Small Data in NLU: Proposals towards a Data-Centric Approach

Alessandra Zarcone[1], **Jens Lehmann**[2] **and Emanuël A. P. Habets**[1]
[1] Fraunhofer IIS, Am Wolfsmantel 33, 91058 Erlangen, Germany
[2] Fraunhofer IAIS, Schloss Birlinghoven, 53757 Sankt Augustin, Germany
Corresponding author: alessandra.zarcone@iis.fraunhofer.de

Abstract

Domain-specific voice assistants often suffer from the problem of data scarcity. Publicly available, annotated datasets are in short supply and rarely fit the domain and the language required by a specific use case. Insufficient attention to data quality can generally be problematic when it comes to training and evaluation. The Computational Linguistics (CL) community has gained expertise and developed best practices for high-quality data annotation and collection as well as for for qualitative data analysis. However, the recent model-centric focus in AI and ML has not created ideal conditions for a fruitful collaboration with CL and the more data-centric fields of NLP to tackle data quality issues. We showcase principles and methods from CL / NLP research, which can potentially guide the development of data-centric NLU for domain-specific voice assistants - but have been typically overlooked by common practices in ML / AI. Those principles can potentially be of help to shape data-centric practices also for other domains. We argue that paying more attention to data quality and domain specificity can go a long way in improving the NLU components of today's voice assistants.

1 NLU from a model-centric to a data-centric approach

The rise of neural Natural Language Processing (NLP) has brought the focus of the community to network architecture and training methods and away from hand-coded features and representations. Such tendency has originated with the first end-to-end approaches to NLP (Church and Mercer, 1993; Jelinek, 1990; Brown et al., 1988), as hand-coded feature representations (for example from expert linguists) became unpopular and models were able to induce features directly from data. Jelinek was famously quoted to say "*Whenever I fire a linguist, our system performance improves*" (Jelinek, 1988). From this followed a general skepticism towards domain experts, not only in NLP but in AI and Machine Learning (ML) in general (Sambasivan et al., 2021). While domain expertise has been considered less relevant, a corresponding tendency has seen an "over-focus on numbers, on beating the state of the art [...] the Kaggle game" (Manning, 2015): given a benchmark, the goal is to find the best model architecture to beat the state of the art. When it comes to developing voice assistants, and in particular for Natural Language Understanding (NLU), playing the Kaggle game actually means playing the Snips game - to mention a popular dataset (Coucke et al., 2018) which has been used as a de facto standard benchmark for intent classification. Snips however only covers a handful of intents for the English language and has been shown to contain unrealistic utterances (see below). When benchmarks do not reflect the use of natural language in the target domain, a model's ability to beat them should always be taken with a grain of salt (Bender and Koller, 2020), as it might tell us very little about how the model will perform in real life.

More recently, a call for a more data-centric AI (Northcutt et al., 2021; Sambasivan et al., 2021) has shifted the focus from architectures, training methods, and leaderboards back to data quality,

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

tasks, features and representations. For academia, this means striving for a better understanding of the relationship between data and models. At the same time, for industry, it is finally an opportunity to understand how to solve data bottleneck problems making the best of small amounts of data. A data-centric approach requires a higher involvement of domain expertise than a model-centric approach. When it comes to language data, though, a lack of appreciation for domain experts among AI / ML practitioners has been a missed opportunity. Experts in CL and spoken dialogue systems have always focused on data, in particular, on dataset reliability (or lack thereof - see below) and annotator bias (Artstein and Poesio, 2008) and on the quality of crowdsourced data (Fort et al., 2011; Snow et al., 2008), have advocated for data ecology and domain scalability, collecting insights and methods to deal with issues not dissimilar from those the Data-Centric AI movement raises awareness for - insights which could prove beneficial also for other domains outside NLP.

2 Towards data-centric NLU

In the following, we go through different steps of an NLU pipeline for the development of domain-specific voice assistants to showcase principles and methods from CL and data-centric NLP research, which are typically overlooked by common practices and current tools in ML / AI and in industry. We formulate proposals for the integration of practices from computational linguistics and NLP, which can be beneficial to the field of data-centric AI and can lead to better and possibly cheaper language-based applications. Thereby, we make a case for a closer alliance between ML / AI and the CL and data-centric NLP community.

