Comparing System-response Retrieval Models for Open-domain and Casual Conversational Agent

Franck Charras[1], Guillaume Dubuisson Duplessis[1], Vincent Letard[2], Anne-Laure Ligozat[3], and 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
{charras, gdubuisson, letard, annlor, rosset}@limsi.fr

Abstract. This paper studies corpus-based process to select a system-response usable both in chatterbot or as a fallback strategy. It presents, evaluates and compares two selection methods that retrieve and adapt a system-response from the OpenSubtitles2016 corpus given a human-utterance. A corpus of 800 annotated pairs is constituted. Evaluation consists in objective metrics and subjective annotation based on the validity schema proposed in the RE-WOCHAT shared task. Our study indicates that the task of assessing the validity of a system-response given a human-utterance is subjective to an important extent, and is thus a difficult task. Comparisons show that the selection method based on word embedding performs objectively better than the one based on TF-IDF in terms of response variety and response length.

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

1 Introduction

This work aims at improving the design of dialogue strategies for conversational agents (e.g., the famous Eliza [19]) that are involved in entertaining interactions occurring in a social environment, e.g., a home, a cafeteria, a museum. This paper targets dyadic social open-domain conversations between a human and a system. We aim at designing a system that can be used either as a chatterbot system, or as a fallback strategy for out-of-domain human utterances in a wider dialogue system. In particular, we consider selection-based models as a promising approach to deal with some human utterances in social interaction (see, e.g., [9, 10][4][2],[16],[7]). The main purpose of this kind of model is to select an appropriate response given a human utterance from a database of indexed dialogue examples, and adapt it by taking into account the history of dialogue. To this end, these data-driven models rely on large and varied corpora of Human-Human interactions as well as on natural language processing tools to avoid the need of a costly and time-consuming human intervention.

This paper focuses on the objective and subjective evaluations of two different approaches of response selection that exploit the English part of the OpenSubtitles2016 corpus [15]. More specifically, we constitute a corpus of human-utterance/system-reponses pairs annotated with subjective evaluation following the "validity" annotation schema proposed in the (RE-)WOCHAT shared task [6]. Next, we offer a study of inter-annotator agreement of the subjective evaluation of selected system-response. Then, we analyse the main differences between two different approaches for response selection (one based on TF-IDF and the other on word embeddings).

The remainder of this paper is organized as follows: after the discussion of related work (Section 2), we introduce the two response selection approaches our work is based on (Section 3). Next, we describe our experimentation protocol (Section 4), along with the corpus constitution and the annotation process. The results are presented and discussed in Section 5. Finally, Section 6 concludes this paper.

2 Related Work

Conversational systems are recently gaining a renewed attention in the research community, 50 years after the famous ELIZA system [19]. This is shown by the recent effort to generate and collect data from the (RE-)WOCHAT workshops [6] involving various systems such as IRIS [4], Joker [7], Politician [12], pyEliza [5], Sarah [1] and TickTock [20] systems. In particular, some approaches aim to select an appropriate response from a corpus given a human utterance, and adapt it by taking into account the history of dialogue. A number of selection-based approaches aim at automatically authoring a conversational strategy based on large dialogue corpora such as movie scripts [4, 16], movie subtitles [2, 7] and in-domain Human-Human dialogue corpus [9, 10].

Selection-based approaches rely on the exploitation of a large and varied corpus of Human-Human or Human-Machine interactions. While there are more and more available data for building data-driven dialogue systems (see, e.g., the extensive study by Serban et al. [18]), there are few data available to author social open-domain conversational system such as chatterbot. Recent efforts aim to generate, collect, and evaluate chat-oriented dialogue data, e.g., the UCAR corpus [7][4] and the (RE-)WOCHAT workshops [6][5]. A number of current approaches exploit large corpora of transcribed/scripted interactions such as (but not limited to) the Movie DiC Corpus [3] or the recently released OpenSubtitles2016 corpus [15].

