereas "The sky is blue." cannot. Relevance of an utterance is defined as a coherent utterance that also provides interesting information.

The second questionnaire is a global self-report regarding the conversation participants just had on a Likert scale of 5 points. It consists of several items aiming at (i) evaluating the interaction in terms of amusement, engagement, enjoyability, (ii) evaluating the system in terms of its capabilities (global politeness, understandability, coherence, relevance and repetition) as well as its attitude (e.g., friendliness), and (iii) evaluating the attitude and mental states of the participants (such as bored, embarrassed, upset, friendly, enthusiastic, embarrassed, surprised, feeling a desire to talk to the system). A representative example of a Likert item of this second questionnaire is: "The interaction with the system has been: 1-'not amusing at all', 2-'somewhat not amusing', 3-'neither amusing nor not amusing', 4-'somewhat amusing', 5-'very amusing'". Additionally, participants had the possibility to let anonymous free form comments about the experiment.

6.3. Participants

27 volunteers participated in this experiment (37% female, 63% male). They are mainly non-native English speakers. Ages of the participants are comprised between 20 and 59. 81% of the participants have ages ranging from 20 to 39; 19% have ages ranging from 40 to 59.

7. Results and Analyses

7.1. Description of the Collected Corpus

Participants were allowed to interact several times with the system. A session consists of a dialogue with the system along with the two evaluation questionnaires. Table 2 presents some figures about the collected corpus. We collected 41 sessions carried out by 27 participants. 48% of the participants only interacted once with the system. On

Dialogues	41
Turns	1384
Unique utterances	917 (system: 358, human: 573)
Tokens	5493 (unique: 989)

Table 2: Figures about the collected corpus

Type	average (std/min/max)
Turns	33.76 (19.32 / 10 / 82)
Tokens	133.98 (73.2 / 29 / 331)
Tokens/human	85.02 (45.5 / 20 / 208)
Tokens/system	48.95 (29.1 / 8 / 130)
Tokens/human utt.	5.04 (2.9 / 1 / 23)
Tokens/system utt.	2.90 (2.9 / 0 / 20)
Duration	7min19s (9m41s / 1m38s / 56m49s)
Response time	
...human	25.7s (108.8 / 0.002 / 2429.5)
...system	1.8s (0.19 / 0.002 / 13.7)

Table 3: Figures about the collected dialogues. utt. = utterance

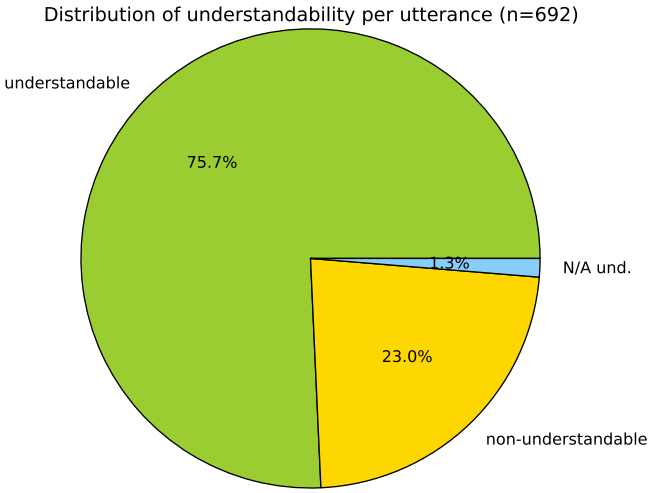

Figure 2: System utterance-level evaluation of understandability

average, participants carried out 1.8 sessions (median=1.0, std=1.1). The maximum number of sessions by a unique participant is 5.

Table 3 presents some figures characterising the collected dialogues. On average, a dialogue with the system lasts around 7 minutes and contains approx. 34 turns. Unsurprisingly, the human participant takes approx. 10 times much time than the system to produce a response, right after a system utterance (on average 25.7s for the human, 1.8s for the system).

The corpus is available at the URL: `https://ucar.limsi.fr/`.