2.1 Data collection

Data used for training and evaluation of NLU modules should ideally match the expected interactions of humans with a voice assistant (in the desired domain or use case). However, collecting and annotating data can be resource-intensive, unless data from previous user interactions with a human associate is available. A typical solution is crowdsourcing (Snow et al., 2008), often combined with the Wizard-of-Oz paradigm (Budzianowski et al., 2018; Garcia et al., 2020). A common approach is to create a few template-based scenarios (e.g., *The restaurant should be in the **expensive** price range and should serve **Italian** food. Book a table for **5** people at **11:30 on Sunday***) and generate a number of variations just changing the entity fillers. The scenarios are then used as prompts to crowdsource dialogue data (Wang et al., 2012). The Wizard-of-Oz paradigm has been widely used by the spoken dialogue systems community, which has always taken ecological validity seriously (Rieser and Lemon, 2011; Schlangen, 2019). However, taking shortcuts in the data collection for NLU but also for ML in general may come at the cost of data quality. In a field where method sections in publications are typically short, these issues often go unnoticed, but a closer analysis shows how template-based scenarios result in scripted, repetitive dialogues, where the entity mentions are phrased in the same way and the same order as in the prompt, as in the MultiWOZ corpus (Budzianowski et al., 2018; De Vries et al., 2020). Snips also contains sentences in discordance with script and world knowledge (e.g., *In twenty-three hours and 1 second my daughter and I want to eat at a restaurant*), which have probably been generated semi-automatically. Situated scenarios, providing a more indirect description of the task, may go a long way in improving the quality of the collected data while still keeping the cost low (Frommherz and Zarcone, 2021).

Proposal: Bring back awareness of of data quality issues back into NLU and ML data collection practices, for example requiring that data collection methods are reported in detail.

2.2 Annotation reliability

Human-labeled training and evaluation data for supervised learning needs to be reliable. Data is said to be reliable if annotators can be shown to have internalized a similar understanding of the annotation guidelines and agree on the labels assigned, producing consistently similar results. This is only possible if the annotators follow an annotation scheme (set of labels and guidelines) that is valid, i.e., captures the "truth" of the phenomenon (Artstein and Poesio, 2008). Finding a valid scheme for arbitrary, use-case specific labels - such as intent labels - corresponds to defining labels human annotators can agree on, often with iterative annotation pilots that include discussion of edge cases and revision of the scheme and guidelines (Pustejovsky and Stubbs, 2012). Agreement, however, is

not always presented as a basic or a key feature of annotation tools (it is not integrated in Doccano or brat[1]) and is typically included only in more complex tools such as WebAnno or INCEpTION[2].

Furthermore, the arbitrariness of labels used in NLU (in particular intents and entities), which are commonly tailored to the use case, makes it difficult to obtain a reliable annotation without spending resources on defining a clear scheme. Whether *Turn on the AC* is classified as *AC_on* or as a more generic *device_on* intent (leaving the extraction of *AC* to an entity recognition module) is the designer's decision entirely and is not a representation of the meaning of the utterance or the user's intention. Relying on arbitrary domain-specific intents also affects scalability: datasets with different sets of intents may need to be relabeled before being used to train the NLU classifier for a specific system. Some suggest getting rid of intents altogether[3] and relying on dialogue modeling, which is typically based on dialogue act detection. There is not a standard annotation scheme for dialogue acts either: some schemes were developed for human-human interaction data (Bunt et al., 2020), some to meet domain-specific engineering requirements (Budzianowski et al., 2018), other (Pareti and Lando, 2018) may be domain-agnostic enough to be broadly applicable to human-machine interaction.

A clear scheme to achieve reliable annotation is beneficial not only for text processing but in any case where ML needs to learn from labeled data.

Proposal: Perform annotation pilots as standard practice in ML, report agreement scores as a measure of reliability. Rely on domain-agnostic labels whenever possible.

2.3 Benchmarking

We have already discussed how leveraging cost-effective methods for dataset collection may result in scripted or generally unrealistic data. Furthermore, many of such datasets are used as benchmarks for tasks different from what they were designed for, leading to yet more issues with the conclusions we can draw from the obtained results. For example, using Snips (Coucke et al., 2018), a dataset collected for non-incremental intent classification, to train or evaluate an *incremental* intent classifier may lead to inflated accuracy scores which are due to artifacts present in the data. For the partial utterance "I want to hear", the classifier will predict *PlayMusic* as the phrase is exclusively found in utterances for that intent in Snips. But that does not mean a human would expect the class *PlayMusic* for "I want to hear" until the words *song, tune, album* are heard, especially for a voice assistant that may play back different kinds of information beyond music (Hrycyk et al., 2021). The choice of datasets for ML training and evaluation is thus crucial to understand to what degree the performance of a classifier is attributable to overfitting or to artifacts in the evaluated dataset or if the classifier has learned a generalizable representation of the classes.

