# Detecting User Initiative in Social Chatbots

Stanford CS224N Custom Project

**Lora Xie**
Department of Computer Science
Stanford University
loraxie@stanford.edu

**Jack Liu**
Department of Computer Science
Stanford University
jiayiliu@stanford.edu

## Abstract

Mixed initiative, or two parties taking turns leading the conversation, is an important feature of human-human conversations, but one that social chatbots have difficulty with. In this project, we train a neural model that detects the user' initiative level, as a basis for the future task of conditioning bot response on user initiative. We have annotated our own data and finetuned two pretrained models, BERT and DialogRPT, the latter with stacking ensemble learning. Our models achieve moderate improvement over the baseline. We discuss the limitations of our models in detail and lay out a plan for future work.

## 1 Key Information to include

- Mentor: Ethan Chi
- External Collaborators (if you have any): No
- Sharing project: No

## 2 Introduction

In a successful, smooth, and comfortable conversation between two people, it is often the case that one person exhibits a higher level of initiative (leads the conversation) and the other exhibits a lower level of initiative (follows the conversation). During a human-chatbot conversation, we wish the conversation to exhibit the same mixed initiative characteristics as a human-human conversation, since the point of a chatbot is to simulate human conversation. With the mixed initiative feature implemented, we can expect the chatbot to create more realistic responses and keep a better conversation flow.

The first step in creating an initiative-considering chatbot is to identify the level of initiative of any given sentence. There has been previous attempts to measure sentence initiative with sentence characteristics (see section 3 Related Work); however, that approach provides only a relative guess, not a concrete measurement. Therefore, we first improved and expanded on previous initiative rubric, then hand annotated 600 turns of conversation based on the rubric as our training and testing data. Lastly, we built and fine tuned two pretrained neural models, one with Bert and another with DialogRPT, to classify the level of initiative of a given sentence. With these models, we have a reached a highest accuracy of 0.72 and lowest Mean Squared Error of 0.84, which is a significant improvement comparing to our baselines.

## 3 Related Work

The original initiative detection method based on sentence characteristics is presented in *Hardy, et. al.*[1]. In their approach, number of words ( tokens), the number of noun phrases ( NPs), and negative log likelihood (NLL) of a sentence are shown to be proportional to the initiative level.

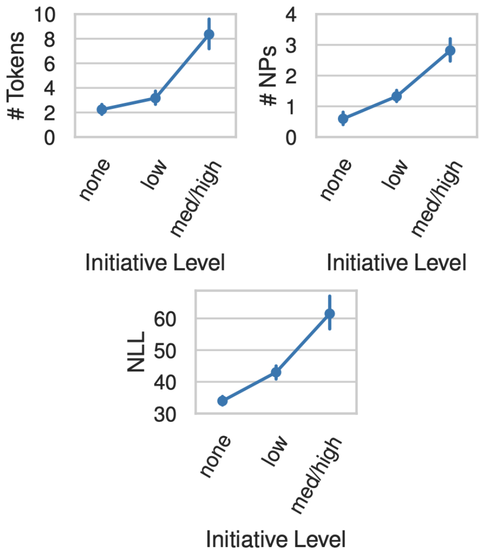

Figure 1: Automated Metrics vs. Hand-Labeled Initiative Levels. Bars show 95% confidence intervals[1]

In the original paper, the accuracy of this method is not recorded. However, this relationship between initiative level and sentence characteristics is still a very useful insight as we use this to produce one set of our baselines (see section 4.1 Baselines).

# 4 Approach

## 4.1 Baselines

We implement three sets of metrics to compute a total of seven different baselines.

The first set of metrics is based on the three evaluation metrics presented in *Hardy, et. al.*[1]. As previously mentioned, these metrics are sentence length ( Tokens), number of noun phrases ( NPs), and negative log likelihood of the sentence (NLL). We use *SpaCy* [2] part-of-speech tagging to count the number of noun phrases, and *Lin's* implementation [3] of Huggingface's pre-trained GPT2 transformer model[4] to compute the sentence probability.

The second set of metrics is based on sentence embeddings. To compute a sentence embedding, we first use the pre-trained GloVe [5] word vectors to find the embedding vector of each word of a sentence. We then perform vector addition on these embedding vectors to produce a sum vector. Lastly, normalize this sum vector by dividing the sum vector by the number of words in this sentence.

