A decision model for designing NLP applications

Eason Chen*, Yuen-Hsien Tseng
National Taiwan Normal University
eason.tw.chen@gmail.com

ABSTRACT
Among the NLP models' usages, we notice that some programs provide multiple output options, and some offer only a single result to the end-users. However, there is little research about which situations providing multiple outputs from NLP models will benefit the user experience. Therefore, in this position paper, we summarize the progress of NLP applications, which shows parallel outputs from the NLP model at once to users. Then a decision model is presented that can assist in deciding whether a given condition is suitable to show multiple outputs at once from the NLP model. We hope developers and UX designers can examine the decision model and create an easy-to-use interface that can present numerous results from the NLP model at once to create a more satisfactory user experience.

CCS CONCEPTS
• Human-centered computing • Human computer interaction (HCI) • HCI theory, concepts and models

KEYWORDS
Natural Language Processing, Natural Language Generation, Decision Model, User Experience

1 Introduction
Natural Language Processing (NLP) has been widely used in various applications, such as text generation [1], chatting services [2], revising writing [3], identifying named entities [4], searching information [5], translation [6], question answering, or even creating images [7].

Among these usages, we notice that some programs provide multiple output options and some provide single options to the end-users. However, there is little research on which situations providing multiple outputs from NLP models will benefit the user experience. Therefore, in this position paper, we are trying to summarize the existing progress about NLP applications which shows parallel outputs at once from the NLP model to users, and then propose a guideline that future UX designers and app developers can reference to decide whether their applications are suitable for this method. Furthermore, we hope future researchers can reference the decision model in this paper to explore the potential of other NLP models' usage that utilizes parallel outputs to create a more satisfactory user experience.

2 Related Work
In this part, we will discuss different NLP model usage scenarios and how showing multiple outputs from the NLP model at once can benefit user experience in these contexts.

2.1 Text to Text Generation
To generate text, Natural Language Generation (NLG) models adopt special decoding strategies [8] to randomly generate one token at a time based on the inputted text and currently generated tokens until a specific stop criterion is met. The generated outputs will be different in these NLG models even when the same input sequence is given.

GPT-2 models are the most widely used Natural Language Generation models [1]. In demos of NLG, authors often show parallel outputs to demonstrate the ability to generate various outputs by GPT-2 models. For example, at Allen NLP’s demo website [9], users can decide from the following word choices in the sentence based on the inputted text and currently generated tokens (Fig 1).

![Figure 1: screenshot of the GPT-2 demo from Allen NLP](image-url)
Several researches used GPT-2’s parallel and independent output features to provide end-users with multiple outputs and a better user experience.

2.2 Parallel Email Writing Suggestions

The research from Buschek, Zürn [3] indicated that when providing phrase suggestions to users when writing emails (Fig 2), users’ adoption rate from the tips will increase, and their ideation will benefit when providing multiple suggestions simultaneously. Nevertheless, the authors also point out that efficiency will decrease when showing more outputs.

![Figure 2: interface that provides multiple writing suggestions at once when writing an email [3].](image)

2.3 Number of replies and NLG Chatbot’s UX

Previous research undertaken by this author [2] has shown that when chatting with GPT-2 based NLG chatbots, by responding to users’ utterances with multiple replies to create a group chat atmosphere, the chatting experience score will increase significantly (Fig 3).

According to the author’s survey, users tend to pay attention to acceptable replies and ignore unsuitable replies when chatting. That is, users will interpret the information they receive in a way that they can understand. Since there are multiple replies, users might think the inappropriate replies are not to them, but are response to other chatbots in the group. Also, users can pick the replies they like to continue the conversation.

![Figure 3: The different numbers of reply conditions on a dialog system from Chen’s research [2].](image)

2.4 Question Answering and Search Engine

Apart from NLG, providing multiple results also occurs at other services. For example, search engines, like the Google Search result (Figure 4), often show numerous candidate results for users to decide which search result is the most relevant one.

![Figure 4: Screenshot of Google Search Results](image)

2.5 NER when classify the type of keywords

The Named Entity Recognition research by Ratinov and Roth [4] states that some words have several meanings and that they need to be viewed based on their context. That is, given one word might belong to different named entities. For example, in the famous baseball comedy act “Who is on first?”, “Who” might be a Name or just a question word. Therefore, we assume that in some cases it might be better to provide the NER’s probability output directly to humans and decide.