Several approaches have been undertaken to evaluate chatterbot and selection-based systems. Some approaches focus on the level of engagement after each turn. For instance, a 5-Likert scale from "strongly disengaged" to "strongly engaged" is explored by the TickTock system [20]. Other approaches are interested in the

[4] Corpus available at: https://ucar.limsi.fr/

[5] Corpus available at: http://workshop.colips.org/re-wochat/data/

system utterance in terms of “appropriateness”, “breakdown”, “coherence” or “validity”. Gandhe and Traum [10] propose to evaluate selection-based systems in terms of appropriateness via a static context evaluation. It consists in providing the selection-based models with the same set of contexts as input (i.e., a sequence of utterances forming the dialogue history) in order to select responses. Then, human judges are asked to evaluate the selected responses given the input context in terms of appropriateness (on a 5-Likert scale from “very inappropriate” to “very appropriate”). They report an inter-annotator agreement showing that judging appropriateness is a difficult task for human judges. Higashinaka et al. [11] propose to evaluate chatterbot responses to human utterance on a 3-level scale of “breakdown” (not a breakdown/possible breakdown/breakdown). A breakdown qualifies a system utterance after which it is difficult to continue the conversation. They report an inter-annotator agreement showing that it is hard for human judges to agree on which system responses constitute a breakdown or not. Dubuisson Duplessis et al. [7] investigate the evaluation of system utterance by their human interlocutor right after the interaction by considering three dimensions: understandability, coherence and relevance. In the context of the (RE–)WOCHAT workshops, an evaluation of system utterances on a 3-level scale of validity is proposed [6] (invalid, acceptable, valid). Report shows encouraging results regarding inter-annotator agreement.

3 Selection-based Approaches

In this paper, we consider approaches based on a response selection mechanism. Our approaches belong to the category of example-based dialogue modeling [14]. The main idea of this approach is to exploit a database of semantically indexed dialogue examples to manage dialogue. A main feature of our work is the complete automation of the conversation strategy authoring process from the creation of the database of dialogue examples based on a corpus of dialogues to the conversational management process. The main purpose of the dialogue management process is to select an appropriate response from a database of dialogue examples given the human utterance. To this end, our approaches discern three main steps: (i) the selection of candidate system responses from the database of examples, (ii) the selection of the most appropriate response, and (iii) the transformation of the selected response by taking into account the human utterance.

In this paper, we consider one selection approach based on TF-IDF and another approach based on word embedding. These approaches are then used to select a response in a subset of the OpenSubtitles2016 corpus [15] (described in section 4). Dialogue examples are initiative/response pairs from the corpus. Given a human-utterance both approaches retrieve the most similar utterances in the dialogue example database and select the most appropriate corresponding response. These approaches differ in the way they select the candidate dialogue examples and the subsequent response. However, they use the same transformation of the selected response. It consists in a substitution of named entities appearing in the human utterance in the selected response (a detailed account can be found in [7], section 5.3). For instance, if the input utterance is “I love Los Angeles .” (and “Los Angeles” has been detected as a location), it is substituted in the selected response “<location> is a nice place !” to produce the final response “Los Angeles is a nice place !”. It should be noted that this heuristic allows incorrect replacement (e.g., an input such as “Hello Bob !” can generate a response such as “Hi Bob !” which is only appropriate if both interlocutors are named “Bob”).

3.1 TF-IDF Approach

Our first approach exploits a TF-IDF similarity measure to retrieve candidate initiative/response pairs. It allows to retrieve pairs in which the initiative is lexically close to the human-utterance. A detailed account of this approach can be found in [7]. The similarity between the human-utterance and each initiative utterance from the base is the mean of the tf-idf for each token (utterances are considered as the documents, *base* is the set of all the documents):

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

Then, a response is selected from the pool of candidate pairs. We intend to choose the most representative of the responses. To this end, we compute the mean word vector of all the possible responses and use it as the ideal mean of the responses [7]. Responses from the pool of candidate pairs are then ranked by the cosine similarity against the ideal mean of responses in order to determine the most appropriate response. Finally, the selected response is transformed using the NE heuristic substitution previously described.

