Crowdsourcing Natural Language Data at Scale: A Hands-On Tutorial

Alexey Drutsa
Yandex
Moscow, Russia

Dmitry Ustalov
Yandex
Saint Petersburg, Russia

Valentina Fedorova
Yandex
Moscow, Russia

{adrutsa,dustalov,valya17}@yandex-team.ru

Olga Megorskaya
Yandex
Saint Petersburg, Russia
omegorskaya@yandex-team.ru

Daria Baidakova
Yandex
Moscow, Russia
dbaidakova@yandex-team.ru

Abstract

In this tutorial, we present a portion of unique industry experience in efficient natural language data annotation via crowdsourcing shared by both leading researchers and engineers from Yandex. We will make an introduction to data labeling via public crowdsourcing marketplaces and will present the key components of efficient label collection. This will be followed by a practical session, where participants address a real-world language resource production task, experiment with selecting settings for the labeling process, and launch their label collection project on one of the largest crowdsourcing marketplaces. The projects will be run on real crowds within the tutorial session and we will present useful quality control techniques and provide the attendees with an opportunity to discuss their own annotation ideas.

Tutorial Type: Introductory

1 Description

Training and evaluating modern Natural Language Processing (NLP) models require large-scale multilingual language resources of high quality. Traditionally, such resources have been created by groups of experts or by using automated silver standards. Crowdsourcing has become a popular approach for data labeling that allows annotating language resources in a shorter time and at a lower cost than the experts while maintaining expert-level result quality. Examples include Search Relevance Evaluation, Machine Translation, Question Answering, Corpus Annotation, etc. However, for running crowdsourcing successfully, it is essential to pay attention to task design, quality control, and annotator incentives. This tutorial aims to teach attendees how to efficiently use crowdsourcing for annotating language resources on a large scale. The tutorial is composed of

1. a theoretical part aimed at explaining the methodology for labeling process in crowdsourcing and main algorithms required to obtain high-quality data (including aggregation, incremental relabeling, and quality-dependent pricing), and

2. practice sessions for setting up and running language resource annotation project on one of the largest public crowdsourcing marketplaces.

The goals of our tutorial are to explain the fundamental techniques for aggregation, incremental relabeling, and pricing in connection to each other and to teach attendees the main principles for setting up an efficient process of language resource annotation on a crowdsourcing marketplace. We will share our best practices and allow the attendees to discuss their issues with language data labeling with crowdsourcing.

To establish trust and allow the attendees to evaluate the crowdsourced results even after the tutorial carefully, we decide to use English as the language of our tutorial datasets. According to our six years of experience, we would emphasize that the same techniques can successfully apply to virtually any language and domain that the crowd performers command, including Russian, Turkish, Vietnamese, and many other languages. The opportunity to attract crowd performers from under-represented languages, backgrounds, and demographics brings the possibility to create more useful language resources and evaluate NLP systems fairly in more challenging multilingual setups.

Proceedings of NAACL-HLT 2021: Tutorials, pages 25–30
June 6–11, 2021. ©2021 Association for Computational Linguistics

1.1 Introduction to Crowdsourcing

We will start with an *introduction* that includes crowdsourcing terminology and examples of tasks on crowdsourcing marketplaces. We will also demonstrate why crowdsourcing is becoming more popular in working with data on a large scale, showing successful crowdsourcing applications for language resource development, and describing current industry trends of crowdsourcing use.

1.2 Key Components for Efficient Data Collection

We will discuss thoroughly *the key components* required to collect labeled data: proper decomposition of tasks (construction of a pipeline of several small tasks instead of one large human intelligent task), easy to read and follow task instructions, easy to use task interfaces, quality control techniques, an overview of aggregation methods, and pricing.

Quality control techniques include approaches "before" task performance (selection of performers, education and exam tasks), the ones "during" task performance (golden sets, motivation of performers, tricks to remove bots and cheaters), and approaches "after" task performance (post verification/acceptance, consensus between performers).

We will share best practices, including critical aspects and pitfalls when designing instructions & interfaces for performers, vital settings in different types of templates, training, and examination for performers selection, pipelines for evaluating the labeling process. Also, we will demonstrate typical crowdsourcing pipelines used in industrial applications, including Machine Translation, Content Moderation, Named Entity Recognition, etc.

1.3 Hands-on Crowdsourcing Practice

We will conduct *a hands-on practice session*, which is the vital and the longest part of our tutorial. We will encourage the attendees to apply the techniques, and best practices learned during the first part of the tutorial. For this purpose, we propose the attendees run their own crowdsourced Spoken Language Recognition pipeline on actual crowd performers. As the *input* the attendees have audio files of variable quality in English, as the *output* they should provide high-quality transcriptions for these recordings obtained via crowdsourcing. Each attendee will be involved in brainstorming the suitable crowdsourcing pipeline for the given task and configuring and launching the annotation project online on the real crowd while optimizing quality and cost.

