IIAI Open Conference Publication Series
IIAI Letters on Informatics and Interdisciplinary Research
Vol.005, LIIR221
DOI: https://doi.org/10.52731/liir.v005.221

Epistemic Stance and Contextualization on MLM and NSP: How Japanese Chatbots Recognize the Long-Distance Cohesion between Utterances

Kaoru Amino [*]

Abstract

Error analysis in non-task-oriented dialogue systems has been discussed from many perspectives, mainly in the fields of Artificial Intelligence and Informatics. However, the current trend of error analysis focuses only on local elements and fails to incorporate the whole discourse, as shown in the masked language model and next language prediction. From a linguistic perspective, there are several reasons for errors and unnatural flow in conversations with a Chatbot. These can be stated as: 1) narrowly defined fragments of discourse and the concept of cohesion, 2) a lack of social intelligence in Chatbots due to the limited variety of corpora, and 3) the algorism uncertainty based on the limited variety of data.

This paper analyses the range of references seen in a Chatbot conversation, via qualitative and quantitative methods, and observes how errors are related to the coverage of references, why and how it occurs, and how this kind of error is related to current architectures.

The hypotheses are examined using two processes: 1) comparing the length of references between Chatbot and human interactions, 2) the frequency of errors in Chatbot conversations based on recognition of turn issues (such as insufficient recognition of references, recognition limited to two turns, and the fixed feedback move in a three-turn exchange structure) based on data from "Airfriend" and real, human-produced conversational data.

Keywords: cohesion of discourse, error analysis, socially positioned utterance

1 Background of Error Analysis

1.1 Error analysis in previous works

Chatbots that can have conversations with people online and are used for specific purposes, including services, help with examinations are called "Supervised AI,". In contrast, Chatbots not used for any of these purposes are called "unsupervised AI", and do not have a specific prototype about conversational flow. Unsupervised AI Chatbots do not generally follow any specific instructions and are often used for entertainment or acting as

[*] ShanghaiTech University, Shanghai, China

surrogate companions. Thus, unsupervised AI Chatbots are frequently compared to human conversation and judged by their ability to converse as naturally as real person.

Error analysis is usually examined by data annotators. Because they are not linguistics or discourse analysis specialists, they tend to label data without using specific frameworks or discoursal flow terminology. In one study of error analysis, Higashinaka et al. (2016) roughly classified discoursal errors into five categories: 1) repetition of the same content, 2) contradicting content, 3) sudden utterance, 4) neglecting interlocutor's utterance, 5) does not answer question. The categories were developed based on comment given by data annotators, who participated in their investigation of concordance rate. However, their study did not clarify the exact reason and origin of these errors. Nevertheless, these variously annotated errors may have an integrated cause from linguistic perspective, in terms of turn-taking. For example, the fifth category (Does not answer question) can be interpreted as a feedback issue of the 3rd turn of exchange structure, which is already established before any responses are given. Categories 2 (Contradicting content), 3 (Sudden utterance), and 4 (Neglecting interlocutor's utterance) may be errors due to recognition problems limited to two turns, which end up neglecting the discoursal flow beyond two turns.

1.2 The concept of turns

When discussing discoursal references in conversation, there are two concepts that define it from a linguistic perspective. Adjacency pairs consist of two utterance lengths, which are adjacent to each other [11], with different speakers producing each utterance. Another element of the adjacent pair, a chronological sequence of utterances that contains discriminative relations, where the corresponding range of the second pair is restricted by the meaning of first pair. Based on this definition, there are adjacency pairs, such as "judgement–agreement" and "explanation–understanding," in addition to the typical adjacency pair of "greeting–greeting." [12]. These is also three turn exchange structure which is related to two turns: initiation, response, and feedback (IRF) [3].

In an analysis of Chatbot error, it is shown that how conversational error and how interesting a Chatbot is depends on the versatility of the feedback move, which is applicable to every situation, irrelevant to the content of the user's response [1]. Although this versatile style aims to cover the wide range of responses as much as possible, it lacks preciseness and accuracy in its responses to each individual user, because the architecture is designed to give specific fixed pattern feedback.

In addition to the adjacency pair and IRF exchange structure concepts, another frame to consider regards the amount of cohesion between references. Sometimes the same topical flow of conversation appears after two turns. It is said that flashback refers to the technique of inserting a scene from the past into the present, as when describing a photo of the "good old days," or the protagonist's former characteristics [9]. While this kind of distance between referenced sentences should be considered a key element of overall discoursal flow, Chatbots based on adjacency between two turns are unlikely to grasp this element in authentic discourse. Moreover, the current architecture used in Chatbot systems may actually cause these errors during turn-recognition.

