Putting Humans in the Natural Language Processing Loop: A Survey

Zijie J. Wang*   Dongjin Choi*   Shenyu Xu*   Diyi Yang
College of Computing, Georgia Tech
{jayw, jin.choi, shenyuxu, dyang888}@gatech.edu

Abstract

How can we design Natural Language Processing (NLP) systems that learn from human feedback? There is a growing research body of Human-in-the-loop (HITL) NLP frameworks that continuously integrate human feedback to improve the model itself. HITL NLP research is nascent but multifarious—solving various NLP problems, collecting diverse feedback from different people, and applying different methods to learn from human feedback. We present a survey of HITL NLP work from both Machine Learning (ML) and Human-computer Interaction (HCI) communities that highlights its short yet inspiring history, and thoroughly summarize recent frameworks focusing on their *tasks*, *goals*, *human interactions*, and *feedback learning methods*. Finally, we discuss future studies for integrating human feedback in the NLP development loop.

1 Introduction

Traditionally, Natural Language Processing (NLP) models are trained, fine-tuned, and tested on existing dataset by machine learning experts, and then deployed to solve real-life problems of their users. Model users can often give invaluable feedback that reveals design details overlooked by model developers, and provide data instances that are not represented in the training dataset (Kreutzer et al., 2020). However, the traditional linear NLP development pipeline is not designed to take advantage of human feedback. Advancing on conventional workflow, there is a growing research body of Human-in-the-loop (HITL) NLP frameworks, or sometimes called mixed-initiative NLP, where model developers continuously integrates human feedback into different steps of the model deployment workflow (Figure 1). This continuous feedback loop cultivates a human-AI partnership that enhance model performance and build users' trust in the NLP system.

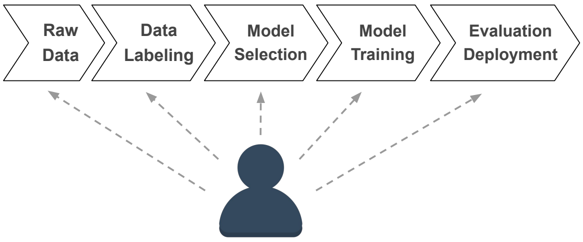

Figure 1: Collaboration between humans and models under a human-in-the-loop Natural Language Processing paradigm. Humans provide various types of feedback in different stages of the workflow to improve the model's performance, interpretability, and usability.

Just like traditional NLP frameworks, there is a high-dimensional design space for HITL NLP systems. For example, human feedback can come from end users (Li et al., 2017) or crowd workers (Wallace et al., 2019), and human can intervene models during training (Stiennon et al., 2020) or deployment (Hancock et al., 2019). Good HITL NLP systems need to clearly *communicate* to humans of what the model needs, provide intuitive *interfaces* to collect feedback, and effectively *learn* from them. Therefore, HITL NLP research spans across not only NLP and Machine Learning (ML) but also Human-computer Interaction (HCI). A meta-analysis on existing HITL NLP work focusing on bridging different research disciplines is vital to help new researchers quickly familiarize with this promising topic and recognize future research directions. To fill this critical research gap, we provide a timely literature review on recent HITL NLP studies from both NLP and HCI communities.

This is the first survey on the HITL NLP topic. We make two main contributions: (1) We summarize recent studies of HITL NLP and position each work with respect to its *task*, *goal*, *human interaction*, and *feedback learning method* (Table 1);

[*] denotes equal contribution

Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing, pages 47–52
April 20, 2021. ©2021 Association for Computational Linguistics

(2) We highlight important research directions and open problems that we distilled from the survey.

2 Human-in-the-Loop NLP Tasks

In this section, we categorize surveyed HITL paradigms based on their corresponding tasks.

2.1 Text Classification

Text classification is a classic NLP task to categorize text into different groups. Many HITL frameworks are developed for this problem, where most of them start with *training* a text classifier, then recruiting humans to *annotate* data based on the current model behavior, and eventually *retraining* the classifier on the larger dataset continuously. For example, Godbole et al. (2004) develop a HITL paradigm where users can interactively edit text features and label new documents. Also, Settles (2011) integrates active learning in their framework—instead of arbitrarily presenting data for users to annotate, samples are selected in a way that maximizes the expected information gain. With active learning, labelers can annotate fewer data to achieve the same model improvement of a framework using random sampling.

