Predicting Orderliness of Articles in wikiHow

Aditya Kashyap* **Arun Kirubarajan*** **Li Zhang***
University of Pennsylvania
{kashyap, kiruba, zharry}@seas.upenn.edu

Abstract

wikiHow is a website that offers over 250,000+ instructional how-to articles containing rich world knowledge. One important type of such knowledge is the temporal ordering of events, which is widely present in some but not all wikiHow contents. Concretely, there are roughly two types of wikiHow articles: ordered step-by-step instructions and unordered parallel suggestions. To facilitate related research, we offer labels of these two types for over 210,000 wikiHow articles with an estimated precision of around 90%, by fine-tuning a state-of-the-art language model using 1,000 hand-annotated examples. We also provide ablation studies for insights into what portions of the texts the model considers making these predictions.

1 Introduction

wikiHow[1] is a website consisting of more than 210,000 professionally-edited how-to articles spanning a tremendous amount of domains, existing in 18 languages. Examples of articles include mundane common-sense such as "How to Save Water" and niche expertise such as "How to Crochet a Teddy Bear". Each how-to article includes a variety of instructional information of different modalities, including textual description of steps to accomplish a goal, corresponding images, instructional videos, and so on.

As its texts alone contain rich world knowledge (Pareti et al., 2014), we claim that wikiHow can be a useful corpus for a variety of tasks, especially temporal event ordering, a classical and important task (Pustejovsky et al.; Chambers et al., 2007; Monroe et al., 2013; Ning et al., 2018a,b). However, most existing temporal event datasets suffers from two major shortcomings. First, they are mostly on a verb level, represented by the order of occurrence of verb pairs in a corpus. This forestalls the learning of ordering of events represented by phrases or sentences: for example, while "prepare" often happens before "execute", "prepare the clean-up" would logically happen after "execute the detonation". Second, most datasets use count of occurrences or probability as the label, which introduce much noise that prevents the resource from being used as a reliable benchmark. We claim that an event temporal ordering dataset included from wikiHow can address these two problems effectively, since wikiHow includes instructional manuals that are intended to be followed in a given order.

However, an immediate obstacle is that not *all* wikiHow articles are ordered. Based on our preliminary observation of 100 random wikiHow articles, about 40% can be categorized as **ordered** step-by-step instructions (e.g. recipes, manuals for woodwork and gardening, etc.), whereas 60% can be categorized as **unordered** parallel suggestions (e.g. ways to be attractive, confident, etc.). We also find that most adjacent step pairs in the so-called ordered articles mostly should indeed be ordered, while the converse in the unordered articles. This observation suggests the possibility of learning a simple model to predict the **orderliness** of articles in a binary classification format, which correspond to the orderliness of steps within with high probability. We show that we are able to learn such a model with 88% cross-validation precision using only 1,000 hand-annotated examples[2]. Further, we show using ablation studies that consider-

* Equal contribution.
[1] www.wikihow.com

[2] We release our annotations and model predictions here: https://drive.google.com/file/d/1miFUCvLKF7jeeqwR0_bWp-rNaJIGgy9_/view.

ing only the first word of the steps is enough for the model to achieve similar precision.

[3]. Further, we show using ablation studies that considering only the first word of the steps is enough for the model to achieve similar precision.

This paper outlines two key contributions: 1) we offer labels of orderliness for with an estimated precision of around 90% by fine-tuning a state-of-the-art model and 2) we offer ablation studies to interpret the model's behavior.

2 Formulation

We reduce the task of predicting wikiHow article orderliness to binary classification since the discrete classes of **ordered** and **unordered** articles present a binary selection process. To distinguish between the two classes, we make the assumption that if 50% of the step-wise event pairs exhibit some direct ordering, we consider the article to be **ordered**. This is because the majority of the article's steps necessitate an order into to ensure the succesful completion of the article's goal. Otherwise, we consider the article to be comprised of **unordered** steps, since it implies most of the article can be executed in parallel without compromising the goal of the article.

The choice to frame the problem as binary classification allows for the use of previous work in the modelling process. In particular, the reduction allows us to fine-tune pre-existing state-of-the-art language models, which we have shown to identify lexical clues and understand the semantic knowledge of given natural language wikiHow steps. This is evidenced through our ablation studies in Section 4.

