Crowdsourcing Diverse Paraphrases for Training Task-oriented Bots

Jorge Ramírez,[1] **Auday Berro,**[1] **Marcos Baez,** [1] **Boualem Benatallah,** [2,1] **Fabio Casati** [3]
[1] LIRIS – University of Claude Bernard Lyon 1; [2] University of New South Wales; [3] ServiceNow
jorge-daniel.ramirez-medina@univ-lyon1.fr

Abstract

A prominent approach to build datasets for training task-oriented bots is crowd-based paraphrasing. Current approaches, however, assume the crowd would naturally provide diverse paraphrases or focus only on *lexical* diversity. In this WiP we addressed an overlooked aspect of diversity, introducing an approach for guiding the crowdsourcing process towards paraphrases that are *syntactically* diverse.

Background & Motivation

Task-oriented chatbots (or simply bots) enable users to interact with software-enabled services in natural language. Such interactions require bots to process utterances (i.e., user input) like *"find restaurants in Milan"* to identify the user's intent. A prominent approach to build datasets for intent recognition models involves acquiring an initial set of seed utterances (for the intents) and then grow it by *paraphrasing* this set via crowdsourcing (Yaghoub-Zadeh-Fard et al. 2020b).

An important dimension to measure quality in this context is *diversity*, i.e., the breath and variety of paraphrases in the resulting corpus, which dictates the ability to capture the many ways users may express an intent. In this context, paraphrasing techniques generally rely on approaches that aim at introducing *lexical* and *syntactic* variations (Thompson and Post 2020). Lexical variations refer to changes that affect individual words, such as substituting words by their synonyms (e.g., *"search restaurants in Milan"*). Syntactic variations, instead, refer to changes in sentence or phrasal structure, such as transforming the grammatical structure of a sentence (e.g., *"Where can we eat in Milan?"*). While the development of techniques to introduce such lexical and syntactic variations is the focus of ongoing work in automatic paraphrasing (Berro et al. 2021), they are currently greatly under-explored in the crowdsourcing community.

Among the few contributions towards diversity, a prominent data collection framework involves turning crowd-based paraphrasing into an iterative and multi-stage pipeline. Here, multiple rounds of paraphrasing are chained together, and the seed utterances for a round come from a previous round by using different seed selection strategies (e.g., simply choosing all paraphrases from the previous round (Negri et al. 2012), random sampling (Jiang, Kummerfeld, and Lasecki 2017), or identifying outliers (Larson et al. 2019)). The focus of these strategies is to ultimately reduce the bias effect of factors like the seed utterances and examples shown to workers (Wang et al. 2012). Diversity can be further improved by focusing on the actual crowdsourcing task. This task could constraint the crowd from using frequently-used words (Larson et al. 2020) or suggest words that workers may incorporate in their paraphrases (Yaghoub-Zadeh-Fard et al. 2020a). While valuable, these contributions assume workers would naturally produce diverse paraphrases or focus primarily on lexical variations.

In this paper we describe our preliminary work towards a multi-stage paraphrasing pipeline that can guide the crowdsourcing process towards producing paraphrases that are syntactically diverse and balanced.

Crowdsourcing Diverse Paraphrases

Figure 1 depicts our approach and where it sits in an iterative and multi-stage pipeline for crowd-based paraphrasing based on prior art (Negri et al. 2012; Kang et al. 2018; Larson et al. 2019). In this pipeline, a typical round r of data collection (black arrows) takes as input a dataset of seeds utterances X and a curated collection of paraphrases Y (initially, Y can be empty). The crowdsourcing task in the *paraphrase generation* step asks a worker to provide a set of n paraphrases y_j for an utterance x. The resulting collection of unverified paraphrases $\bar{Y}$ is fed to the *paraphrase validation* step, where another crowd helps to check for correctness. The correct paraphrases are then appended to the collection of curated paraphrases Y. The *seed selection* step updates (or fully replaces) the seeds in X by sampling from the correct paraphrases to create the set of seeds for the next round.