2.6 Translation

Current translation services such as Google Translate provide multiple translation results and their frequency to the end-user. Thus, users can decide which word they should use in their respective contexts. (Fig 5)

![Figure 5: Screenshot of Google Translate that it lists possible translation results and their frequency in the bottom right corner.](image)

2.7 GAN network (text-to-image)

Apart from pure NLP, when using other models such as text-to-image GAN network [7], the program often shows parallel outputs at once, too. Then, because users know the model is not perfect, users can cherry-pick the best output they received and

### 2.2 Parallel Email Writing Suggestions

The research from Buschek, Zürn [3] indicated that when providing phrase suggestions to users when writing emails (Fig 2), users’ adoption rate from the tips will increase, and their ideation will benefit when providing multiple suggestions simultaneously. Nevertheless, the authors also point out that efficiency will decrease when showing more outputs.

![Figure 2: interface that provides multiple writing suggestions at once when writing an email [3].](image)

### 2.3 Number of replies and NLG Chatbot’s UX

Previous research undertaken by this author [2] has shown that when chatting with GPT-2 based NLG chatbots, by responding to users’ utterances with multiple replies to create a group chat atmosphere, the chatting experience score will increase significantly (Fig 3).

According to the author’s survey, users tend to pay attention to acceptable replies and ignore unsuitable replies when chatting. That is, users will interpret the information they receive in a way that they can understand. Since there are multiple replies, users might think the inappropriate replies are not to them, but are response to other chatbots in the group. Also, users can pick the replies they like to continue the conversation.

![Figure 3: The different numbers of reply conditions on a dialog system from Chen’s research [2].](image)

### 2.4 Question Answering and Search Engine

Apart from NLG, providing multiple results also occurs at other services. For example, search engines, like the Google Search result (Figure 4), often show numerous candidate results for users to decide which search result is the most relevant one.

![Figure 4: Screenshot of Google Search Results](image)

### 2.5 NER when classify the type of keywords

The Named Entity Recognition research by Ratinov and Roth [4] states that some words have several meanings and that they need to be viewed based on their context. That is, given one word might belong to different named entities. For example, in the famous baseball comedy act “Who is on first?”, “Who” might be a Name or just a question word. Therefore, we assume that in some cases it might be better to provide the NER’s probability output directly to humans and decide.

### 2.6 Translation

Current translation services such as Google Translate provide multiple translation results and their frequency to the end-user. Thus, users can decide which word they should use in their respective contexts. (Fig 5)

![Figure 5: Screenshot of Google Translate that it lists possible translation results and their frequency in the bottom right corner.](image)

### 2.7 GAN network (text-to-image)

Apart from pure NLP, when using other models such as text-to-image GAN network [7], the program often shows parallel outputs at once, too. Then, because users know the model is not perfect, users can cherry-pick the best output they received and
discard those unacceptable outputs, just like the example from flax-community/dalle-mini at Fig 6.

![DALL·E mini](https://huggingface.co/spaces/flax-community/dalle-mini)

Figure 6: Text-to-Image demo from https://huggingface.co/spaces/flax-community/dalle-mini

3 Discussion and Guidelines

Based on the previous related works, we will summarize the best practices in which selecting cases showing multiple output options at once is more appropriate than offering only one output option. Then we will propose a decision model that determines whether a given case is reasonable to provide parallel outputs.

3.1 Conditions that Can show multiple outputs at once

There are three key conditions that show multiple outputs at once that might benefit the user experience.

First, if the results from the same input are independent or highly randomizable, showing multiple outputs from NLP models might be better. For example, we can create a separate and independent output for a classification model that uses sigmoid instead of SoftMax as the final layer.

Another example is how NLG models such as GPT-2 use decoding strategies to generate the next token randomly.

Second, if the output has no standard answer, such as in the case of the generation model, showing multiple outputs will work out better. In a purely rational analysis [2], if the general probability that one output fits the user’s needs is p(x) (assumed 0.3), then the probability that the outcome is unacceptable is 1 - p(x) (assumed 0.7). Nevertheless, the only thing a user wants is a proper output. If we present several outputs at once, the probability that the user can not find any appropriate outputs will be greatly decreased, from 1 - p(x) to (1 - p(x))^2 (assumed 0.7^2 = 0.16807). In other words, the chance that the user can find at least one acceptable output will be significantly increased from p(x) to 1 - (1 - p(x))^2 (1 - 0.7^2 = 0.83193).