Since creating a project from scratch might be time-consuming, we will propose our attendees choose from the most popular pre-installed templates (text input or audio playback). We will also provide the attendees with pre-paid accounts and data sets for annotation. By the end of the practice session, the attendees will learn to construct a functional pipeline for data collection and labeling, become familiar with one of the largest crowdsourcing marketplaces, and launch projects independently.

1.4 Advanced Techniques

We will discuss *the major theoretical results*, computational techniques and ideas which improve the quality of crowdsourcing annotations, and *summarize the open research questions on the topic*.

Crowd Consensus Methods. Classical models: Majority Vote, Dawid-Skene (Dawid and Skene, 1979), GLAD (Whitehill et al., 2009), Minimax Entropy (Zhou et al., 2015). Analysis of aggregation performance and difficulties in comparing aggregation models in unsupervised setting (Sheshadri and Lease, 2013; Imamura et al., 2018). Advanced works on aggregation: combination of aggregation and learning a classifier (Raykar et al., 2010), using features of tasks and performers for aggregation (Ruvolo et al., 2013; Welinder et al., 2010; Jin et al., 2017), aggregation of crowdsourced pairwise comparisons (Chen et al., 2013) and texts (Li and Fukumoto, 2019).

Incremental Relabeling (IRL). Motivation and the problem of incremental relabeling: IRL based on Majority Vote; IRL methods with worker quality scores (Ipeirotis et al., 2014; Ertekin et al., 2012; Abraham et al., 2016); active learning (Lin et al., 2014). Connections between aggregation and IRL algorithms. Experimental results of using IRL at crowdsourcing marketplaces.

Task Pricing. Practical approaches for task pricing (Wang et al., 2013; Cheng et al., 2015; Yin et al., 2013). Theoretical background for pricing mechanisms in crowdsourcing: efficiency, stability, incentive compatibility, etc. Pricing experiments and industrial experience of using pricing at crowdsourcing platforms.

Task Design for NLP. Most crowdsourcing tasks are domain-specific (Callison-Burch and Dredze, 2010; Biemann, 2013) and designed manually, yet the task design can be made more efficient by using the generic workflow patterns (Bernstein et al., 2010; Gadiraju et al., 2019), computer-supported methods (Little et al., 2009), and crowd-supported methods (Bragg et al., 2018).

1.5 Concluding Remarks

Finally, we will finish with analyzing obtained results from the launched projects. This step demonstrates the process of verification of collected data. Together with the attendees, we will discuss which aggregation algorithms can be applied, analyze outcome label distribution, check performer quality and contribution, elaborate on budget control, detect possible anomalies and problems. We will then share practical advice, discuss pitfalls and possible solutions, ask the attendees for feedback on the learning progress, and answer final questions.

By the end of the tutorial, attendees will be familiar with

- key components required to produce language resources via crowdsourcing efficiently;
- state-of-the-art techniques to control the annotation quality and to aggregate the annotation results;
- advanced methods that allow to balance out between the quality and costs;
- practice of creating, configuring, and running data collection projects on real performers on one of the largest global crowdsourcing platforms.

2 Outline

Our tutorial includes the following sessions:

- Introduction to Crowdsourcing (15 min)
- Key Components for Efficient Data Collection (30 min)
- Practice Session I (60 min)
- Lunch Break (45 min)
- Advanced Techniques (45 min)
- Practice Session II (30 min)
- Results Evaluation and Concluding Remarks (15 min)

3 Prerequisites for the Attendees

We expect that our tutorial will address an audience with a wide range of backgrounds and interests. Thus, even a beginner, each participant will be able to practice their skills in producing language resources via a crowdsourcing marketplace (this practical part will constitute most of our tutorial timeline).

Our tutorial contains an introduction that positions the topic among related areas and gives the necessary knowledge to understand the main components of data labeling processes. Thus, the entry threshold is shallow to start learning and understanding the topic. Only minimal knowledge on collecting labels is required: no knowledge on crowdsourcing, aggregation, incremental relabeling, and pricing is needed.

We plan to share rich experiences of constructing and applying large-scale data collection pipelines while highlighting the best practices and pitfalls. As a result, any person who develops a web service or a software product based on labeled data and NLP will learn how to construct a language data annotation pipeline, obtain high-quality labels under a limited budget, and avoid common pitfalls.