Proposal: Comparing ML evaluation with a human upper baseline based on human accuracy and agreement scores.

2.4 Data scarcity and domain specificity

Data scarcity is a classic obstacle to data-driven NLU. In the following, we review possible solutions to this problem from the point of view of data-centric AI.

Data augmentation / data programming Data programming efforts (see for example Snorkel[4]) provide a first attempt at augmenting data or at semi-automatically labeling data in a programmatic way, for example, by coding labeling functions to quickly label items that match a pattern or by replacing a word with one of its synonyms. However, we still lack an understanding of what and how many data manipulations have a significant effect on a model's performance.

Data quality Data quality can be assessed before training a model, or after. Assessing data quality before training, for example, using existing methods from corpus linguistics such as Type-Token ratio or MTLD (McCarthy and Jarvis, 2010), can provide some intuitive measure of the lexical diversity of the data, but can probably not offer much beyond what an experienced practitioner with in-domain knowledge can already see in the data (lack of lexical variety, repetitive syntactic structures). The eye of a domain expert can also spot unwanted bias, for example if commands in the passive voice are

[1] `https://github.com/doccano/doccano/`, `https://brat.nlplab.org/`

[2] `https://webanno.github.io/webanno/`, `https://inception-project.github.io/`

[3] `https://rasa.com/blog/its-about-time-we-get-rid-of-intents/`

[4] `https://www.snorkel.org/`

only present for one intent class but not for another or if only the masculine form is used. Assessing data quality after training, however, requires tracing a model's decision back to what the model has learned during training, which is particularly challenging for deep learning models. Cognigy[5] adopts a traffic-light system to show the designer if the quality of the training data is sufficient - however, the mechanism behind this system is not transparent for the user. Developers will likely benefit from iterative bootstrapping processes such as active learning (Laws et al., 2011; Yang et al., 2018), where at each iteration a small amount of maximally informative examples is chosen for annotation. The annotator's uncertainty (disagreement) or a classifier's uncertainty can be useful proxies to estimate the informativeness of a data point.

Algorithm-oriented solutions Possible solutions to solve the bottleneck of the data scarcity of labeled training data for a specific domain are few-shot and transfer learning approaches (Bengio, 2012). Alam et al. (2021) show that in order to annotate time expressions accurately in the voice assistant domain it is possible to leverage larger existing labeled datasets for different domains (mostly the news domain) and to fine-tune a model trained on out-of-domain data with a small amount of in-domain data, achieving considerable improvements with a small amount of in-domain data (less than 30 sentences).

Proposal: Develop practices to systematically augment data and quality measures to assess what properties of the data are helpful to improve a model's performance. Combine augmentation with algorithmic solutions for domain adaptation to leverage larger datasets by using smaller quantities of in-domain data.

2.5 The role of the human

The human factor is often seen as a fallible one, which impedes scalability. Ironically enough, though, the problems of domain specificity and data scarcity are often solved by involving human annotators - as long as something can be quickly transformed in a microtask for crowdsourcing, it is considered to be scalable and affordable. It is often the case, though, that the humans recruited via crowdsourcing platforms are underpaid and subject to tedious or even worse disturbing tasks, such as labeling offensive content (Barbosa and Chen, 2019).

It is possible that in some cases, the involvement of a human in the loop is unavoidable. However, it seems like a missed opportunity to simply resort to crowdsourcing for cheap annotation labor rather than thinking of how humans (possibly, domain experts) can be involved in the most effective way possible in automatic data processes and how human-ML teamwork can be achieved. Possible directions can range from iterative annotation processes (as the ones required by Active Learning) to fast prototyping, which allows testers to interact with the first version of a voice assistant and designers to use the interaction data as training data to improve the NLU of a voice assistant. As we learn more about how to interpret the behavior of a deep learning model and how to systematically manipulate data to improve the performance of the model, developing techniques that exploit the best of both worlds (the human intuition and the machine's scalable power) may be the key to solve many data scarcity and domain specificity problems and develop better ML / AI systems.

Proposal: Rather than outsourcing annotation to a cheap non-expert workforce, design tools that are understandable by domain experts without an ML background to facilitate their involvement, with the goal of understanding better how to improve the data and consequently the model's performance.

3 Conclusions

We have sketched some proposals to integrate existing methods and principles from CL / NLP into data-centric ML practices. Creating standardized methods for handling, evaluating, and augmenting data can increase replicability and help the community gain insights into the relation between data manipulation and ML model performance. This is particularly needed in a field such as NLU and voice assistant development, where the technology is developing quickly, but data constitutes the main bottleneck. We argue for closer collaboration between, on the one hand CL / NLP and, on the other hand, ML and AI to create standardized data manipulation and evaluation methods. We suggest that designs that take into account the inherent limitation of human annotators and ML models can pave the way for data- and human-centric approaches that make the best of both worlds.