3.2 Word Embedding Approach

The second approach relies on word and utterance embeddings, using the *doc2vec* model [13]. Word and utterance embeddings are jointly learnt as the coefficients of a shallow neural network trained to predict a word, given its context and the utterance it belongs to. We focused especially in harvesting the utterance embeddings, as their cosine similarity can translate lexical and semantic similarity. We used the implementation provided by Gensim [17], with the length of the context window set to 2. After training the model, we infer the embeddings of the human utterances, and use it to retrieve the closest initiative in the example database, with a nearest neighbour search. The selected answer is again adapted with NE substitution.

4 Experimentation

This experimentation aims at evaluating objectively and subjectively a set of open-domain human-utterance/system-response pairs in written English where the response part is generated by one of the previously presented selection methods. The first part of the pair is drawn from human utterances available in the RE-WOCHAT corpus of chatbot dialogues [6]. This experimention includes three main steps: (i) the selection of a subset of 400 human iniatives from the RE-WOCHAT corpus, (ii) the creation of a pair dataset by generating system responses with the two selection methods, and (iii) the annotation of the system response of each pair.

4.1 Selection of the Human Utterances

In this experiment, we only considered the human utterance from the RE-WOCHAT corpus collected with English systems (namely, IRIS [4], Joker [7], pyEliza [5], Sarah [1] and TickTock [20] systems).

We sampled the utterances with a probability that grow proportionally with their number of neighbours in the vector space described in 3.2. We defined neighbours as vectors within a sphere of radius 0.5. As a result, the groups of similar utterances (for instance all questions that began with “Do you like”) were more likely to be sampled together while the isolated, rarest utterances were more likely to be excluded.

All in all, we selected a corpus of 400 human utterances. Table 1 provides some figures about this corpus. 97% of the automatically picked up human utterances are unique[6]. A human-utterance contains approximately 6 tokens with a minimum of 1 token and a maximum of 20 tokens.

4.2 Creation of the Pair Dataset

The English version of the OpenSubtitles2016 corpus [15] was used as the corpus to select response to human-utterance. This corpus is made from subtitles of television dramas. It provides a large amount of transcribed interactions that can be useful for dialogue modelling. It consists of pre-processed subtitles formatted as sequences of tokenized sentences with timing information and meta-data about the subtitle (e.g., identifiant of the TV episode). We completed this preprocessing by applying a named entity (NE) recognition for each utterance. This was done with the Stanford NER [8]. NEs are memorised for each utterance and replaced by their type. Table 1 presents some figures about the selection corpus. It shows that this corpus contains a wide variety of utterances. It is worth noting that it includes nearly 14 millions unique utterances.

The creation of the dataset of human-utterance/system-response pairs consists in generating a response for each one of the 400 human utterances and for each selection method. In all, we generated a corpus of 800 pairs (2 × 400). It should be noted that pairs are not strictly speaking initiative/response pairs. For instance, the human-utterance may be a response to a previous utterance such as “yeah” or “all right !”. Table 1 presents some figures about the dataset of generated responses. A comparison between the TF-IDF-based method and

[6] In this paper, utterance equality is the same as string equality.

	Selection corpus	Human utterances	Response Selection	
			TF-IDF	Doc2Vec
Utterances	23,651,642	400	400	400
Unique utterances	13,913,129 (58.8%)	388 (97%)	265 (66.3%)	348 (87%)
Tokens	173,075,274	2429	2357	2961
Unique tokens	432,811	460	291	774
Tokens per utterance				
. . . average/median	7.32/6	6.07/6	5.89/5	7.40/6
. . . std	5.34	3.10	3.06	5.04
. . . min/max	0/1140	1/20	0/20	1/28

Table 1. Figures about the selection corpus (subset of OpenSubtitles2016 [15]), about the dataset of human utterances and about the datasets of generated responses. In bold, objective features discerning the two selection methods.