1.3 Masked language model and next language prediction

The bidirectional encoder representations from transformers system were constructed by Google in 2018 and became famous for achieving the highest prediction accuracy rate in various tasks. The system is powered by an architecture called "Transformer," which understands the conversational context by learning from the bidirection. Specifically, Transformer conducts a dual prediction process that uses both the masked language model and next sentence prediction [8]. In the masked language model, the masked word in the previous sentence is predicted using possible words deduced from the sequence of surrounding words in the sentence. In this system, the accuracy of word prediction is based on frequency of collocations that is used to deduce the context [7]. Thus, in next sentence prediction, the probability of cohesion between two sentences is identified by the collocation of the sentences, by using the frequency of collocation algorithm. However, linguistically, this architecture has some potential problems regarding how cohesion between sentences is captured. The concept of cohesion in linguistics is not based purely on the prediction of the next word or sentence, nor is it based on frequency of collocation.

2 Research Question

2.1 Research question and hypothesis

"Yui spot" can be used to roughly classify discoursal errors, such as "repetition of the same content" or "contradicting content," from a non-linguistic perspective [6]. Moreover, the definition of cohesion tends to produce local fragments within two turns; therefore, it fails to incorporate conversational cohesion and naturalness in the discourse and deduce the overall context. This approach therefore limits the view toward the utterances, and also limits perspectives to analysis on error or unnaturalness.

In considering overall discoursal flow, it is hypothesized that referents that are far away from each other are not recognized by current natural language processing approaches. In this study, errors due to the narrow recognition of contextual references and the limitations of cognitive cohesion based on the masked language model and next sentence prediction are examined using a case-study. How errors are related to the coverage of reference, why and how it occurs, and how this kind of error is related to the current architecture is examined. The findings are used to suggest various improvements to the current architecture.

In more detail, this investigation is to clarify the next research hypotheses, a certain number of the errors found in Chatbot can be caused by the lack of recognition of reference beyond 2 or 3 turns and limited its scope only to adjacency pair or the fix conversational flow of IRF.

To verify this hypothesis, this study examines qualitatively and quantitatively how errors are related to the coverage of reference, why and how it occurs, which are caused by the narrow recognition on reference and to the limitation of the cognitive cohesion based on masked language model or next sentence prediction, via next two research processes;

By comparing distance between references in utterances from the Chatbot and a user, whether the Chatbot only refers the limited scope of turns to calculate the distance of references can be elucidated.

Examination of how many errors is based on mistakenly recognized turn-issues, such as insufficient recognition of reference beyond three turns, the recognitive scope limited to two turns, or fixed feedback on a three-turn exchange structure.

2.2 Data

The data used to observe the Chatbot discourse contained 500 turns extracted from both Chatbot conversations and conversations between people. The turns were drawn from the Chatbot "Airfriend" and its conversational corpus with users "*Dansei no Kotoba, Shokuba-hen* (male words, working place version)." Airfriend is a Chatbot developed by a Japanese company and has a very natural conversational style in terms of its use of honorifics and casual style, known as "Sentence Final Particle." This conversation-based corpus incorporates sources such as the Manga corpus. Although the data is defined as "male words, working place conversation" it actually includes equal amounts of both female and male participants. Therefore, the utterance portions are represented equally, and it also includes casual conversation that occurs during lunch and break times to provide a natural overview of the workplace. Based on these two data types, two kinds of examinations related to comparisons of referential distance and frequency of each error type were conducted.

3 Analysis

3.1 Referential distance between Chatbot and human beings

As stated in Section 2.1-2.2, the 500 turns were extracted from each data of both Chatbot and human conversations to compare the referential distances. Since two-turn (adjacency pair) or three-turn utterances (IRF) are the scope of next analysis (in Section 3.2) for each error type, this examination calculates the conversational distance beyond three turns. The long-distance references, which refers to those from more than three turns, were extracted and categorized up to six turns or more. The frequency of each referential distance was then compared to each other. Consequently, the statistical result seen in Figure 1 shows