Note all the "sentences" being treated in these metrics are the human utterance of each bot-human turn in a dialogue (See 5.1 Data for more information about turns). However, intuitively, human initiative should be effected by the bot utterance and conversations of the previous turns as well. This is an idea we will explore with our model.

For these two sets of metrics, we use scikit-learn's [6] random forest classifier, decision tree classifier, and SGD classifier (a linear classifier) to model the level of initiative of a sentence. Levels of initiatives are measured from 1 - 5 (See 5.1 Data for more information).

The third set of metric is simply predicting everything as the most frequent initiative level.

We train these models with $80\%$ of our data and test with the remaining $20\%$. The results (baselines) in accuracy and mean squared error (MSE) based on our current data are described in Figure 2 below:

| Method | Accuracy | MSE |
|---|---|---|
| Most Frequent Level | 0.33 | 5.03 |
| Evaluation Metrics Random Forrest | 0.54 | 1.51 |
| Evaluation Metrics Decision Tree | 0.50 | 1.42 |
| Evaluation Metrics Linear Regression | 0.21 | 2.21 |
| GloVe Embedding Random Forrest | 0.59 | 1.79 |
| GloVe embedding Decision Tree | 0.45 | 2.28 |
| GloVe embedding Linear Regression | 0.59 | 1.66 |

Figure 2: Baselines

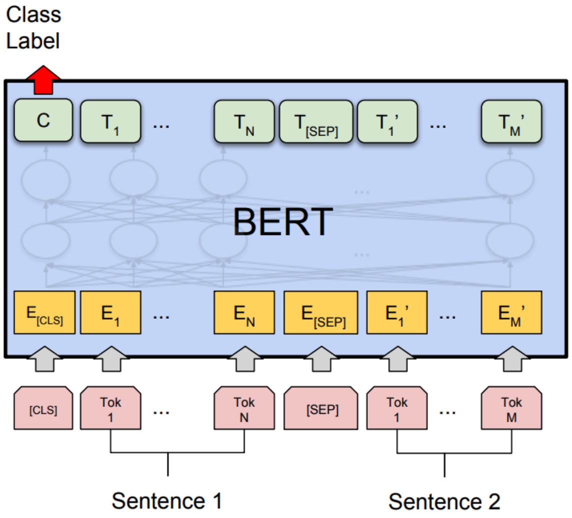

Figure 3: BERT Classification

### 4.2 Approach

Our main approach is to finetune for initiative-detection on large pretrained models. With BERT, we used encoding for the [CLS] token in its output and feed it into a multi-class classification neural layer. We used Huggingface's `BertForSequenceClassification` module for this. With DialogRPT[7], which was a GPT model trained on Reddit posts and predicts human feedback of dialogue responses, we hypothesize that three of the feedback dimensions it evaluates are related to initiative level: updown (how likely a response gets the most upvotes), width (how likely a response gets the most direct replies), and depth (how likely the response gets the longest follow-up thread). We finetune each of the three sub-models (updown, width, and depth) to predict a class label in $\{1, \dots, 5\}$. We test with and without freezing the decoding layers of the GPT base model. We also apply stacking ensemble learning [8], where we train an additional linear layer to combine the predictions made by each of the three sub-models to produce a final prediction.

## 5 Experiments

### 5.1 Data

We trained our baselines and model with the ConvAI2 dataset [9], which comprises of dialogues between human evaluators and chatbots. We manually annotated the user-initiative of 600 turns. We first screened for incoherent and/or confusing bot utterance and adversarial/complaining user utterance, since in turns involving these elements it is not clear whether discussing initiative would be meaningful. Moreover, empirically speaking, these turns tend to have poor cross-annotator agreement. We then labeled the rest of the turns from 1 to 5 using the below rubric for evaluation. This rubric is based on *Hardy, et. al.*, which only evaluated human response to bot questions. We followed the