the Doc2Vec-based method shows that they vary on two main points. First, the Doc2Vec-based method generates responses that are more varied than the TF-IDF-based method. Indeed, this is shown by the ratio of unique utterances generated by the method (87% for the first one against 66.3% for the second one). Besides, this can also be observed from the number of unique tokens used by the two methods (774 for the first method against 291 for the second one). This increase in variety can be explained by the fact that our TF-IDF-based method tends to select the most average response (cf. section 3.1). Another explanation may be found in the fact that the TF-IDF-based method is confined to the lexical coherence while the Doc2Vec-based method tends to go toward "semantic" coherence. Then, the Doc2Vec-based method generates longer responses than the TF-IDF-based method. The responses selected by the Doc2Vec-based method have the same median length than the initiaves in terms of tokens per utterance. However, it should be noted that the length of these responses seem to vary more importantly, as shown by the largest standard deviation (5.04). It should be noted that our TF-IDF-based method selects responses that are globally smaller in number of tokens per utterance than the human utterances from our corpus.

4.3 Annotation Process

The annotation process consists in the subjective evaluation of system responses in each pair, manually carried out by human annotators. To this purpose, we adopted the annotation schema suggested in the RE-WOCHAT shared task [6]. Annotators were asked to select one (and only one) tag from the following list: "valid", "acceptable" and "invalid". A response is "valid" if it is semantically and pragmatically valid given the human-utterance. For example, valid responses to the utterance "Do you have a girlfriend?" include: "No, not yet.", "Yes, of course.", "You are too curious!". A response that is not semantically valid but can be acceptable considering the human-utterance is "acceptable". For example, acceptable responses to the utterance "did you finish your homework already?"

include: “all of it?”, “why are you asking that?”, “let us better talk about football”. A response that is definitely invalid considering the human-utterance is annotated with “invalid”. For example, invalid responses to the utterance “are you lazy?” include: “one, two.”, “I know they will.”, “your friends are out there.”.

4 participants were involved in the annotation effort. We constituted 2 groups of 2 annotators. Annotators were explained the schema and were provided with the official guidelines[7]. We constituted two complementary subsets of 400 out of 800 pairs from our dataset. Pairs are assigned randomly to one (and only one) of the subset. Subsets contain pairs generated by the two selection methods. One group of annotators were assigned with the first subset, while the other were assigned with the second subset.

5 Results

5.1 Inter-Annotator Agreement

We first investigated the Inter-Annotator Agreement (IAA) on the annotation task among the two pairs of annotators. Table 2 reports the Cohen’s Kappa globally, per group and per selection method. We found, at most, a fair agreement among the experiments. The Cohen’s kappa calculated over all annotations is 0.369. The approach by embeddings yields a slightly higher agreement ($\kappa = 0.382$) than the tf-idf approach ($\kappa = 0.355$).

Next, we took a closer look at the disagreement cases to better understand why annotators disagree. Table 3 presents the computation of IAA for several subsets of our annotated corpus. It turns out that the annotators almost never disagree by opposing *valid* and *invalid* labels (15% of the disagreement cases). Indeed, 85% of the disagreements occurs with an *acceptable* label. Beside, when considering only the pairs of annotations that do not contain the label *acceptable*, we report a kappa equals to 0.732, against merely 0.293 and 0.301 when respectively excluding the annotations with the labels *valid* and *invalid.*

Table 4 presents some examples of pairs from our experiment in which system responses have been assessed as “valid”. These examples are related to human-utterances that are yes/no question, open question, assertion or social conventions. On the other hand, Table 5 reports some examples of pairs in which system-response have been annotated as “invalid”.

Finally, we explored the impact of the length of the human-utterance on the IAA. We found that the IAA is linked to the length of the human-utterance. In particular, Figure 1 reveals that Cohen’s Kappa plummets for very short human-utterances. It shows that IAA ranges from 0.17 for short human utterances (1 or 2 tokens) to 0.48 for longer human utterances (≥ 10 tokens). IAA is at least fair for human utterances exceeding 3 tokens.