3 Annotation

Our dataset is comprised of three key fields for wikiHow articles: the title of the article (most often the goal), the natural language steps of the procedure, and the annotated label for whether the article's steps were ordered or unordered. Out of the 1000 total annotated examples, we observe that 414 articles were identified to have ordered steps whereas 586 articles presented little to no ordering between steps.

Annotations were collected by presenting each annotator a set of article titles and step descriptions, where annotators could classify each instance as an ordered or unordered article. The annotators were comprised of individuals with varying experience in natural language processing, including the authors of this paper. However, due to the sporadic themes and goals of randomly sampled articles, little to no external knowledge was used to ground the annotation process. Although

Goal	Steps	Label
How to Make Sweet Curry Powder	If using fresh spices, grind or mill them in readiness. Mix the spices together in a small mixing bowl. Pour into a suitable airtight container for storage. Label and store.	Ordered
How to Find Perfect Ice Skates	Consider what kind of skating you do. Once that is done, get some skates. Go to a sports store like, Models, Sports Authority, or Dick's Sporting Goods. Go to the ice skate area and look at all of the skates. Choose 3 different types of skates in your size.	Ordered
How to Make Him Want You	Act confident. Be kind. Be open to new things. Be positive. Give him his space. Keep your conversations interesting.	Unordered
How to Deal With a Friend That You Lost	Accept it will take time. Find new ways to fill the space. Learn from the experience. Reach out to new people.	Unordered

Table 1: Examples of ordered and unordered article training examples from wikiHow.

[3]We release our fine-tuning code on GitHub: https://github.com/kirubarajan/wikiHow_orderliness.

this may have contributed to noisy labels in unknown or obscure domains, this overall ensured that semantics alone would be able to provide the insight required for annotation and prediction. We first sampled 30 articles to compute a preliminary Inter-Annotator Agreement (IAA) score among 6 annotators. We observed that of the 30 articles, annotations only differed for 3 articles, yielding an IAA score of 90%.

Of the 1000 articles, we split 700 instances as a training set and 100 instances as a development set. We used a held-out test set of 40 annotated examples to evaluate the model against, which we repeated five times. The test set was not annotated by the paper's authors in order to mitigate any source of bias.

4 Modelling and Evaluation

The RoBERTa language model (Liu et al., 2019) with an appended binary classification layer was fine-tuned on the wikiHow dataset. All the input text was first converted to lowercase and then tokenized using a WordPiece tokenizer. A learning rate of $5 * 10^{-5}$ was chosen and the model was trained for 3 epochs. 5-fold cross validation was performed across the 1,000 annotated training examples, with a precision of 0.87, recall of 0.81, and F1-score of 0.84. The cross validation performances given variable amount of training examples are shown in Figure 1.

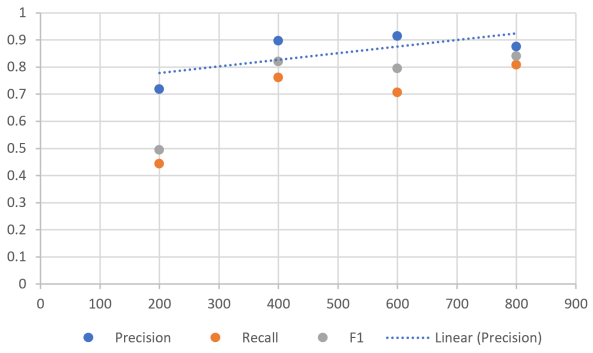

Figure 1: Performances of orderliness prediction given variable amount of training examples.

4.1 Ablation

In order to understand the amount of signal the classification model obtained from different parts of the wikiHow article, several ablation experiments were carried out. For each of the following points, the entire article except what is listed was masked out.

1. The first step of each wikiHow article.
2. The first word of each step for every wikiHow article.
3. Nouns in the title and the steps of the wikiHow article.
4. Verbs in the title and the steps of the wikiHow article.
5. The full text of steps of each wikiHow article.
6. The full text of the title of each wikiHow article.

The results for these experiments are shown in Table 2.

Experiment	Precision	Recall	F1-Score
First Step	0.87	0.48	0.62
First word of each step	0.92	0.68	0.78
Nouns	0.84	0.68	0.75
Verbs	0.81	0.80	0.80
Only Steps (No Title)	0.89	0.83	0.86
Only Title (No Steps)	0.88	0.68	0.77
Entire Article	0.87	0.81	0.84

Table 2: The precision, recall and F1 scores of the various models in predicting whether an article is ordered or not.