7.2. Variety of System Responses

We investigated some objective indicators to evaluate system responses. First, it is worth noticing than the system contributions are shorter than the human ones in terms of token size. From table 3 and given the fact that every dialogue contains the same number of system turns and human participant turns, we can see than the system produces less tokens than the human on the whole dialogue (on average, 48.95 VS 85.02). The same can be said at the utterance level (on average, 2.90 VS 5.04): system utterances are shorter than human utterances.

Next, we looked into the repetition of the system inside a same dialogue. To that purpose, we computed a unique utterance[4] ratio which corresponds to the proportion of unique system utterances in a given dialogue. Alternatively, it is given by the formula: #unique system utterance /#system utterances. For the entire corpus, the ratio is on average 0.92 (median=0.93; std=0.08; min=0.73; max=1.0). In other words, 92% of the utterances of the system are unique in a dialogue on average. Then, we examined the variety of system responses in the entire corpus, i.e. across dialogues. To that aim, we computed the ratio given by the number of unique system utterances in the corpus against the number of unique human utterances in the corpus. It turns out that the ratio for this corpus is 0.62. This is an objective evidence supporting the fact that system utterances are less varied than human ones. These results about the repetition of the system are completed by the subjective evaluation of human participants in section 7.3.

[4]Here, utterance equality is the same as string equality.

7.3. Subjective Study

7.3.1. Capabilities of the System

Participants were asked to assess the global capabilities of the system in producing a response via 5-point Likert items. When asked about the politeness of the system, participants reported that the answers of the system were "3: neither polite nor impolite" (mode=median=3), where 1 is "very impolite" and 5 is "very polite". Concerning the understandability of the responses of the system, participants reported that the answers of the system were "4: somewhat understandable" (mode=4; median=3), where 1 is "very incomprehensible" and 5 is "very understandable". Regarding the coherence of the responses of the system, participants reported that the answers of the system were "2: somewhat incoherent" (mode=median=2), where 1 is "very incoherent" and 5 is "very coherent". Similar results have been observed for the relevance of the responses of the system. These results are corroborated by the utterance-level evaluation. It shows that 75.7% of the responses of the system have been rated as understandable, while 23% have been considered as incomprehensible (see figure 2). On the other hand, only 41.2% of the responses have been rated as coherent, and 55.5% as incoherent. This comes down to 32.2% for relevant responses.

As of the repetition of the system (see figure 3), participants reported that the system has been "2: repeating itself" (mode=2; median=3), where 1 is "repeating itself a lot" and 5 is "not repeating itself at all". However, these results should be mitigated. Indeed, results presented in