[7] Available at: http://workshop.colips.org/re-wochat/documents/Annotation_Guidelines.pdf

	Group 1+2	Group 1	Group 2
All annotations	0.369	0.401	0.337
doc2vec approach	0.382	0.434	0.327
tf-idf approach	0.355	0.361	0.347

Table 2. Cohen's Kappa calculated globally and over several subsets of the annotations. It reflects, at most, a fair inter-annotator agreement.

	Samples remaining	Group 1+2	Group 1	Group 2
Label *acceptable* removed	457	0.734	0.736	0.734
Label *valid* removed	572	0.293	0.343	0.246
Label *invalid* removed	266	0.301	0.347	0.265

Table 3. Cohen's Kappa calculated over several subsets of the annotations, when excluding pairs that were annotated with a certain label. It shows that, while the annotators agree on what are both ends of the spectrum, the intermediate class *acceptable* is blurry.

H: are you upset ?
S: well , i 'm not .

H: i am happy at the moment .
S: i 'm not happy at Addison .

H: what is your middle name ?
S: Ezekiel .

H: i have to go , bye
S: bye .

Table 4. Pair examples in which system response has been rated as "valid" by annotators ("H" is the human, "S" is the system)

H: do you like travelling ?
S: Alfie , this is Tom .

H: very interesting indeed
S: hello .

H: what is your middle name ?
S: no .

H: bye SYSTEM_NAME
S: bye , SYSTEM_NAME , see you when you 'll be back .

Table 5. Pair examples in which system response has been rated as "invalid" by annotators ("H" is the human, "S" is the system)

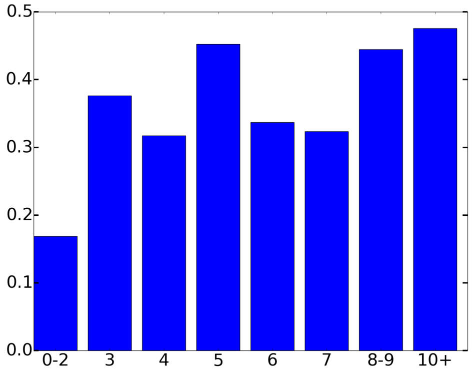

Fig. 1. Cohen's kappa for annotations grouped by length of the human-utterance. It plummets for very short sentences.

5.2 Comparison of the Selection Methods

In addition to exploring the annotation scheme in terms of “validity”, this study also aims at comparing subjectively and objectively the two selection methods (TF-IDF-based and Doc2Vec-based).

First, we investigated the distribution of “validity” annotations globally, per selection method and in function of the agreement of the annotators. Figure 2 reports the observed distributions. Globally, about half of the answers were rated as “invalid” across all dialogues and approaches. We observe a slightly higher proportion of answers rated as “valid” when using the word embeddings. Furthermore, when considering only the dialogues that the annotators agreed on, we find a significantly higher proportion of “invalid” ratings: about two “invalid” rating for one “valid” or “acceptable”.

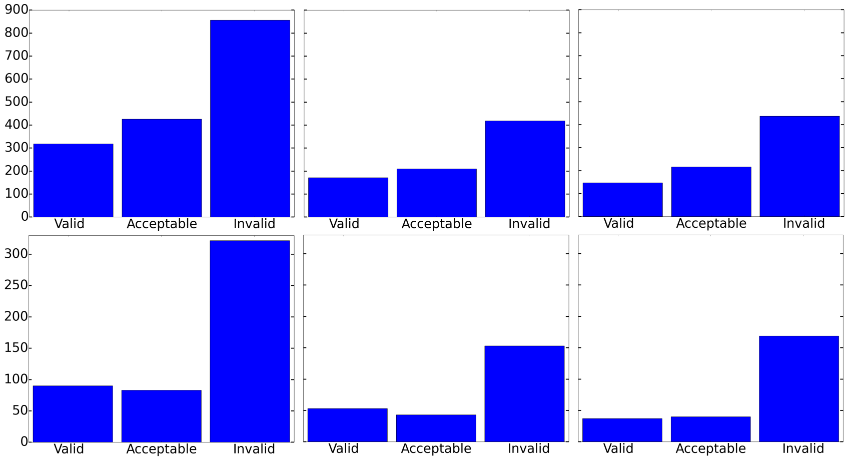

Fig. 2. First line: Distribution of all annotations (left), for doc2vec utterances (middle) and tf-idf utterances (right). Second line: Same distributions, on annotations that annotators agreed on.