5 Discussion

The fine-tuned model achieves an impressive cross-validation precision of 0.87 in retrieving ordered articles, with merely 800 labeled training examples. The high performance can likely be attributed to the bias of wikiHow's text style, where lexical usage corresponds strongly to the orderliness of the articles. As BERT models are known to be able to exploit such bias (Poliak et al., 2018; Si et al., 2019; Zellers et al., 2018), it is not surprising that the task is solved to an satisfactory extent. Nevertheless, the result is encouraging, as the model can be applied to the entire 210,000 wikiHow articles with high precision. We release the labels of orderliness of these articles, hoping to facilitate research in event temporal ordering.

As for the ablation studies, the model presented with the entire wikiHow article performed as well (F1 score of 0.84) as the model presented with only the article steps (F1 score of 0.86). This goes to show that given the steps of a wikiHow article, the corresponding title becomes redundant in predicting whether the article is ordered or not. The results for the ablation study using only titles of articles (f1 score of 0.77) suggest that the topic contained in the title of the article is often a great indicator of whether the steps are ordered/unordered. Therefore, looking at the results of the ablation study "Only Nouns" (containing only common nouns and proper nouns present in the articles) shows that the topic discussed in the article (cooking, gardening vs self-improvement) frequently provide clues on the temporal sequence of its steps. Similarly, the results of the ablation study "Only Verbs" indicates that the sequence of verbs in an article (e.g cleaning, washing, cutting, cooking ,etc. vs improve, show, try, etc.) carry information used by the model. Looking at only the first step of the article is often not sufficient (f1 score of 0.62) in predicting whether the respective article is ordered. However, looking at the first word of each step in the article can help the model often reach the right decision (f1 score of 0.78) given that most steps start with a verb.

Future work for this research involves understanding the logical dynamics of steps in a given article. For this paper, we made the simplifying assumption that if over 50 percent of an article's steps are not ordered, then the article is considered to be unordered. However, there can exist certain exceptions to this rule. For example, some articles may provide non-linearity in the steps (i.e. the steps are not ordered) but all steps are necessary to complete the task. This sort of logical nuance isn't explored due to the formulation's simplifying assumption. One can posit that learning these underlying structures in step ordering for an article would better help facilitated research in automatic procedure prediction. This can be explored by identifying and labelling articles with their type of logical structure to complete the specified goal.

6 Conclusion

We offer labels of article orderliness with high precision for over 210,000 wikiHow, by fine-tuning a state-of-the-art pre-trained model on only 1,000 hand-annotated examples. Using ablation studies, we also show that the model can learn most information needed using the first word in each step, likely due to the idiosyncratic style of wikiHow articles. We hope these labels and observations would facilitate research in event temporal ordering and schema learning.

References

Nathanael Chambers, Shan Wang, and Dan Jurafsky. 2007. Classifying temporal relations between events. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 173–176.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Megan Monroe, Rongjian Lan, Hanseung Lee, Catherine Plaisant, and Ben Shneiderman. 2013. Temporal event sequence simplification. *IEEE transactions on visualization and computer graphics*, 19(12):2227–2236.

Qiang Ning, Hao Wu, Haoruo Peng, and Dan Roth. 2018a. Improving Temporal Relation Extraction with a Globally Acquired Statistical Resource. In *Proc. of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 841–851, New Orleans, Louisiana. Association for Computational Linguistics.

Qiang Ning, Hao Wu, and Dan Roth. 2018b. A multi-axis annotation scheme for event temporal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1318–1328, Melbourne, Australia. Association for Computational Linguistics.

Paolo Pareti, Benoit Testu, Ryutaro Ichise, Ewan Klein, and Adam Barker. 2014. Integrating know-how into the linked data cloud. In *International Conference on Knowledge Engineering and Knowledge Management*, pages 385–396. Springer.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, et al. The timebank corpus.

Chenglei Si, Shuohang Wang, Min-Yen Kan, and Jing Jiang. 2019. What does bert learn from multiple-choice reading comprehension datasets? *ArXiv*, abs/1910.12391.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.