Third, if the outcomes are presented directly to humans, showing multiple outputs from NLP models will be superior. According to the spirit of Augmented Intelligence [10], a human can generally make better decisions than AI, so AI can serve as an assistant when humans make their decisions. Therefore, providing an easy-to-use interface to allow users to decide which output from the NLP model is suitable will benefit the user experience.

3.2 Conditions that are inappropriate to show multiple outputs at once

In contrast, there still may be many conditions that are not suggested to show multiple outputs from the NLP model at once.

First, if the output is neither independent nor randomizable, showing multiple outputs from NLP models will not benefit the user. An example of this is a single regression output.

Second, if the context cannot provide multiple outputs to the end-user, like the voice assistant, showing multiple outputs from NLP models will not come in handy.

Third, if the efficiency is more important than the quality, or the model is already of adequate quality, showing multiple outputs will worsen the user experience since users need time to find the output they want [3].

Apart from that, if other programs, such as the downstream Machine Learning models, process the output, showing multiple outputs from NLP models might be beneficial or not, depending on one’s design. If the downstream model uses the outputs as feature vectors or can distinguish which outputs are better, it might benefit the final output’s quality. For example, Yang and Tseng [11] designed a GPT-2 and BERT dual-model reply generation system in which the GPT-2 NLG model offers several outputs to the BERT regression model. After receiving the input message, GPT-2 will first generate many outputs and use BERT to predict replies’ coherence score. Then, the system will sort the coherence score and use the top n replies as its output.

3.3 Decision Model

![Decision Model](https://example.com/decision-model.png)

Figure 7: The decision model for whether or not to present multiple outputs from the NLP model.
Based on the above discussions, we propose the decision model shown in the Figure 7 to distinguish whether an NLP application is suitable for presenting multiple outputs.

The decision model can be used to determine whether a given application is appropriate to provide multiple outputs from the NLP model at once.

First, suppose one is using an NLG chat-oriented dialog system[2]. In that case, the context allows showing multiple outputs at once, the NLP model’s outputs are independent, and humans will read outputs. Thus, the decision model will show a positive result, suggesting the user presents multiple outputs simultaneously. However, if one is using a voice assistant, the context cannot show multiple outputs at once, resulting in a negative result. Furthermore, if the input leads to a simple standard answer, such as when asking the local time, showing multiple outputs is unnecessary, and the outcome from the decision model will be negative.

Second, when using a QA or Translation NLP model, showing multiple outputs at once will be better because the input might be ambiguous, namely, leads to no standard answer. However, showing numerous results from the NLP model is only acceptable when inputting short content or having high-quality requirements because manually viewing each result is time-consuming. If the QA or Translation model is good enough, the interface can show one result to save users’ time and provide an option for users to view more results from their input.

Third, if the NLP model is being used for creative purposes, such as writing articles [9] or creating images [7], showing many outputs together might benefit the creator to decide which output he wants.

Lastly, if the NLP models’ outputs will be read by downstream programs but not humans, showing multiple outputs concurrently or not will depend on the programs’ design. If the downstream programs can judge the quality of the outputs [11], the designer can let the NLP model show multiple outputs.

### 4 Conclusion and Future Work

In this paper, we present a decision model that can decide whether a given condition is suitable to show multiple outputs at once from the NLP model. We hope developers and UX designers can examine the decision model and create an easy-to-use interface that can present numerous results from the NLP model at once. Moreover, we hope future researchers can reference and reinforce the decision model from this paper to explore the potential of other NLP models’ usage that can show parallel outputs at once to create a more satisfactory user experience.

In the future, we are trying to use state-of-the-art NLP models to build and evaluate an open-domain dialog system. We plan to let our applications follow the decision model proposed in this paper in the system. The program will sometimes provide multiple outputs depending on several conditions, such as the input from the user or the purpose of the conversation.

### 4 Acknowledgements

This work was supported by the Ministry of Science and Technology of Taiwan (R.O.C.) under Grants 110-2813-C-003-033-E. We would like to thank Tsung-Ren Huang for his suggestion to write this paper. Furthermore, we thank the anonymous reviewers for their helpful suggestions.

### REFERENCES