Then, we investigated objective features that make it possible to distinguish the two methods. Table 1 summarises the objective features that we found and discussed previously in section 4.2. The vocabulary found in the answers obtained from embeddings contains more than twice more words than with the tf-idf selection. Moreover, it produces answers almost half longer on average. We also found that it uses more diverse answers, when the tf-idf tends to repeat the same utterances for different human-utterances.

5.3 Discussion

We generated and annotated with the (RE-)WOCHAT annotation schema two set of responses to open-domain human utterances via two selection-based system (one based on TF-IDF similarity, one based on Doc2Vec). We have explored the "validity" part of the evaluation annotation schema proposed for the REWOCHAT shared task to evaluate the responses of the system. We reached a fair agreement with Cohen's Kappa globally ranging from 0.337 to 0.401 between two groups of annotators. We took a closer look at the annotation by (i) studying agreement in function of the labels ("invalid", "acceptable", "valid"), and by (ii) computing inter-annotator agreement in function of the size of the human utterances. It turns out that the global Cohen's Kappa did not exceed 0.5 contrary to what was found by [6] ($\kappa = 0.567$). However, it should be noted that the evaluation methodologies differ, in particular on the fact that history was limited to one human utterance in our case whereas annotators had access to the entire dialogue history in [6]. Our study shows that it is very hard to reach an agreement on system-response for very short human-utterance (< 3 tokens). Results also indicate that annotators cannot distinguish clearly "acceptable" from "valid" or "invalid". However, they agree on what defines the two ends of the scale (namely "invalid" and "valid"). This study thus confirms that it is hard for human judges to assess the validity of a system response considering a human-utterance, even with a schema limited to 3 tags. Similar results have been observed with evaluation schema taking other perspectives (namely, "appropriateness" [10] or "breakdown" [11]).

In addition, we have compared two methods to select a response from a large subset of the OpenSubtitles2016 corpus given an open-domain human-utterance. We found two objective criteria that distinguish the two methods, namely the response variety (ratio of unique utterance selected and size of the vocabulary) and the response size in number of tokens. Responses selected by the Doc2Vec-based method are more varied and longer than responses selected by the TF-IDF method, two criteria that have been spotted as important to foster human participation in chatterbot usage [7]. However, we did not found a clear difference on the subjective part. Approximately half of the responses selected by both methods are either valid or acceptable, whereas the other half is invalid.

6 Conclusion and Future Work

In this paper, we have presented two types of selection method that retrieve and adapt a system-response selected from the huge English version of the OpenSubtitles2016 corpus given an open-domain human-utterance . We have constituted a manually annotated corpus of 800 human-utterance/system-response pairs to objectively and subjectively compare these two methods. We have investigated the inter-annotator agreement of a 3-item validity schema used to evaluate system-response proposed by the (RE-)WOCHAT shared task. Main results include the fact that the doc2vec-based selection method performs objectively better than the TF-IDF method in terms of response variety and response length. However, we did not observe a clear difference between the two methods on the subjective evaluation. Besides, our study indicates that the task of assessing the validity of a system-response considering a human-utterance is subjective to an important extent, and is thus a difficult task.

Future work includes the improvements in the pre-processing phase of our approach of large resources of interactions such as OpenSubtitles2016 corpus to enrich the sentences (e.g., with polarity) or to improve the dialogue structure (e.g., structuring the corpus with turn and scenes). We also consider the study of other corpora and other language as an interesting perspective. We are planning to take into account dialogue history in the selection mechanism and studying the impact of the size of this history on the objective and subjective quality of the selected responses.

Acknowledgements

Authors warmly thank Jean-Baptiste Kerbœuf and Korantin Leveque for their contribution on the annotation task. 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).