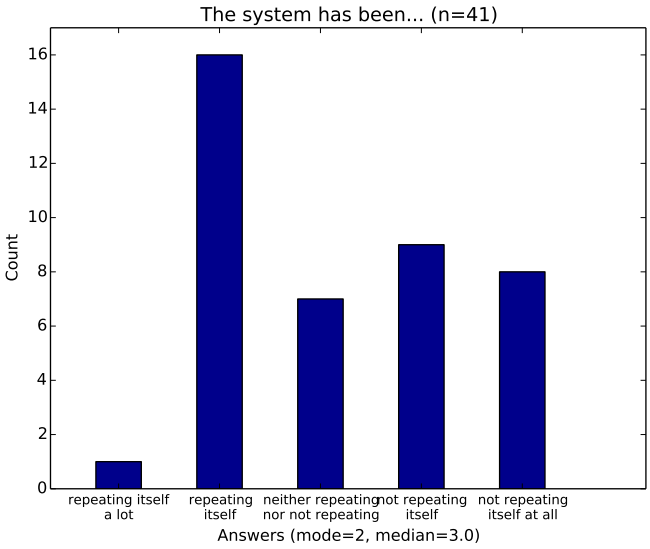

Figure 3: Perceived system repetition by human participants

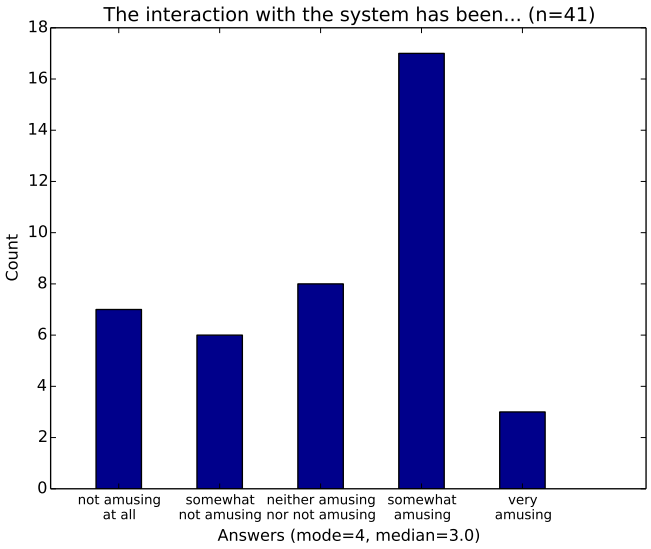

Figure 4: Amusement reported by human participants

section 7.2. indicate that the system seems to be repeating itself mainly across dialogues rather than inside the same dialogue. As a matter of fact, it turns out that participants that only interacted once with the system have the median answer "4: not repeating itself".

Lastly, the participants mainly reported the system as being "3: neither friendly nor unfriendly" and "3: neither rude nor nice".

7.3.2. Self-report from Human Participants

Participants mainly rejected negative mental states by answering that they were (i) "3: neither bored nor not bored" (mode = median = 3), (ii) "3: neither embarrassed nor not embarrassed" (mode = median = 3), and (iii) "1: not upset at all" (mode = 1; median = 2). Similarly, participants mainly selected positive attitudes in answering that they were "4: somewhat friendly" (mode=median=4) and "3: neither rude nor nice" (mode = 3; median = "2: somewhat nice"). However, participants answered being "3: neither enthusiastic nor unenthusiastic" (mode = median = 3).

Concerning the chatterbot system, participants were "4: somewhat surprised" (mode = 4; median = 3) by the responses of the system. Besides, participants mainly agreed to feeling a desire to talk to the system (mode = 4; median = "3: neutral"), where 1 corresponds to "strongly disagree" and 5 to "strongly agree". A clear distinction on the desire to talk to the system appears between participants that only interacted once with the system, and those who interacted several times. By considering results on the first session carried out by each participants, we realise that participants that interacted once strongly disagreed on feeling a desire to talk to the system (mode = 1; median = 2) while participants that interacted more than once agreed (mode = 4; median = 3).

7.3.3. Subjective Evaluation of the Interactions with the System

Participants were asked to report their opinions on three notions related to the interaction they just had with the system: amusement, engagement and enjoyability.

Figure 4 presents the results concerning the amusement of the interaction with the system. Participants mainly reported that the interactions were "4: somewhat amusing" (mode = 4; median = 3). We looked into features of dialogues (length in turns or tokens, duration, response time, vocabulary overlap, and ratios about understandability/coherence/relevance/politeness) that could explain the amusement reported by human participants. The most significative feature that we found is the coherence ratio, i.e. the proportion of system utterances in a dialogue that have been rated as coherent by the human participant. Alternatively, it is given for a specific dialogue by the formula: $\#\text{coherent system utterances}/\#\text{system utterances}$. The coherence of the system responses seems to play a role in the amusement of the human participant. Indeed, it exists a significative linear correlation between the coherence ratio and the reported amusement ($r = 0.52$, t-test p-value$= 0.0004$), computed with Pearson's product-moment correlation. In addition, we examined potential links between mental states reported by the participants and the amusement. Surprising the human participant with system responses seems to favour the reported amusement of the participant. This is supported by the existence of a significative linear correlation between the reported amusement and the reported "surprised" mental state ($r = 0.58$, t-test p-value$= 6.35 \times 10^{-5}$).