Our approach assumes an initial (X, Y) as input and aims to steer the crowd towards specific *patterns* or encourage workers to contribute novel syntactic variations to the input dataset. For these goals, we introduce a pattern selection step and propose novel prompts for paraphrase generation.

Pattern selection. To capture and control syntax, we follow (Iyyer et al. 2018) and define a pattern as the top two levels of a constituency parse tree (this depth mostly has clause/phrase level nodes, making syntax comparisons less strict but still effective). The pattern selection step thus analyzes the paraphrases in Y and identifies *target patterns* to support the paraphrase generation step towards these goals.

How to identify target patterns? For example, we may choose the k least-frequent patterns in Y as targets, or the

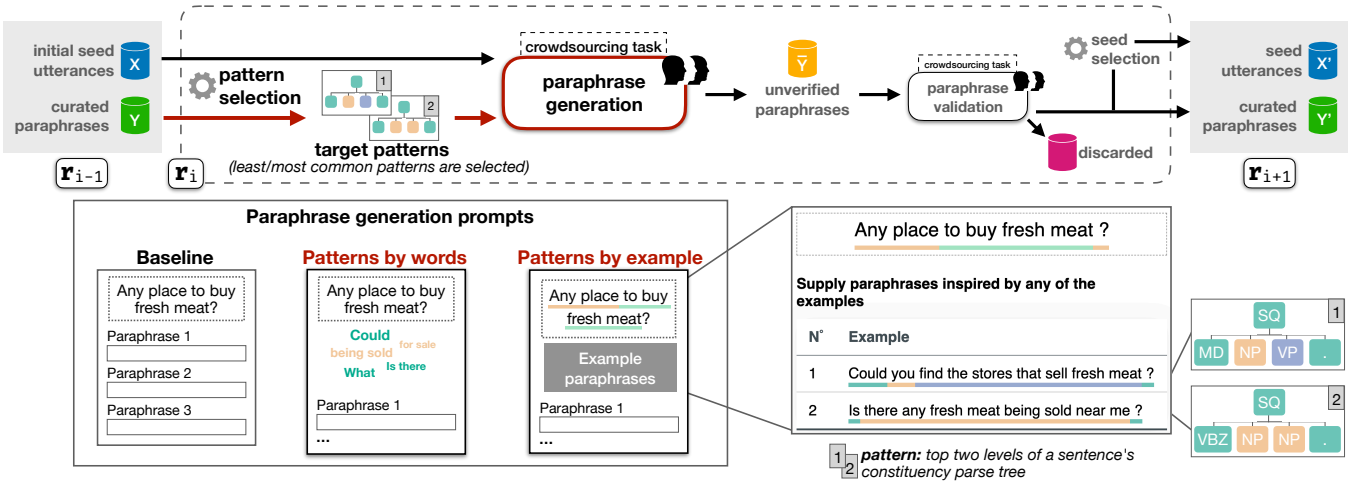

Figure 1: Our approach (in red) sits in a pipeline for paraphrasing. *Pattern selection* identifies target patterns (capturing syntax). *Paraphrase generation* leverages these patterns to craft prompts aiming for specific syntax or to elicit novel syntactic variations.

k most-frequent ones, any choice informing the generation step differently. Bottom-k patterns may be used to guide workers to provide paraphrases matching any target pattern to collectively contribute more balanced syntax. While the top-k may be used as "taboo" to avoid frequent syntax.

Paraphrase generation. Prompts can easily include words and ask workers to avoid/incorporate them in their paraphrases. This is not straightforward for patterns, as patterns directly are not informative for non-experts.

How can the prompts leverage target patterns to steer towards (or encourage novel) syntax? To achieve these goals, this work proposes prompts that 1) impose constraints or 2) give recommendations, both by showing example words or paraphrases sampled from target patterns. The specific goal shapes the prompt and what the target patterns represent. For example, if we want to steer the crowd towards uncommon syntax, we can set target patterns to the least frequent patterns in Y. The prompts can then show example paraphrases sampled from these patterns and ask workers to contribute paraphrases matching a pattern in any example (i.e., constraining workers to a specific syntax and compensating for less frequent syntax in Y). Alternatively, we may aim for novel syntax, so the prompts may use example paraphrases/words as recommendations to inspire workers and encourage them to contribute novel (or "unseen") patterns.