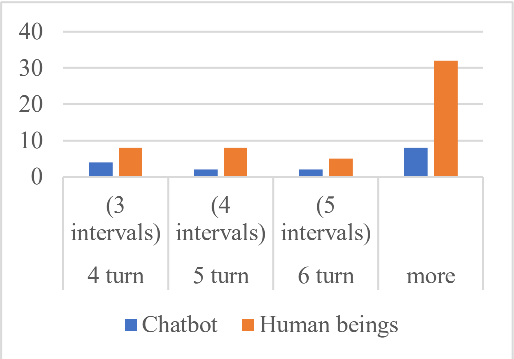


Figure 1: Occurrence of each referential distance

that conversations in the human corpus show longer discoursal references compared to that of the Chatbot. This means that the referential distance enabled by Chatbot remains at the two- or three-turn level. This finding supports the hypothesis, in which Masked Language Model and Next Sentence Prediction can only track the proximate discourse, as shown in Figure 1.

3.2 Error classification

As stated in Section 2.1, the error type was classified into five typologies. The types and their corresponding frequencies are shown below.

(a) Utterances, which do not recognize the previous context

(b) Interpretation of words limited by adjacency pair

(c) Fixed feedback in three-turn structure (IRF)

(d) Meaningless backchannels

(e) Lack of social common sense/knowledge

Some example conversations categorized into (a) and (b) are given below.

Example 1: non-recognizable context (a).

01 A: As for working, I'm a bit busy, but I had something in my private life.
02 U: Oh, did you separate from your girlfriend?
03 A: No, I didn't.
(19:06)

→04 A: However, while we were hanging out with my girlfriend, we loved each other
05 lot.
06 U: Your girlfriend?
07 A: Yes, I feel glad that you understand it.
(19:09)

*A・・・Airfriend, U・・・User

In Example 1, although the Chatbot states in line 02 that it is still hanging out with its girlfriend, this statement is contradicted in line 04 -05, where Chatbot says "while we were hanging out," which implies that they are not together anymore. Thus, the utterance referred to previously tends not to be recognized in by Chatbot.

Example 2: Interpretation of words limited by adjacency pair (b).

01 A: As for the other issue, I want to be friend with a Japanese person.

02 U: You are Japanese, aren't you?

03 A: No, but I want to be friends with them.

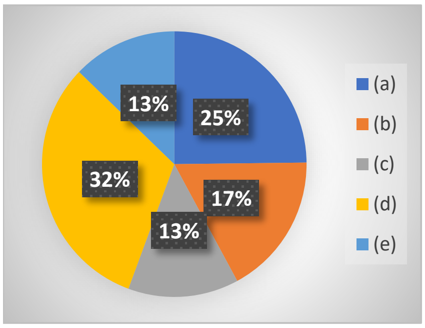
04 U: <u>Then what nationality (*Nani-jin* 何人) do you have?</u>

→05 A: <u>Ummm, now I have around four people</u> (<u>*Yo-nin* 四人</u>).

In Example 2, the Chatbot interprets "nationality (*nani-jin* 何人) " as "how many (*nan-nin* 何人) in line 04. In the reply in line 05, it responds with a number of people without considering the previously stated context. This kind of error in turn recognition is due to the limited number of turns (in this case, two turns), so this type of error is seen frequently in the data. Table 1 and Graph 1 show the number of categorized errors from, (a) to (e).

Table 1: Frequency of each error type

Error type	(a)	(b)	(c)	(d)	(e)
Occurrence	33	23	18	42	17



Graph1: Ratio of occurrence on error type

Taken together, these results show that utterances, specifically errors associated with a lack of referential distance, with limited scope of turn-recognition and fixed IRF (a+b+c) make up more than half of the error types. However, the meaningless versatile backchannels also have a high frequency in the data. Furthermore, lack of social common sense could also be related to the lack of referential distance, because the architecture fails to realize the cohesion in meaning without it. Thus, this error type (a+e) occupies approximately 40% of total errors.

4. Summary and Discussion

4.1 Summary

This study shows that errors due to narrow recognition of references and the limitations of cognitive cohesion based on the masked language model and next sentence prediction can introduce errors into a Chatbot's architecture. This issue is mainly due to the lack of limited scope of references within a short (e.g., two-turn) conversation.

Compared to utterance made by people in real conversations, interactions with Chatbots tend to be limited to two- to three-turn sequences, which generally do not allow a natural progress of topics and discourse. In addition, most errors found in Chatbot discourse are due to a lack of reference to previously stated contexts. This includes separation of discourse flow from numerous two- to three-turn segments, which is partially related to an architectural lack of sufficient data, in terms of social common sense.