Figure 5 presents the results concerning the engagement of the interaction reported by the human participants. They mainly reported that the interactions were "1: not engaging at all" (mode = 1; median = "2: somewhat not engaging"). In the same way as amusement, engagement is connected to the coherence of the system utterances. This is confirmed by the existence of a significative linear correlation between

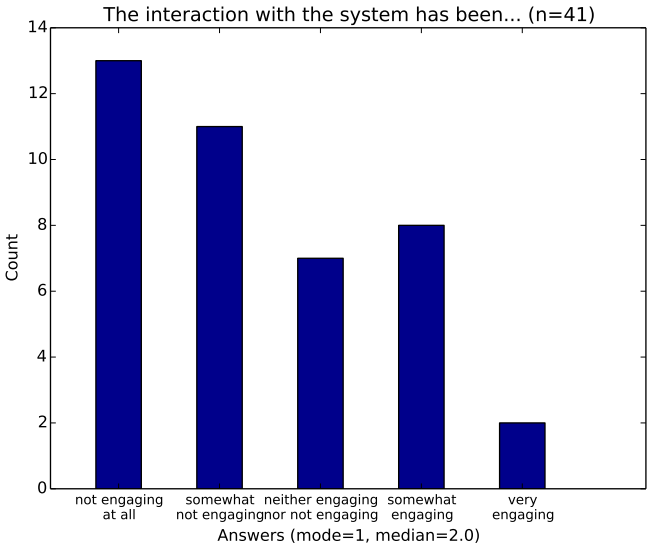

Figure 5: Interaction engagement reported by human participants

the coherence ratio and the reported engagement ($r = 0.60$, t-test p-value$= 2.87 \times 10^{-5}$). Besides, we looked into the links between self-reports and engagement. It turns out that the main link we found is between the engagement and the "desire to talk" felt by the human. It is supported by a significative linear correlation between those two ($r = 0.65$, t-test p-value$= 4.84 \times 10^{-6}$).

Lastly, participants reported that the interactions were "2: somewhat unenjoyable" (mode = median = 2). This may seem surprising compared to the results about amusement. We did not find significative link between features of dialogues and the reported enjoyability. In particular, we did not observe significative links between enjoyability and the capabilities of the system in terms of politeness, understandability, coherence and relevance (unlike what was found for amusement and engagement). In addition, reported enjoyability does not seem to play a role between participants that only interacted once with the system, and those that interacted several times (contrary to the "talk desire", cf. section 7.3.2.). One explanation may be that enjoyability is more connected to the feeling of the participant toward the chatterbot nature of the experiment rather than on the experiment itself. This seems supported by the fact that the only significative link we found between enjoyability and mental states is with the reported enthusiasm of the participant. We found a significative linear correlation between those two aspects ($r = 0.58$, t-test p-value$= 6.75 \times 10^{-5}$).

7.4. Discussion: Lessons from a Data Collection with a Chatterbot

This study has permitted to evaluate the general capabilities of our system. It has pointed out some limitations of the automatically authored open-domain conversational strategy in a chatterbot usage. First, we have observed that the system mainly generates understandable responses to a human utterance, with room for improvement. Next, our system has unsurprisingly shown clear limitations on the coherence and relevance aspects that should be improved in the near future. In particular, our results show a direct link between the reported system response coherence, and the amusement and engagement of the human in the interaction. Last, the system objectively repeats itself across dialogues, and it is perceived by the human participants. However, the repetition inside a unique dialogue seems acceptable.

This study has brought some useful insights regarding the chatterbot usage of the authored open-domain conversational strategy. First, it has confirmed that the engagement of interaction reported by the human participants in chatterbot usage is dependent on other factors than the system response selection process. In particular, participants have been complaining in the free form comments about the lack of initiative of the system (which is only reactive). This is corroborated by the fact that we found a significant correlation between the talk desire felt by human participants and the reported interaction engagement. Next, participants have been complaining about the shortness of system responses, explaining that there was "not enough matter to bounce on to continue the interaction". This idea seems supported by correlations between the surprise felt by human participants and the reported amusement. This may indicate that the ability of the system to produce surprising and intriguing responses may increase amusement and probably interaction engagement. Lastly, our study indicates that enjoyment of a chatterbot interaction by a human participant may reside in external factors independent from the system. We have not found significative links between reported enjoyment of the interaction and dialogue features or system capabilities (such as dialogue length and coherence). However, we have found a significant correlation with the reported mental state "enthusiasm".