4 Reading List

We offer an optional reading list for the tutorial attendees. These references allow one to understand crowdsourcing annotation basics for maximizing the learning outcomes from our hands-on tutorial. We will nevertheless cover these materials during the workshop.

Quality Control. Dawid and Skene (1979); Li and Fukumoto (2019)

Task Design for NLP. Bernstein et al. (2010); Callison-Burch and Dredze (2010); Biemann (2013)

Incentives. Snow et al. (2008); Wang et al. (2013)

5 Tutorial Presenters

Alexey Drutsa (PhD), Yandex

Alexey is responsible for data-driven decisions and the ecosystem of Toloka, the open global crowd platform. His research interests are focused on Machine Learning, Data Analysis, Auction Theory; his research is published at ICML, NeurIPS, WSDM, WWW, KDD, SIGIR, CIKM, and TWEB.

Alexey is a co-author of three tutorials on practical A/B testing (at KDD ’18, WWW ’18, and SIGIR ’19), five hands-on tutorials on efficient crowdsourcing (at KDD ’19, WSDM ’20, SIGMOD ’20, CVPR ’20, and WWW ’21), and a co-organizer of the crowdsourcing workshop at NeurIPS2020. He served as a senior PC member at WWW ’19 and as a PC member at several NeurIPS, ICML, ICLR, KDD, WSDM, CIKM, and WWW conferences; he was also a session chair at WWW ’17. He graduated from Lomonosov Moscow State University (Faculty of Mechanics and Mathematics) in 2008 and received his PhD in Computational Mathematics from the same university in 2011.

🔗 `https://research.yandex.com/people/603399`

✉ `mailto:adrutsa@yandex-team.ru`

Dmitry Ustalov (PhD), Yandex

Dmitry is responsible for crowdsourcing studies and product metrics at Toloka. His research, focused on Natural Language Processing and Crowdsourcing, has been published at COLI, ACL, EACL, EMNLP, and LREC. He has been co-organizing the TextGraphs workshop at EMNLP, COLING, and NAACL-HLT since 2019 and the crowdsourcing workshops at NeurIPS and VLDB since 2020. Dmitry teaches quality control in the crowdsourcing course at the Yandex School of Data Analysis and Computer Science Center. He was also a co-author of the crowdsourcing tutorials at WWW ’21, SIGMOD ’20, and WSDM ’20. Dmitry received a bachelor’s and master’s degrees from the Ural Federal University (Russia), PhD in Computer Science from the South Ural State University (Russia), and post-doctoral training from the University of Mannheim (Germany).

🔗 `https://scholar.google.com/citations?user=wPD4g7AAAAAJ`

✉ `mailto:dustalov@yandex-team.ru`

Valentina Fedorova (PhD), Yandex

Valentina is a research analyst at the Crowdsourcing Department of Yandex. She works on research in Crowdsourcing, including aggregation models and algorithms for incremental labeling. Her research has been presented at ICML, NIPS, KDD, SIGIR, and WSDM. She is a co-author of tutorials on crowdsourcing at SIGMOD ’20, WSDM ’20, and KDD ’19. Valentina graduated from Lomonosov Moscow State University (Faculty of Applied Mathematics and Computer Science) and obtained her PhD in Machine Learning from Royal Holloway University of London in 2014. She is reading lectures on response aggregation and IRL for the crowdsourcing course at the Yandex School of Data Analysis (Moscow, Russia) and Computer Science Center (Saint Petersburg, Russia).

🔗 `https://research.yandex.com/people/603772`

✉ `mailto:valya17@yandex-team.ru`

Olga Megorskaya, Yandex

Olga Megorskaya, CEO of Toloka. Under Olga’s leadership, Toloka platform has grown the number of crowd performers involved in data labeling from several dozen in 2009 up to 4.1 million in 2020 and became a global infrastructure for data labeling available for all ML specialists. Olga is responsible for providing human-labeled data for all AI projects at Yandex. She is in charge of integrating crowdsourcing into other business processes, such as customer support, product localization, software testing, etc. She graduated from the Saint Petersburg State University as a specialist in Mathematical Methods and Modeling in Economics. Also, she is a co-author of research papers and tutorials on efficient crowdsourcing and quality control at SIGIR, CVPR, KDD, WSDM, and SIGMOD.