2.2 Parsing and Entity Linking

Besides classifying documents, recent research shows great potential of HITL approach in enhancing the performance of existing parsing and entity linking models. Advancing traditional Combinatory Categorial Grammars (CCG) parsers, He et al. (2016) crowdsource parsing tasks—a trained parser is uncertain about—to non-expert mechanical turks, by asking them simple what-questions. Also, with more strategic sampling methods to select instances to present to humans, a smaller set of feedback can quickly improve the entity linking model performance (Klie et al., 2020).

2.3 Topic Modeling

In addition to use HITL approach to enhancing learning low-level semantic relationships, researcher apply similar framework to topic modeling techniques that are used to analyze large document collections (Lee et al., 2017). For example, Hu et al. (2014)'s systems allow users to refine a trained model through adding, removing, or changing the weights of words within each topic. Recent work also focuses on human-centered HITL topic modeling methods. Kim et al. (2019) develop an intuitive visualization system that allows end users to up-vote or down-vote specific documents to inform their interest to the model. Smith et al. (2018) conduct users studies with non-experts and develop a responsive and predictable user interface that supports a broad range of topic modeling refinement operations. These examples show that NLP HITL systems can benefit from HCI design techniques.

2.4 Summarization and Machine Translation

HITL can be used in text summarization and machine translation. For instance, Stiennon et al. (2020) collects human preferences on pairs of summaries generated by two models, then train a reward model to predict the preference. Then, this reward model is used to train a policy to generate summaries using reinforcement learning. Kreutzer et al. (2018) collect both explicit and implicit language human feedback to improve a machine translation model by using the feedback with reinforcement learning. Experiments show that these models have higher accuracy and better generalization.

2.5 Dialogue and Question Answering

Recently, many HITL frameworks have been developed for dialogue and Question Answering (QA) systems, where the AI agent can have conversation with users. We can group these systems into two categories: *online feedback loop* and *offline feedback loop*. With online feedback loop, the system continuously uses human feedback to update the model. For example, Liu et al. (2018) collects dialogue corrections from users during deployment, and then use online reinforcement learning to improve the model. With offline feedback loop, model is updated after collecting a large set of human feedback. For instance, Wallace et al. (2019) invites crowd workers to generate adversarial questions that can fool their QA system, and use these questions for adversarial training. Offline feedback loop can be more robust for dialogue systems, because user feedback can be misleading so directly updating the model is risky (Kreutzer et al., 2020).

2.6 Human-in-the-Loop Goals

Among surveyed papers, the most abundant motivation for using a HITL approach in NLP tasks is to improve the **model performance**. For example, with a relatively small set of human feedback, HITL can significantly improve the model accuracy (Smith et al., 2018), model robustness

	TASK					GOAL			INTERACTION							UPDATE		
Work	Text Classification	Parsing and Entity Linking	Topic Modeling	Summarization and Machine Translation	Dialogue and Question Answering System	Model Performance	Model Interpretability	Usability	Mediums – Graphical User Interface	Mediums – Natural Language Interface	User Feedback Type – Binary	User Feedback Type – Scaled	User Feedback Type – Natural Language	User Feedback Type – Counterfactual Example	Intelligent Interaction	Data Augmentation – Offline Model Update	Data Augmentation – Online Model Update	Model Direct Manipulation
Godbole et al. (2004)	●					●			●		●				●	●		
Settles (2011)	●					●			●		●				●	●		
Simard et al. (2014)	●					●			●		●	●				●		
Karmakharm et al. (2019)	●					●		●	●		●					●		
Jandot et al. (2016)	●						●		●		●					●		
Kaushik et al. (2019)	●					●				●			●	●		●		
He et al. (2016)		●				●			●		●					●		
Klie et al. (2020)		●				●		●	●		●					●		
Lo and Lim (2020)		●				●			●		●				●	●		
Trivedi et al. (2019)		●				●			●		●		●			●		
Lawrence and Riezler (2018)		●				●				●	●		●	●		●		
Kim et al. (2019)			●			●		●	●		●						●	
Kumar et al. (2019)			●			●			●		●	●					●	
Smith et al. (2018)			●			●	●	●	●		●	●					●	
Stiennon et al. (2020)				●		●			●			●			●	●		
Kreutzer et al. (2018)				●		●			●			●			●			●
Hancock et al. (2019)					●	●				●			●				●	
Liu et al. (2018)					●	●				●	●		●		●		●	●
Li et al. (2017)					●	●				●	●		●		●	●	●	●
Wallace et al. (2019)					●	●				●	●		●	●		●		

Table 1: Overview of representative works in HITL NLP. Each row represents one work. Works are sorted by their task types. Each column corresponds to a dimension from the four subsections (task, goal, human interaction, and feedback learning methods).

and generalization (Stiennon et al., 2020). Besides model performance, HITL can also improve the **interpretability** and **usability** of NLP models. For instance, Wallace et al. (2019) guides humans to generate adversarial questions that fool the question answering model—these adversarial questions are also used as probes for researchers to study the underlying model behaviors. In Smith et al. (2018)'s topic modeling work, user studies have shown that users gain more trust and confidence through the HITL system.

