A Benchmark for Recipe Understanding in Autonomous Agents

Author: Robin De Haes
Supervisors: Prof. Dr. Paul Van Eecke and Dr. Jens Nevens

Vrije Universiteit Brussel, Belgium, Artificial Intelligence Lab
robindehaes@msn.com

Procedural text is text composed of a sequence of instructions that describe how to perform a specific task [7, 6]. Understanding such text can be useful from both a scientific and a practical viewpoint. From a scientific viewpoint, procedural text understanding requires modeling and reasoning with a dynamically changing world and can therefore be seen as a representative task for demonstrating an agent's ability to understand the impact of actions and events [4, 8]. From a practical viewpoint, having agents that understand procedural text could progress the development of robotic workers that can execute a variety of manual tasks [11]. To facilitate and measure progress on such natural language understanding of the procedural type, this thesis proposes the MUHAI Recipe Execution Benchmark. This benchmark consists of a corpus of 30 recipes, that should be mapped to a graph-based machine-readable representation language, and a standardized evaluation procedure using a kitchen simulator with metrics that are specifically designed for measuring procedural execution success.

Recipes have been chosen as the data source because they are an inherently challenging type of procedural text. They generally require preciseness, contextual reasoning and the application of commonsense knowledge, because anaphora, ellipses, hyponyms and meronyms are abundantly present in recipe texts [10, 3]. When a recipe for example requires egg whites to be beaten, it will often omit that this involves cracking the eggs, separating the egg whites from the yolks and grabbing a whisk to perform the beating action. Moreover, the same linguistic expression of 'eggs' might even semantically refer to whole eggs in their shell at first while later it might be used to refer to the bowl of beaten egg whites that has been obtained.

However, having a sufficiently challenging data corpus is not enough to be a suitable testbed for measuring and advancing progress. In fact, prior research efforts have already delivered such corpora [12, 13, 5, 9]. Therefore, the main scientific contribution of the MUHAI Recipe Execution Benchmark is the standardized evaluation suite accompanying the corpus. To perform model evaluation, the benchmark comes with a cross-platform symbolic simulator that can read, execute and evaluate mapped recipes. During this evaluation, performance is measured via existing non-simulation-based metrics in combination with novel simulation-based metrics. This allows for multiperspective performance estimates which maximizes transferability to real-world utility.

Smatch [2] is an existing semantic graph matching metric that is included because of its high adoption rate and general familiarity within the community.

However, it has the downside of weighing all mistakes in the graph equally while the severity of mistakes in real-world cooking can vary a lot. Moreover, it does not take into account that the execution of certain actions can be permuted without it having an effect on reaching the end goal. Therefore, the benchmark also includes novel simulation-based metrics that specifically aim to gauge to what extent the end goal of execution, i.e., obtaining the final dish, is reached.

The simulation-based metric 'goal-condition success' represents the ratio of goal-conditions that were reached to the number of goal-conditions that are required to reach the end goal. It thus approximates how many steps an agent is removed from being able to create the final dish. The goal-conditions have been established for each recipe in the corpus beforehand and can be summarized as minimal simulation states that should be traversed at some point during execution. An intermediate goal-condition for a chocolate-frosted cookie would for example be the presence of a container with frosting, since an agent will never be able to create a chocolate-frosted cookie if it did not create the frosting at some point.

The simulation-based metric 'dish approximation score' represents how well the dish that is prepared by an agent approximates a gold standard version of the dish. It is a weighted similarity measure that is based on the properties of both dishes in simulation. These properties can be quantitative, e.g., their weight, or qualitative, e.g., the presence of sugar. The weight or importance of each property in the score computation is based on how significantly the property would impact the taste of the dish.

Finally, prior research has shown that community adoption of a benchmark is as integral to the general utility of a benchmark as its quality is [1]. Therefore, the MUHAI Recipe Execution Benchmark is made publicly available online, accompanied by extensive documentation and open-source code. This does not only promote usage but also promotes extensions by other researchers. Whether this will lead to community adoption and significant advances in natural language understanding can only be determined in the future. Nevertheless, its transparent and execution-oriented design makes the benchmark useful for the development of new recipe understanding approaches and as a source of inspiration for new benchmarks.

Resources The benchmark with all its documentation and source code can be found at https://ehai.ai.vub.ac.be/recipe-execution-benchmark.