🔗 `https://research.yandex.com/people/603770`

✉ `mailto:omegorskaya@yandex-team.ru`

Daria Baidakova, Yandex

Daria is responsible for consulting and educating Toloka requesters on integrating crowdsourcing methodology in AI projects. She also manages crowdsourcing courses at top data analysis schools (Yandex School of Data Analysis, Y-Data, etc) and organizes tutorials and hackathons for crowdsourcing specialists. Daria is a co-author of four hands-on tutorials on efficient crowdsourcing (at WSDM ’20, CVPR ’20, SIGMOD ’20, WWW’21) and a co-organizer of the crowdsourcing workshop at NeurIPS’2020. Prior to her work at Yandex, she conducted several education projects for youth while working at the UAE Minister's of Youth office in 2016–2017. She graduated from the London School of Economics and Political Science with MSc in Social Policy & Development (2018), and from New York University with BA in Economics (2017).

🔗 https://www.linkedin.com/in/dashabaidakova

✉ mailto:dbaidakova@yandex-team.ru

References

Ittai Abraham et al. 2016. How Many Workers to Ask?: Adaptive Exploration for Collecting High Quality Labels. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 473–482.

Michael S. Bernstein et al. 2010. Soylent: A Word Processor with a Crowd Inside. In *Proceedings of the 23Nd Annual ACM Symposium on User Interface Software and Technology*, pages 313–322.

Chris Biemann. 2013. Creating a system for lexical substitutions from scratch using crowdsourcing. *Language Resources and Evaluation*, 47(1):97–122.

Jonathan Bragg et al. 2018. Sprout: Crowd-Powered Task Design for Crowdsourcing. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, pages 165–176.

Chris Callison-Burch and Mark Dredze. 2010. Creating Speech and Language Data With Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 1–12.

Xi Chen et al. 2013. Pairwise Ranking Aggregation in a Crowdsourced Setting. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, pages 193–202.

Justin Cheng et al. 2015. Measuring Crowdsourcing Effort with Error-Time Curves. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 1365–1374.

A. Philip Dawid and Allan M. Skene. 1979. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):20–28.

Seyda Ertekin et al. 2012. Learning to Predict the Wisdom of Crowds. Proceedings of the Collective Intelligence 2012.

Ujwal Gadiraju et al. 2019. Crowd Anatomy Beyond the Good and Bad: Behavioral Traces for Crowd Worker Modeling and Pre-selection. *Computer Supported Cooperative Work (CSCW)*, 28(5):815–841.

Hideaki Imamura et al. 2018. Analysis of Minimax Error Rate for Crowdsourcing and Its Application to Worker Clustering Model.

Panagiotis G. Ipeirotis et al. 2014. Repeated labeling using multiple noisy labelers. *Data Mining and Knowledge Discovery*, 28(2):402–441.

Yuan Jin et al. 2017. Leveraging Side Information to Improve Label Quality Control in Crowd-Sourcing. In *Proceedings of the Fifth Conference on Human Computation and Crowdsourcing*, pages 79–88.

Jiyi Li and Fumiyo Fukumoto. 2019. A Dataset of Crowdsourced Word Sequences: Collections and Answer Aggregation for Ground Truth Creation. In *Proceedings of the First Workshop on Aggregating and Analysing Crowdsourced Annotations for NLP*, pages 24–28.

Christopher H. Lin et al. 2014. To Re(label), or Not To Re(label). In *Proceedings of the Second AAAI Conference on Human Computation and Crowdsourcing*, pages 151–158.

Greg Little et al. 2009. TurKit: Tools for Iterative Tasks on Mechanical Turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, pages 29–30.

Vikas C. Raykar et al. 2010. Learning From Crowds. *Journal of Machine Learning Research*, 11:1297–1322.

Paul Ruvolo et al. 2013. Exploiting Commonality and Interaction Effects in Crowdsourcing Tasks Using Latent Factor Models. In *NIPS '13 Workshop on Crowdsourcing: Theory, Algorithms and Applications*.

Aashish Sheshadri and Matthew Lease. 2013. SQUARE: A Benchmark for Research on Computing Crowd Consensus. In *Proceedings of the First AAAI Conference on Human Computation and Crowdsourcing*, pages 156–164.

Rion Snow et al. 2008. Cheap and Fast—but is It Good?: Evaluating Non-expert Annotations for Natural Language Tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 254–263.

Jing Wang et al. 2013. Quality-Based Pricing for Crowdsourced Workers. NYU Working Paper No. 2451/31833.

Peter Welinder et al. 2010. The Multidimensional Wisdom of Crowds. In *Advances in Neural Information Processing Systems 21*, pages 2424–2432.

Jacob Whitehill et al. 2009. Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise. In *Advances in Neural Information Processing Systems 22*, pages 2035–2043.

Ming Yin et al. 2013. The Effects of Performance-Contingent Financial Incentives in Online Labor Markets. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, pages 1191–1197.

Dengyong Zhou et al. 2015. Regularized Minimax Conditional Entropy for Crowdsourcing.