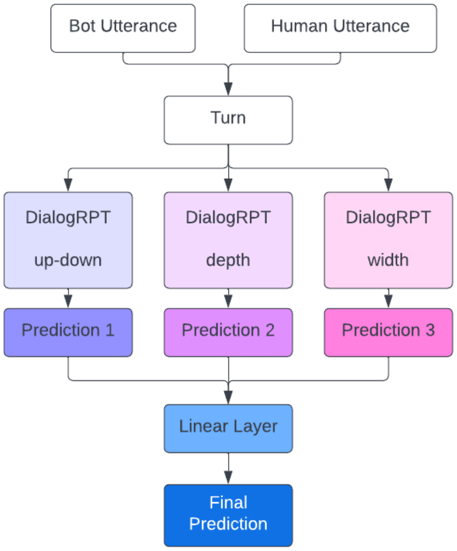

Figure 4: Stacking Ensemble Learning

overarching definition of initiative as how much a response alters the course of the conversation and contributes to the topic, and we expanded the application to other non-question-answering types of responses. For the purpose of detecting initiative level, we trained and evaluated with only turns labeled between 1 and 5. The complete rubric is as follows:

- 1 - None: does not alter course of conversation, no or minimum contribution to topic
  - Yes/No responses to binary questions, may return the question w/o modifying it
    - E.g. > do you have any animals ? > no, and you?
  - Uninformative answers
  - For statements: Agreement / disagreement / acknowledgement
    - E.g. > i do not have any hobbies . > Really?
    - E.g. > Hi, I am listening to some classical music. > you told me that already.... nevermind
  - Conventional greeting/closing/etiquette
    - E.g. > my wife just became legally deaf > Sorry for that
- 2 - Low: slight contribution to topic OR slightly alter course of conversation
  - Responses to closed-ended questions without extra information, may return the question w/o modifying it
    - E.g. > I am sorry. Do you like music? > Yes i do like music. What about you?
  - Clarifying questions (either the question was unclear, or confirming mostly by repetition to be sure)
    - E.g. > I like all kinds. I like rock, and i like to listen to music. > wow, you love rock music??
    - E.g. > I am 24 and I have a dog named her name is named name > Really her name is name?
  - Dismissing bot's or own previous turn
    - E.g. > i do , but i do not have a lot of money . > Forget that picture it was not meant for you
- 3 - Medium: substantive contribution to topic OR changing the course of convo
  - Responses to open-ended questions
  - Responses that share unprompted information / opinion

        - E.g. > i'm not sure , but i do not have a lot of money . > Spend it wisely !

- 4 - High: Explicitly taking the lead
    - Questions & questioning
        - E.g. > I'm not going to be late. > Where are you goin?
        - E.g. > pretty good  things have been slow > which ones?
    - Commanding/requesting a topic or other kind of request naturally (not necessarily completely the same topic)
        - E.g. > I am 22 and i am not sure what to do with them. > do you like coimputer games 4
        - E.g. > yes , but i do have a greenhouse . > Send me picture of you please
        - E.g. > I like to read. > So you should ask me as well
- 5 - Abrupt: Unnatural, impolite, forceful, uncooperative
    - Commanding/requesting a topic unnaturally
        - E.g. > where are you from > i am from east asia where exactly are you from?
        - E.g. > I love sports > We should talk about music
    - Testing the bot on info about user previously revealed to the bot in an impolite fashion
    - Not answering the question / completely unrelated response

Note that if an utterance matches to more than one categories listed above, we take the category with the higher value as the utterance's level of initiative.

Human labeling the dataset based on a rubric is a very subjective process. Therefore, to ensure the consistency of the labeling process between different people, we have to revise this rubric multiple times since the milestone version. With this version of the rubric, we are able to achieve an inter-rater reliability, measured by Cohen's Kappa, of 0.916, which means the labeling between different people are mostly consistent.

Besides hand labeling data, we are also able to utilize the Switchboard Dialog Act Corpus as our silver data. This corpus consists of transcripts of 1,155 minutes of telephone conversation between two people, producing 221,616 utterances. With each utterance, the corpus provides a list of DAMSL speech act tag. After analyzing these tags, we mapped most of these tags with an initiative level. Thus, for each utterance, we are able to produce an initiative labeling from the provided tags.

### 5.2 Evaluation method

We used two automatic quantitative evaluation metrics. First is simply accuracy (percentage of correct levels predicted) as this is a single-label classification task. Secondly, because the levels are on a spectrum and have an inherent ordering, we also used Mean Squared Error to capture the distance of predictions from the correct levels.

Since there is no prior work on initiative detection, we compare our results to the baselines results, which were also measured with the aforementioned two metrics.