3 Human-machine Interaction

This section discusses the mediums that users use to interact with HITL systems and different types of feedback that the system collect.

3.1 Interaction Mediums

Graphical User Interface (GUI) provides a user interface that allows users to interact with systems through graphical icons and visual indicators. Some HITL NLP systems allow users to directly label samples in the GUI (Godbole et al., 2004). The GUI also makes feature editing possible for end-users who do not develop the model from initial (Simard et al., 2014). Some work even uses the GUI for users to rate training sentences in the text summarization task (Stiennon et al., 2020) and rank generated topics in the topic modeling task (Kim et al., 2019). One obvious advantage of the GUI is that it helps visualize NLP models, enhancing the interpretability of the model. In addition, the GUI supports *Windows, Icons, Menus, Pointer* (WIMP) interactions, providing users more accurate control for refining the models.

Natural Language Interface is an interface where the user interacts with the computer through natural language. As this interface usually simulates having a conversation with a computer, it mostly comes with the purpose of building up a dialogue system (Hancock et al., 2019). The natural language interface not only supports users to provide explicit feedback (Liu et al., 2018), such as positive or negative responses. It also allows users to give implicit feedback with natural language sentences (Li et al., 2017). Compared to the GUI, the natural language interface is more intuitive to use as it simulates the process of human's conversation and thus needs no additional tutorial. In particular, it naturally fits in dialogue systems.

3.2 User Feedback Types

Binary Feedback has two categories which are usually opposite to each other, such as "like" and "dislike". It can be collected by both the GUI and the natural language interface. GUIs can collect binary user feedback from the user's adding or removing labels (Settles, 2011) and features (Godbole et al., 2004). The natural language interface can also support binary user feedback collection with simple short natural language response, such as "agree" and "reject" (Liu et al., 2018).

Scaled Feedback has scaled categories and is usually in numerical formats, such as the 5-point scale rating. It often can only be collected through the GUI as it is difficult to express accurate scaled feedback in natural language. Such user feedback is collected in the GUI when users rate their preferences of training data or model results (Kreutzer et al., 2018) and adjust features on a numerical scale (Simard et al., 2014). Similar to binary user feedback, scaled user feedback can provide explicit feedback for the system to update the models (e.g. adjusting the weight of one feature from 1 to 3 on a scale of 5 points). Besides, the scaled ratings of user preferences can also be used as implicit guidance for improving the model.

Natural Language Feedback, comparing to binary user feedback and scaled user feedback, is better for representing users' intention but vague and hard for the machine to interpret. It can only be collected through the natural language interface. Users provide this type of user feedback by directly inputting natural language sentences to the system (Hancock et al., 2019). By analyzing the user input sentences, the system implies the user's intention and accordingly update the model.

Counterfactual Example Feedback, similar to the natural language user feedback, are usually collected through the natural language interface. The HITL NLP systems collect and analyze user-modified counterfactual text examples and retrain the model accordingly (Kaushik et al., 2019; Lawrence and Riezler, 2018).

3.3 Intelligent Interaction

As discussed in section 2, *active learning* is one commonly used technique we observed in our surveyed systems. Active learning allows the system to interactively query a user to label new data points with the desired outputs (Godbole et al., 2004). By strategically choosing samples to maximize information gain with fewer iterations, active learning not only reduces human efforts on data labeling but also improves the efficiency of the interface.

4 How to Use User Feedback

This section summarizes how existing HITL NLP systems utilize different types of feedback.

4.1 Data Augmentation

One popular approach is to consider the feedback as a new ground truth data sample. We describe two types of techniques to use augmented data set: *Offline Update* re-trains NLP model from scratch after collecting human feedback, while *Online Update* trains NLP models while collecting feedback.

Offline Model Update is usually performed after certain amount of human feedback is collected. Offline update does not need to be immediate, so they are suitable for noisy feedback with complex models which takes extra processing and training time. For example, Simard et al. (2014) and Karmakharm et al. (2019) use human feedback as new class labels and span-level annotations, and retrain their models after collecting enough new data.