References

1. Barbosa-Silva, A., Ott, S., Blagec, K., Brauner, J., Samwald, M.: Mapping global dynamics of benchmark creation and saturation in artificial intelligence. Nature Communications **13**, 1–11 (11 2022). https://doi.org/10.1038/s41467-022-34591-0
2. Cai, S., Knight, K.: Smatch: an Evaluation Metric for Semantic Feature Structures. In: Schuetze, H., Fung, P., Poesio, M. (eds.) Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. pp. 748–752. Association for Computational Linguistics (ACL), Stroudsburg, PA, USA (08 2013)

3. Fang, B., Baldwin, T., Verspoor, K.: What does it take to bake a cake? The RecipeRef corpus and anaphora resolution in procedural text. In: Muresan, S., Nakov, P., Villavicencio, A. (eds.) Findings of the Association for Computational Linguistics: ACL 2022. pp. 3481–3495. Association for Computational Linguistics (ACL), Stroudsburg, PA, USA (05 2022). https://doi.org/10.18653/v1/2022.findings-acl.275
4. Henaff, M., Weston, J., Szlam, A., Bordes, A., LeCun, Y.: Tracking the World State with Recurrent Entity Networks. arXiv preprint arXiv:1612.03969 (05 2017). https://doi.org/10.48550/arXiv.1612.03969
5. Jiang, Y., Zaporojets, K., Deleu, J., Demeester, T., Develder, C.: Recipe instruction semantics corpus (RISeC): Resolving semantic structure and zero anaphora in recipes. In: Wong, K., Knight, K., Wu, H. (eds.) Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing. pp. 821–826. Association for Computational Linguistics (ACL), Stroudsburg, PA, USA (12 2020)
6. Kiddon, C., Ponnuraj, G.T., Zettlemoyer, L., Choi, Y.: Mise en Place: Unsupervised Interpretation of Instructional Recipes. In: Màrquez, L., Callison-Burch, C., Su, J., Pighin, D., Marton, Y. (eds.) Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 982–992. Association for Computational Linguistics (ACL), Stroudsburg, PA, USA (09 2015). https://doi.org/10.18653/v1/D15-1114
7. Maeta, H., Sasada, T., Mori, S.: A Framework for Procedural Text Understanding. In: Proceedings of the 14th International Conference on Parsing Technologies. pp. 50–60. Association for Computational Linguistics (ACL), Stroudsburg, PA, USA (07 2015). https://doi.org/10.18653/v1/W15-2206
8. Mishra, B.D., Huang, L., Tandon, N., Yih, W., Clark, P.: Tracking State Changes in Procedural Text: a Challenge Dataset and Models for Process Paragraph Comprehension. In: Walker, M.A., Ji, H., Stent, A. (eds.) Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 1595–1604. Association for Computational Linguistics (ACL), Stroudsburg, PA, USA (06 2018). https://doi.org/10.18653/v1/n18-1144
9. Mori, S., Maeta, H., Yamakata, Y., Sasada, T.: Flow Graph Corpus from Recipe Texts. In: Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the 9th International Conference on Language Resources and Evaluation. pp. 2370–2377. European Language Resources Association (ELRA), Paris, France (05 2014)
10. Nanba, H., Doi, Y., Tsujita, M., Takezawa, T., Sumiya, K.: Construction of a Cooking Ontology from Cooking Recipes and Patents. In: Brush, A.J., Friday, A., Kientz, J.A., Scott, J., Song, J. (eds.) Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing. pp. 507–516. Association for Computing Machinery (ACM), New York, NY, USA (09 2014). https://doi.org/10.1145/2638728.2641328
11. Ribeiro, J., Lima, R., Eckhardt, T., Paiva, S.: Robotic Process Automation and Artificial Intelligence in Industry 4.0 - A Literature review. Procedia Computer Science **181**, 51–58 (01 2021). https://doi.org/10.1016/j.procs.2021.01.104
12. Tasse, D., Smith, N.A.: SOUR CREAM: Toward Semantic Processing of Recipes. Tech. Rep. CMU-LTI-08-005, Carnegie Mellon University, Pittsburgh, PA, USA (2008)

13. Yamakata, Y., Mori, S., Carroll, J.A.: English Recipe Flow Graph Corpus. In: Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the 12th International Conference on Language Resources and Evaluation. pp. 5187–5194. European Language Resources Association (ELRA), Paris, France (05 2020)