### 5.3 Experimental details

Of the 600 turns we annotated, 480 have a sensible initiative level between 1 and 5. We divided them into a train set of 324 turns, a dev set of 89 turns, and a test set with 67 turns. We used batch size 8, learning rate 5e-05, and 30 epochs for all models. The runtime were all under five minutes.

### 5.4 Results

We found that BERT has the highest accuracy for our initiative detection task, while the ensembled DialogRPT models have the lowest MSE. Both are a moderate improvement over the best-performing baselines with the two metrics (GloVe Embedding Random Forest and Evaluation Metrics Decision Tree). Freezing turns out to be very unhelpful.

| Model | Accuracy | MSE |
|---|---|---|
| GloVe Embedding Random Forest | *0.59* | 1.79 |
| Evaluation Metrics Decision Tree | 0.50 | *1.79* |
| BERT | **0.72** | 1.08 |
| updown | 0.60 | 1.12 |
| updown (freeze) | 0.29 | 1.89 |
| depth | 0.61 | 0.94 |
| depth (freeze) | 0.31 | 1.60 |
| width | 0.60 | 0.97 |
| width (freeze) | 0.39 | 1.42 |
| ensemble | 0.68 | **0.84** |

Table 1: Results of Different Models

## 6 Analysis

We did not observe any systematic bias in our models' prediction towards higher or lower initiative, and most erroneous predictions are off by 1. However, for both BERT and DialogRPT, we see patterns where the models being mistaken when a more straightforward characteristic of the turn dominates the subtlety of the rest of the sentence, for example, in:

Bot: That is great! Do you like movies?
User: Yes, mostly Disney movies

Since there is a clear Y/N question and answer, our models is led to incorrectly predict a lower initiative. Another example is:

Bot: yes, i do, but i do not have a lot of money
Human: Help is free

Here the model is probably fooled by the shortness of the human response.

We speculate that because of the small datasize, the models haven't seen enough of these more complicated examples in their training data to learn how to balance when there are conflicting signals (e.g. a straight forward Y/N answer to a Y/N question vs. an expansion to the Y/N).

## 7 Conclusion

We chieved preliminary results on user initiative detection in chatbots with both BERT and DialogRPT, and given the limitations of our exploration, we believe both models still have significant room for improvement.

There are several limitations to our models. First, we made the simplifying assumption that removing context does not greatly impair initiative detection, which often does not hold, as humans frequently refer to what has been previously spoken, which would seem like an abrupt change of topic if our examination window is limited to only one turn. If we had more time, we would explore incorporating a context window.

Second, we are rather limited in our dataset size, especially after removing turn data in which the bot's utterance is either incoherent or irrelevant, which we found to result in user confusion and thus complaining or corrective behaviors that defy a regular rating with respect to initiative level. The fact that our models' losses and accuracies often start to plateau after as few as 10 epochs might be a sign that our dataset is too small. In an attempt to expand our dataset, we had collected silver data from the Switchboard Dialogue Act dataset but did not end up training on it yet due to time limit. We also plan to explore unsupervised pre-training on the ConvAI dataset and see if it would improve model performance without costing additional data annotation.

Third, we have discovered in the process of annotating that the bots in our data can vary greatly in their style, and strengths/weaknesses, which make their data less generalizable, especially when our dataset is small. A potential solution to make up for the noise created by bot peculiarities is to include some human-human data, which is more high quality and stable.

## References

[1] Amelia Hardy, Ashwin Paranjape, and Christopher Manning. Effective social chatbot strategies for increasing user initiative. In *Special Interest Group on Discourse and Dialogue (SIGDIAL)*, 2021.

[2] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. 2017.

[3] Yuchen Lin. Compute sentence probability using gpt-2 with huggingface transformers. 2020.

[4] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Perric Cistac, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-Art Natural Language Processing. pages 38–45. Association for Computational Linguistics, 10 2020.

[5] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[6] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. volume 12, pages 2825–2830, 2011.

[7] Xiang Gao, Yizhe Zhang, Michel Galley, Chris Brockett, and Bill Dolan. Dialogue response rankingtraining with large-scale human feedback data. In *EMNLP*, 2020.

[8] David H. Wolpert. Stacked generalization. In *Neural Networks*, 1992.

[9] Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W Black, Jason Rudnicky, Alexander andd Williams, Joelle Pineau, Mikhail Burtsev, and Jason Weston. The Second Conversational Intelligence Challenge (ConvAI2). 2019.