References

1. Abu Shawar, B.: Sarah chatbot. In: Proceedings of the Workshop on Collecting and Generating Resources for Chatbots and Conversational Agents-Development and Evaluation Workshop Programme (RE-WOCHAT), Shared Task Report, LREC (2016)
2. Ameixa, D., Coheur, L., Fialho, P., Quaresma, P.: Luke, I am your father: dealing with out-of-domain requests by using movies subtitles. In: Intelligent Virtual Agents. pp. 13–21. Springer (2014)
3. Banchs, R.E.: Movie-dic: a movie dialogue corpus for research and development. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2. pp. 203–207. Association for Computational Linguistics (2012)
4. Banchs, R.E., Li, H.: IRIS: a chat-oriented dialogue system based on the vector space model. In: Proceedings of the ACL 2012 System Demonstrations. pp. 37–42. Association for Computational Linguistics (2012)
5. D’Haro, L.: Py-eliza: A python-based implementation of the famous computer therapist. In: Proceedings of the Workshop on Collecting and Generating Resources for Chatbots and Conversational Agents-Development and Evaluation Workshop Programme (RE-WOCHAT), LREC (2016)
6. D’Haro, L., Abu Shawar, B., Yu, Z.: Shared task on data collection and annotation, re-wochat 2016–shared task description report. In: Proceedings of the Workshop on Collecting and Generating Resources for Chatbots and Conversational Agents-Development and Evaluation Workshop Programme (RE-WOCHAT), LREC (2016)
7. Dubuisson Duplessis, G., Letard, V., Ligozat, A.L., Rosset, S.: Purely corpus-based automatic conversation authoring. In: Chair), N.C.C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Tenth International Conference on

Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA), Paris, France (may 2016)

8. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by Gibbs sampling. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. pp. 363–370. Association for Computational Linguistics (2005)
9. Gandhe, S., Traum, D.R.: Creating spoken dialogue characters from corpora without annotations. In: INTERSPEECH. pp. 2201–2204 (2007)
10. Gandhe, S., Traum, D.R.: Surface text based dialogue models for virtual humans. In: Proceedings of the SIGDIAL (2013)
11. Higashinaka, R., Funakoshi, K., Araki, M., Tsukahara, H., Kobayashi, Y., Mizukami, M.: Towards taxonomy of errors in chat-oriented dialogue systems. In: 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue. p. 87 (2015)
12. Kuboň, D., Hladká, B.: Politician. In: Proceedings of the Workshop on Collecting and Generating Resources for Chatbots and Conversational Agents-Development and Evaluation Workshop Programme (RE-WOCHAT), Shared Task Report, LREC (2016)
13. Le, Q.V., Mikolov, T.: Distributed representations of sentences and documents. In: ICML. vol. 14, pp. 1188–1196 (2014)
14. Lee, C., Jung, S., Kim, S., Lee, G.G.: Example-based dialog modeling for practical multi-domain dialog system. Speech Communication 51(5), 466–484 (2009)
15. Lison, P., Tiedemann, J.: OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In: 10th edition of the Language Resources and Evaluation Conference (LREC). Portorož, Slovenia (May 2016)
16. Nio, L., Sakti, S., Neubig, G., Toda, T., Adriani, M., Nakamura, S.: Developing non-goal dialog system based on examples of drama television. In: Natural Interaction with Robots, Knowbots and Smartphones, pp. 355–361. Springer (2014)
17. Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. pp. 45–50. ELRA, Valletta, Malta (May 2010)
18. Serban, I.V., Lowe, R., Charlin, L., Pineau, J.: A survey of available corpora for building data-driven dialogue systems. CoRR abs/1512.05742 (2015)
19. Weizenbaum, J.: ELIZA - a computer program for the study of natural language communication between man and machine. Communications of the ACM 9(1), 36–45 (Jan 1966)
20. Yu, Z., Papangelis, A., Rudnicky, A.: Ticktock: A non-goal-oriented multimodal dialog system with engagement awareness. In: 2015 AAAI Spring Symposium Series (2015)