4.2 Discussion

The above findings can be used to provide some suggestions on how to improve topic-tracking and recognition in cohesion, by rethinking the concept of cohesion or references. First, it is strongly suggested that the concept of cohesion does not simply rely on the repetition of a selected word and arranging of phrases, which is the basis of current algorism-generated architectures such as MLM or NSP. Additionally, the lack of referential ability in long-distance referencing may not solely be due to architectural inefficiency, it may also be related to more fundamental issues with the concept of the Chatbot itself; specifically, the lack of social intelligence, which positions each utterance as to make a certain meaning in a specific context with a specific conversational participant to interpret it, and the concept of "context" combines the following four elements [4]:

The situation of "setting," where speech-act occurs.

(1) The linguistic expression itself related to the setting of context, which is represented by "contextualization cues [5]."

(2) The Context as the background such as common sense, which surrounds the setting of speech act.

Considering the above elements, installing the function of speech act into architecture could be solved by inputting both data on a certain characteristic and background knowledge on Chatbot and human users. In addition, the context and background as common sense (3) could be solved through utilizing crowd resources to expand the coverage and diversity of data, as well as examining the personal data stored inside cloud. However, Chatbots could be designed to track some "contextualization cues (2)," which are embedded inside linguistic expression itself.

Regarding the contextualization of utterances (2), the idea of epistemic stance [2] as a person's knowledge or belief, including sources of knowledge and degrees of commitment to the truth and certainty of propositions [10] should be taken into account. This means that each utterance first makes its meaning through interpretation of personal belief and degree of commitment. To consider this concept, a Chatbot lacks a core existence, individuality, and ability to interpret utterances from its own contextualized situation and knowledge.

The problem of lack of social intelligence and epistemic stance to interpret meaning in Chatbot conversations will not be easily solved. Further consideration is required to increase accuracy in cohesion between utterances, in order to generate adequate responses in various circumstances, beyond local cohesion.

References

[1] K. Amino, J. Dong, Z. Zhou, Z., "The Analysis of Conversational Error on the Unsuper vised Chatbot: Applying Linguistic Frameworks of IRF and its 6 Acts," Bulletin of the Society of East Asian Language and Culture Studies, vol. 2, Academic Bulletin, Kyusyu University, 2023.

[2] W. Chafe and J. Nichols, "Evidentiality: The Linguistic Coding of Epistemology." Norwood, NJ Ablex, 1986.

[3] M. Coulthard and D. Brazil, "Exchange Structure." In Coulthard, M. (eds.), Advances in Spoken Discourse Analysis. Routledge, 1992, pp. 50-78.

[4] A. Duranti and C. Goodwin, "Rethinking Context: Language as an Interactive Phenomenon." In E. Ochs (Ed.), Culture and language development: Language acquisition and language socialization in a Samoan village. Cambridge: Cambridge University Press, 1992.

[5] J. Gumperz, J. Discourse Strategies. Cambridge University Press, 1982.

[6] R. Higashinaka, K. Funakoshi, Y. Kobayashi, M. Inaba, "The Dialogue Breakdown Detection Challenge: Task Description, Datasets, and Evaluation Metrics." In The Tenth International Conference on Language Resources and Evaluation (LREC' 16), 2016, pp. 3146-3150.

[7] Kobelco Systems Corporation. 2021. "Breakthrough of Natural Language Processing," 15 Aug. 2022; https://www.kobelcosys.co.jp/column/itwords/20211101/

[8] Ledge.ai. 2020. "What is BERT? Explaining the Mechanism of Natural Language Processing, which Google proud of," 12 Aug. 2022; https://ledge.ai/bert/.

[9] Matshakayile-Ndlovu, "The flash-back and the flash-forward, Zambie 29.2, De-pt. Languages Literature, Univ. Zimbabwe, 2001, pp. 189-198.

[10] E. Ochs, E, "Becoming a Speaker of Culture." in Language Acquisition and Language Socialization: Ecological Perspectives, ed C. Kramsch (New York, NY: Continuum), 2002, pp. 99-120.

[11] H. Sacks, E. Schegloff and G. Jefferson, "A Simplest Systematics for the Organization of Turn-Taking for Conversation. Language" vol. 50, no. 4, 1974, pp. 696-735.

[12] E. Takamori, E., “The comparative study of Japanese and Spanish -The feature of feedback towards opinion,” Tokyo University of Foreign Studies, 2004. “unpublished master’s thesis.”