Ongoing Experiments

We are running experiments to explore (i) whether our approach can effectively increase syntactic diversity, and (ii) what task designs are more effective for this goal. Below, we overview of our planned experiments.

Datasets. We selected the ParaQuality dataset (Yaghoub-Zadeh-Fard et al. 2019), which contains seed utterances for intents from different domains, including those for Scopus, Spotify, Open Weather, Gmail among other services.

Experimental conditions[1]. We consider six task designs, each representing different prompts. All prompts share the same basic set of instructions. The ❶ *baseline* prompt simply queries for paraphrases for the given seed. Variations of the baseline include ❷ *word recommendations* (Yaghoub-Zadeh-Fard et al. 2020a) ❸ and *taboo words* (Larson et al. 2020). Our approach ❹ *patterns by example* shows example paraphrases associated with least-frequent patterns and asks workers to use them as inspiration (allowing novel syntax). A variant of this prompt constraints workers to use only patterns present in the examples. The ❺ *taboo patterns* asks for paraphrases with a pattern different than the given example paraphrases (sampled from most-frequent patterns). We also propose ❻ *patterns by words* to show words (sampled from least-frequent patterns) and request workers to use them in their paraphrases. A variant fixes the position of the words and asks workers to fill in the blanks. Informed by pilots, all conditions include validators to avoid paraphrases that are (clearly) incorrect: (i) check that they are not copies of the examples and are unique after preprocessing (e.g., lemmatizing), and (ii) avoid gibberish, as in (Liu and Liu 2019).

Procedure. We conduct two full rounds of the pipeline in Figure 1, running all conditions. Pattern selection simply counts the frequency of unique patterns using exact matching, and we adopt the approach in (Larson et al. 2019) for paraphrase validation. Seeds selection is based on random sampling of correct paraphrases. For the first round (r_1), we use the seeds and correct paraphrases from ParaQuality as input. We recruit English-speaking workers ranked top-20% in Toloka and collect paraphrases from 10 workers per seed.

Metrics. We consider commonly-used paraphrase diversity metrics: Type-Token Ratio (TTR), Paraphrase In N-gram changes (PINC) (Chen and Dolan 2011), and DIV (Kang et al. 2018). We also consider a measure of pattern diversity based on Jiang, Kummerfeld, and Lasecki (2017): the number of distinct patterns divided by the total number of paraphrases. Following Yaghoub-Zadeh-Fard et al. (2020a), we also measure the accuracy of an intent detection model trained on the datasets resulting from each condition.

Discussion. We have implemented the pipeline and prompts, informed by pilots, and are ready to start the experiments.

[1] Screenshots and details at https://tinyurl.com/hcomp2021div

References

Berro, A.; Fard, M. Y. Z.; Baez, M.; Benatallah, B.; and Benabdeslem, K. 2021. An Extensible and Reusable Pipeline for Automated Utterance Paraphrases. *Proc. VLDB Endow.* 14(12): 2839–2842. URL http://www.vldb.org/pvldb/vol14/p2839-berro.pdf.

Chen, D. L.; and Dolan, W. B. 2011. Collecting Highly Parallel Data for Paraphrase Evaluation. In Lin, D.; Matsumoto, Y.; and Mihalcea, R., eds., *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*, 190–200. The Association for Computer Linguistics. URL https://www.aclweb.org/anthology/P11-1020/.

Iyyer, M.; Wieting, J.; Gimpel, K.; and Zettlemoyer, L. 2018. Adversarial Example Generation with Syntactically Controlled Paraphrase Networks. In Walker, M. A.; Ji, H.; and Stent, A., eds., *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, 1875–1885. Association for Computational Linguistics. doi:10.18653/v1/n18-1170. URL https://doi.org/10.18653/v1/n18-1170.