Online Model Update is applied right after user feedback is given. This is effective for dialogue systems and conversational QA systems where recent input is crucial to machine's reasoning (Li et al., 2017). *Incremental learning* technique is often used to learn augmented data in real-time (Kumar et al., 2019). It focuses on making an incremental change to current system using the newly come feedback information effectively. Interactive topic modeling systems and feature engineering systems widely use this technique. For example, Kim et al. (2019) incrementally updates topic hierarchy by extending or shrinking topic tree incrementally. Also, some frameworks use Latent Dirichlet Allocation (LDA) to adjust sampling parameters with collected feedback in incremental iterations (Smith et al., 2018).

4.2 Model Direct Manipulation

Collected numerical human feedback are usually directly used to adjust model's objective function. For example, Li et al. (2017) collect binary feedback as rewards for reinforcement learning of a dialogue agent. Similarly, Kreutzer et al. (2018) uses a 5-point scale rating as reward function of reinforcement and bandit learning for machine translation. Existing works have focused more on numerical feedback than natural language feedback. Numerical feedback is easier to be incorporated into models, but provides limited information than natural language. For future research, incorporating more types of feedback (e.g., speech, log data) will be an interesting direction to gain more useful insights from humans. With more complex feedback type, it is critical to design both quantitative and qualitative methods to evaluate collected feedback, as they can be noisy just like any other data.

5 Conclusion and Future Directions

In this paper, we summarize recent literature on HITL NLP from both NLP and HCI communities, and position each work with respect to its task, goal, human interaction, and feedback learning method. The field of HITL NLP is still relatively nascent and we see many different design choices. We find improving model performance is the most popular goal among surveyed NLP HITL frameworks. However, researchers have found HITL method also enhances NLP model interpretability (Jandot et al., 2016) and usability (Lee et al., 2017). we encourage future NLP researchers to explore HITL as a mean to better understand their models and improve the experience of end users. One way is to design systems that take feedback from model engineers and end users beyond crowd workers.

Most of the HITL NLP systems are designed by NLP researchers. As human feedback is the core for HITL design, we believe that this field will be greatly benefited from a deeper involvement of the HCI community. For example, with a poorly designed human-machine interface, the collected human feedback are more likely to be inconsistent, incorrect, or even misleading. Therefore, better interface design and rigorous user study to evaluate interfaces can greatly enhance the quality of feedback collection, which in turn improve the downstream task performance.

To shed light on HITL NLP research from a HCI perspective, Wallace et al. (2019) explore the effect of adding model interpretation cues in the HITL interface on the quality of collected feedback; Schoch et al. (2020) investigate the impacts of question framing imposed on humans; similarly, Rao and Daumé III (2018) study how to ask good questions to which humans are more likely to give helpful feedback. In particular, we recommend future researchers to (1) consider integrating interactive visualization techniques into human-machine interfaces; (2) conduct user study to evaluate the effectiveness of their HITL system in addition to model performance; (3) share collected human feedback data and user study protocols with the community.

References

Shantanu Godbole, Abhay Harpale, Sunita Sarawagi, and Soumen Chakrabarti. 2004. Document Classification Through Interactive Supervision of Document and Term Labels. In *Knowledge Discovery in Databases: PKDD 2004*, volume 3202, pages 185–196. Springer Berlin Heidelberg, Berlin, Heidelberg.

Braden Hancock, Antoine Bordes, Pierre-Emmanuel Mazaré, and Jason Weston. 2019. Learning from Dialogue after Deployment: Feed Yourself, Chatbot! *arXiv:1901.05415 [cs, stat]*.

Luheng He, Julian Michael, Mike Lewis, and Luke Zettlemoyer. 2016. Human-in-the-Loop Parsing. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2337–2342, Austin, Texas. Association for Computational Linguistics.

Yuening Hu, Jordan Boyd-Graber, Brianna Satinoff, and Alison Smith. 2014. Interactive topic modeling. *Machine Learning*, 95(3):423–469.

Camille Jandot, Patrice Simard, Max Chickering, David Grangier, and Jina Suh. 2016. Interactive Semantic Featuring for Text Classification. *arXiv:1606.07545 [cs, stat]*.

Twin Karmakharm, Nikolaos Aletras, and Kalina Bontcheva. 2019. Journalist-in-the-Loop: Continuous Learning as a Service for Rumour Analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 115–120, Hong Kong, China. Association for Computational Linguistics.

Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. 2019. Learning the difference that makes a difference with counterfactually-augmented data. *arXiv preprint arXiv:1909.12434*.

Hannah Kim, Dongjin Choi, Barry Drake, Alex Endert, and Haesun Park. 2019. TopicSifter: Interactive Search Space Reduction through Targeted Topic Modeling. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 35–45, Vancouver, BC, Canada. IEEE.

Jan-Christoph Klie, Richard Eckart de Castilho, and Iryna Gurevych. 2020. From Zero to Hero: Human-In-The-Loop Entity Linking in Low Resource Domains. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6982–6993, Online. Association for Computational Linguistics.

Julia Kreutzer, Shahram Khadivi, Evgeny Matusov, and Stefan Riezler. 2018. Can Neural Machine Translation be Improved with User Feedback? *arXiv:1804.05958 [cs, stat]*.

Julia Kreutzer, Stefan Riezler, and Carolin Lawrence. 2020. Learning from Human Feedback: Challenges for Real-World Reinforcement Learning in NLP. *arXiv:2011.02511 [cs]*.

Varun Kumar, Alison Smith-Renner, Leah Findlater, Kevin Seppi, and Jordan Boyd-Graber. 2019. Why Didn't You Listen to Me? Comparing User Control of Human-in-the-Loop Topic Models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6323–6330, Florence, Italy. Association for Computational Linguistics.

Carolin Lawrence and Stefan Riezler. 2018. Counterfactual learning from human proofreading feedback for semantic parsing. *arXiv preprint arXiv:1811.12239*.

Tak Yeon Lee, Alison Smith, Kevin Seppi, Niklas Elmqvist, Jordan Boyd-Graber, and Leah Findlater. 2017. The human touch: How non-expert users perceive, interpret, and fix topic models. *International Journal of Human-Computer Studies*, 105:28–42.

Jiwei Li, Alexander H. Miller, Sumit Chopra, Marc'Aurelio Ranzato, and Jason Weston. 2017. Dialogue Learning With Human-In-The-Loop. *arXiv:1611.09823 [cs]*.

Bing Liu, Gokhan Tur, Dilek Hakkani-Tur, Pararth Shah, and Larry Heck. 2018. Dialogue Learning with Human Teaching and Feedback in End-to-End Trainable Task-Oriented Dialogue Systems. *arXiv:1804.06512 [cs]*.

Pei-Chi Lo and Ee-Peng Lim. 2020. Interactive Entity Linking Using Entity-Word Representations. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1801–1804, Virtual Event China. ACM.

Sudha Rao and Hal Daumé III. 2018. Learning to Ask Good Questions: Ranking Clarification Questions using Neural Expected Value of Perfect Information. *arXiv:1805.04655 [cs]*.

Stephanie Schoch, Diyi Yang, and Yangfeng Ji. 2020. This is a Problem, Don't You Agree?" Framing and Bias in Human Evaluation for Natural Language Generation.

Burr Settles. 2011. Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1467–1478, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Patrice Y. Simard, David Maxwell Chickering, Aparna Lakshmiratan, Denis Xavier Charles, Léon Bottou, Carlos Garcia Jurado Suarez, David Grangier, Saleema Amershi, Johan Verwey, and Jina Suh. 2014. ICE: enabling non-experts to build models interactively for large-scale lopsided problems. *CoRR*, abs/1409.4814.

Alison Smith, Varun Kumar, Jordan Boyd-Graber, Kevin Seppi, and Leah Findlater. 2018. Closing the Loop: User-Centered Design and Evaluation of a Human-in-the-Loop Topic Modeling System. In *Proceedings of the 2018 Conference on Human Information Interaction&Retrieval - IUI 18*, pages 293–304, Tokyo, Japan. ACM Press.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize from human feedback. *arXiv:2009.01325 [cs]*.

Gaurav Trivedi, Esmaeel R Dadashzadeh, Robert M Handzel, Wendy W Chapman, Shyam Visweswaran, and Harry Hochheiser. 2019. Interactive nlp in clinical care: Identifying incidental findings in radiology reports. *Applied clinical informatics*, 10(4):655.

Eric Wallace, Pedro Rodriguez, Shi Feng, Ikuya Yamada, and Jordan Boyd-Graber. 2019. Trick Me If You Can: Human-in-the-Loop Generation of Adversarial Examples for Question Answering. *Transactions of the Association for Computational Linguistics*, 7:387–401.