[5] `https://www.cognigy.com/`

Acknowledgments and Disclosure of Funding

We acknowledge funding by the German Federal Ministry for Economic Affairs and Energy (BMWi) through the SPEAKER project (FKZ 01MK19011).

References

Touhidul Alam, Alessandra Zarcone, and Sebastian Padó. 2021. New domain, major effort? How much data is necessary to adapt a temporal tagger to the voice assistant domain. In *Procceedings of IWCS 2021*, page 144.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Natã M Barbosa and Monchu Chen. 2019. Rehumanized crowdsourcing: A labeling framework addressing bias and ethics in machine learning. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Yoshua Bengio. 2012. Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML workshop on unsupervised and transfer learning*, pages 17–36.

Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Frederick Jelinek, Robert L Mercer, and Paul Roossin. 1988. A statistical approach to language translation. In *Proceedings of COLING 1988*.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Harry Bunt, Volha Petukhova, Emer Gilmartin, Catherine Pelachaud, Alex Fang, Simon Keizer, and Laurent Prévot. 2020. The ISO standard for dialogue act annotation, second edition. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 549–558, Marseille, France. European Language Resources Association.

Kenneth Church and Robert L Mercer. 1993. Introduction to the special issue on computational linguistics using large corpora. *Computational linguistics*, 19(1):1–24.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Maël Primet, and Joseph Dureau. 2018. Snips voice platform: An embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*.

Harm De Vries, Dzmitry Bahdanau, and Christopher Manning. 2020. Towards ecologically valid research on language user interfaces. *arXiv preprint arXiv:2007.14435*.

Karën Fort, Gilles Adda, and K. Bretonnel Cohen. 2011. Last words: Amazon Mechanical Turk: Gold mine or coal mine? *Computational Linguistics*, 37(2):413–420.

Yannick Frommherz and Alessandra Zarcone. 2021. Crowdsourcing Ecologically-Valid Dialogue Data for German. *Frontiers in Computer Science*, 3:55.

Francisco J Chiyah Garcia, José Lopes, Xingkun Liu, and Helen Hastie. 2020. Crwiz: A framework for crowdsourcing real-time wizard-of-oz dialogues. In *Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, France.

Lianna Hrycyk, Alessandra Zarcone, and Luzian Hahn. 2021. Not so fast, classifier – Accuracy and entropy reduction in incremental intent classification. *Proceedings of the 3rd Workshop on NLP for Conversational AI (to appear)*.

Fred Jelinek. 1990. Self-organized language modeling for speech recognition. *Readings in speech recognition*, pages 450–506.

Frederick Jelinek. 1988. Applying information theoretic methods: Evaluation of grammar quality. Talk at the Workshop on Evaluation of NLP Systems, Wayne PA.

Florian Laws, Christian Scheible, and Hinrich Schütze. 2011. Active learning with amazon mechanical turk. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1546–1556.

Christopher D Manning. 2015. Computational linguistics and deep learning. *Computational Linguistics*, 41(4):701–707.

Philip M McCarthy and Scott Jarvis. 2010. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

Curtis Northcutt, Lu Jiang, and Isaac Chuang. 2021. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411.

Silvia Pareti and Tatiana Lando. 2018. Dialog intent structure: A hierarchical schema of linked dialog acts. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*. O'Reilly.

Verena Rieser and Oliver Lemon. 2011. *Reinforcement learning for adaptive dialogue systems: A data-driven methodology for dialogue management and natural language generation*. Springer Science & Business Media.

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

David Schlangen. 2019. Language tasks and language games: On methodology in current natural language processing research. *arXiv preprint arXiv:1908.10747*.

Rion Snow, Brendan O'connor, Dan Jurafsky, and Andrew Y Ng. 2008. Cheap and fast–but is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on Empirical Methods in Natural Language Processing*, pages 254–263.

William Yang Wang, Dan Bohus, Ece Kamar, and Eric Horvitz. 2012. Crowdsourcing the acquisition of natural language corpora: Methods and observations. In *2012 IEEE Spoken Language Technology Workshop (SLT)*, pages 73–78. IEEE.

Jie Yang, Thomas Drake, Andreas Damianou, and Yoelle Maarek. 2018. Leveraging crowdsourcing data for deep active learning an application: Learning intents in Alexa. In *Proceedings of the 2018 World Wide Web Conference*, pages 23–32.