8. Conclusion and Future Work

In this paper, we have presented an unsupervised corpus-based process to author an open-domain conversation strategy usable both in chatterbot systems and as a fallback strategy. This process has been implemented on a corpus of television dramas subtitles.

This system has been used as a chatterbot to collect a corpus of 41 open-domain textual dialogues with 27 human participants. It is available at the following URL: `https://ucar.limsi.fr/`. We have carried out a study that has made it possible to discuss the general capabilities of the system in terms of understandability, repetition and coherence of the selected system response in reaction to a human utterance. In addition, we have collected the subjective evaluation of human participants in terms of amusement, engagement and enjoyability. We have discussed the main factors influencing those dimensions in our chatterbot experiment, namely: the coherence of the responses of the system with regard to the human utterance, their sizes and their ability to surprise the participant, and the lack of initiative of the system.

This study has brought some useful insights to improve our unsupervised approach to open-domain conversation strategy authoring. In future work, we intend to study the impact of the subtitle corpus on the understandability of the

system response. In particular, we envision a more robust automatic filtering process to limit the number of incomprehensible system utterances. Next, we aim at improving coherence of the response by taking into account history of dialogue and not just the last human utterance. Then, we plan to favour response variability in the response selection algorithm to avoid repetition.

Acknowledgements

This work was partly funded by the JOKER project and supported by ERA-Net CHIST-ERA, and the "Agence Nationale pour la Recherche" (ANR, France).

9. Bibliographical References

Ameixa, D., Coheur, L., Fialho, P., and Quaresma, P. (2014). Luke, I am your father: dealing with out-of-domain requests by using movies subtitles. In *Intelligent Virtual Agents*, pages 13–21. Springer.

Banchs, R. E. and Li, H. (2012). IRIS: a chat-oriented dialogue system based on the vector space model. In *Proceedings of the ACL 2012 System Demonstrations*, pages 37–42. Association for Computational Linguistics.

Courtois, B. (1990). Un système de dictionnaires électroniques pour les mots simples du français in dictionnaires électroniques du français. *Langue française*, 87:11–22.

Devillers, L., Rosset, S., Dubuisson Duplessis, G., Sehili, M., Béchade, L., Delaborde, A., Gossart, C., Letard, V., Yang, F., Yemez, Y., Türker, B., Sezgin, M., El Haddad, K., Dupont, S., Luzzati, D., Estève, Y., Gilmartin, E., and Campbell, N. (2015). Multimodal data collection of human-robot humorous interactions in the joker project. In *6th International Conference on Affective Computing and Intelligent Interaction (ACII)*.

Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370. Association for Computational Linguistics.

Gandhe, S. and Traum, D. R. (2007). Creating spoken dialogue characters from corpora without annotations. In *INTERSPEECH*, pages 2201–2204.

Gandhe, S. and Traum, D. R. (2013). Surface text based dialogue models for virtual humans. In *Proceedings of the SIGDIAL*.

Hu, Z., Rahimtoroghi, E., Munishkina, L., Swanson, R., and Walker, M. A. (2013). Unsupervised induction of contingent event pairs from film scenes. In *EMNLP*, pages 369–379.

Lee, C., Jung, S., Kim, S., and Lee, G. G. (2009). Example-based dialog modeling for practical multi-domain dialog system. *Speech Communication*, 51(5):466–484.

Nio, L., Sakti, S., Neubig, G., Toda, T., Adriani, M., and Nakamura, S. (2014). Developing non-goal dialog system based on examples of drama television. In *Natural Interaction with Robots, Knowbots and Smartphones*, pages 355–361. Springer.

Weizenbaum, J. (1966). ELIZA - a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, January.