Jiang, Y.; Kummerfeld, J. K.; and Lasecki, W. S. 2017. Understanding Task Design Trade-offs in Crowdsourced Paraphrase Collection. In Barzilay, R.; and Kan, M., eds., *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, 103–109. Association for Computational Linguistics. doi:10.18653/v1/P17-2017. URL https://doi.org/10.18653/v1/P17-2017.

Kang, Y.; Zhang, Y.; Kummerfeld, J. K.; Tang, L.; and Mars, J. 2018. Data Collection for Dialogue System: A Startup Perspective. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*. New Orleans - Louisiana: Association for Computational Linguistics. doi:10.18653/v1/N18-3005. URL https://www.aclweb.org/anthology/N18-3005.

Larson, S.; Mahendran, A.; Lee, A.; Kummerfeld, J. K.; Hill, P.; Laurenzano, M. A.; Hauswald, J.; Tang, L.; and Mars, J. 2019. Outlier Detection for Improved Data Quality and Diversity in Dialog Systems. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 517–527. Association for Computational Linguistics. doi:10.18653/v1/n19-1051. URL https://doi.org/10.18653/v1/n19-1051.

Larson, S.; Zheng, A.; Mahendran, A.; Tekriwal, R.; Cheung, A.; Guldan, E.; Leach, K.; and Kummerfeld, J. K. 2020. Iterative Feature Mining for Constraint-Based Data Collection to Increase Data Diversity and Model Robustness. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, 8097–8106. Association for Computational Linguistics. doi:10.18653/v1/2020.emnlp-main.650. URL https://doi.org/10.18653/v1/2020.emnlp-main.650.

Liu, P.; and Liu, T. 2019. Optimizing the Design and Cost for Crowdsourced Conversational Utterances. In *KDD-DCCL*.

Negri, M.; Mehdad, Y.; Marchetti, A.; Giampiccolo, D.; and Bentivogli, L. 2012. Chinese Whispers: Cooperative Paraphrase Acquisition. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, 2659–2665. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2012/summaries/772.html.

Thompson, B.; and Post, M. 2020. Paraphrase generation as zero-shot multilingual translation: Disentangling semantic similarity from lexical and syntactic diversity. *arXiv preprint arXiv:2008.04935* .

Wang, W. Y.; Bohus, D.; Kamar, E.; and Horvitz, E. 2012. Crowdsourcing the acquisition of natural language corpora: Methods and observations. In *2012 IEEE Spoken Language Technology Workshop (SLT), Miami, FL, USA, December 2-5, 2012*, 73–78. IEEE. doi:10.1109/SLT.2012.6424200. URL https://doi.org/10.1109/SLT.2012.6424200.

Yaghoub-Zadeh-Fard, M.; Benatallah, B.; Barukh, M. C.; and Zamanirad, S. 2019. A Study of Incorrect Paraphrases in Crowdsourced User Utterances. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 295–306. Association for Computational Linguistics. doi:10.18653/v1/n19-1026. URL https://doi.org/10.18653/v1/n19-1026.

Yaghoub-Zadeh-Fard, M.; Benatallah, B.; Casati, F.; Barukh, M. C.; and Zamanirad, S. 2020a. Dynamic word recommendation to obtain diverse crowdsourced paraphrases of user utterances. In *IUI '20: 25th International Conference on Intelligent User Interfaces, Cagliari, Italy, March 17-20, 2020*, 55–66. ACM. doi:10.1145/3377325.3377486. URL https://doi.org/10.1145/3377325.3377486.

Yaghoub-Zadeh-Fard, M.; Benatallah, B.; Casati, F.; Barukh, M. C.; and Zamanirad, S. 2020b. User Utterance Acquisition for Training Task-Oriented Bots: A Review of Challenges, Techniques and Opportunities. *IEEE Internet Comput.* 24(3): 30–38. doi:10.1109/MIC.2020.2978157. URL https://doi.org/10.1109/MIC.2020.2978157.